EL PROFESIONAL de la INFORMACIÓN

Noviembre 2001 · ISSN 1386-6710 · Vol. 10, nº 11

ARTÍCULOS

Documentos y lenguaje de marcado: conceptos, problemas y tendencias

Por **Mela Bosch**

RECURSOS INFORMATIVOS EN LAS REDES

Papi: acceso ubicuo a recursos de información

Por **Rodrigo Castro-Rojo** y **Diego R. López**

La *Biblioteca virtual Miguel de Cervantes*

Por **Vicente Sabido**

SISTEMAS DE INFORMACIÓN

Microsoft: SharePoint server

Por **Ricardo Eíto Brun**

EL PROFESIONAL *de la* **INFORMACIÓN**

Revista fundada en 1992 por **Tomàs Baiget** y **Francisca García-Sicilia**

Dirección editorial:
Tomàs Baiget
baiget@sarenet.es

Pedro Hípola
phipola@ugr.es

Subdirector:
José A. Senso
jsenso@platon.ugr.es

Redactor jefe:
Jesús Castillo Vidal
jcastillo@baratz.es

Redactores:
Lluís Codina
lluis.codina@cpis.upf.es

Alfons Cornella
cornella@infonomics.net

Ricardo Eíto Brun
ricardo.eito@adecco.es

Cristina Gª Testal
cristina.garcia-testal@uv.es

Mari Carmen Marcos
mcmarcos@posta.unizar.es

José Antonio Ontalba
jontalba@campus.uoc.es

PUBLICIDAD:
Alice Keefer
akeefer@arrakis.es

REDACCIÓN:
Apartado 32.280
08080 Barcelona
Fax: +34-934 250 029
epi@sarenet.es

IMPRIME:
Gráficas Alhambra
Pol. Tecnológico Ogíjares, 50-51
18151 Granada
Tel.: +34-958 506 742
galhambra@alis.net

SUSCRIPCIONES:
Henrik Lang
Swets & Zeitlinger Publishers
http://www.szp.swets.nl/szp/orderjournals/subscr_pi.htm
orders@swets.nl

PRODUCCIÓN Y DISEÑO:
Luis Pardo Gallo
Virgen de Montserrat, 19-1º
18008 Granada
Tel.: +34-958 135 994
luispard@wanadoo.es

El profesional de la información se publica mensualmente
(excepto enero y agosto) por:
Swets Blackwell
Caspe, 46 · 08010 Barcelona (España)
http://www.szp.swets.nl/szp/journals/pi.htm
Atención al suscriptor:
Tel.: +34-932 701 144; fax: +34-932 701 145
Suscripción online:
http://www.szp.swets.nl/szp/orderjournals/subscr_pi.htm

SWETS & ZEITLINGER PUBLISHERS

Swets & Zeitlinger Publishers
PO Box 825, 2160 SZ Lisse (Holanda). Tel.: 900-993 180; fax: +31-252 415 888
orders@swets.nl *http://www.szp.swets.nl*

Director de publicaciones:
John Martin

Depósito legal: B-12303-97

Los contenidos de **El profesional de la información** están referenciados en los siguientes servicios bibliográficos y bases de datos:

Bedoc
http://www.inforarea.es/bedoc.htm

Bulletin Board for Libraries (Bubl)
http://bubl.ac.uk/journals/lis/ae/epdli

Compludoc
http://www.ucm.es/BUCM/complu

Consorci de Biblioteques Universitàries de Catalunya (Cbuc)
http://sumaris.cbuc.es/13866710.htm

Datathéke
http://milano.usal.es/dtt.htm

DoIS (Documents in Information Science)
http://dois.mimas.ac.uk/DoIS/data/julqtichq.html

Ebsco Online
http://www-sp.ebsco.com/online/direct.asp?JournalID=105302

Guíame
http://www.guiame.net/lista-fuentes.html

Índice Español de Ciencias Sociales y Humanidades (Isoc)
http://www.cindoc.csic.es/prod/isoc-cd.html

Information Science Abstracts (ISA)
http://www.infotoday.com/catalog/infosvcs.htm#info3

Information Services in Physics, Electronics and Computing (Inspec)
http://www.iee.org.uk/publish/inspec

IWEb
http://www.doc6.es/iwe

Library and Information Science Abstracts (Lisa)
http://catalogue.bowker-saur.co.uk/data/198.html

Registros Bibliográficos para Bibliotecas Públicas Españolas (Rebeca)
http://www.mcu.es/REBECA/que.html

Resúmenes de Información y Documentación (ReID)
http://www.sisdoc.es/servicios/reid.htm

RoweCom Information Quest
http://www.informationquest.com

Swetsnet*Navigator*
http://www.swetsnetnavigator.com/direct/SB_issue?title=2246163

Los trabajos de la sección "Artículos" son aprobados según el sistema tradicional "peer review": al menos dos expertos en el tema, del *Consejo Asesor* de la revista y/o externos, deben dar el visto bueno antes de su publicación.

Para conseguir que los trabajos no pierdan su actualidad, la dirección y los evaluadores de esta revista ponen especial esfuerzo en revisar los artículos con gran rapidez, consiguiendo un tiempo medio de aceptación o rechazo de los trabajos de sólo unas pocas semanas.

SUMARIO

Código ético del sector de información electrónica

EL SECTOR DE INFORMACIÓN ELECTRÓNICA de la *Asociación Multisectorial de la Información* (*Asedie*) ha elaborado un código deontológico con el fin potenciar la buena conducta profesional y, en particular, para autorregular el sector dentro del marco legal existente en la actualidad, de forma que las disposiciones que contiene constituyan un manual de instrucciones de aplicación de la ley y su normativa de desarrollo.

El código pretende defender la honestidad en la competencia y velar por que la actuación de los asociados se acomode a las normas de ética establecidas, siendo una garantía de trato comercial correcto, tanto para los clientes como para los miembros de *Asedie* y para el mercado en general. En él se regula la relación entre las empresas e instituciones miembros de la *Asociación* de tal forma que se garantiza un trato transparente, y los contactos se desarrollan en armonía con los principios de la libre competencia. Asimismo, se reserva un capítulo para definir las relaciones con los consumidores, en el que los socios se comprometen a tratar de forma confidencial y veraz la información difundida.

Algunas de la personas que han intervenido en la redacción del código: Rosa de la Viesca y Carmen Vidal (Cindoc / Csic), Francisca García-Sicilia (Dialog España), Jaime Tascón (Asedie) y Mª Teresa Abella (Institut Català de Tecnologia, ICT)

A tenor de este Código los titulares de los datos se benefician de una amplia protección que se traduce en el derecho a la información en la recogida de los datos y los derechos de acceso, rectificación, cancelación y oposición.

La creación de un Comité Disciplinario permitirá verificar la aplicación de los preceptos del Código, apreciando el comportamiento de los miembros y sancionando, en su caso, el incumplimiento. De esta forma se pretende preservar la calidad de las empresas dedicadas a la recogida y distribución de información electrónica

Las normas son de obligado cumplimiento para todos los miembros de la *Asociación*.

Asedie.
http://www.asedie.es

Artículos

Documentos y lenguaje de marcado: conceptos, problemas y tendencias

Por **Mela Bosch**

Mela Bosch

Resumen: *Se presenta un panorama de la estructuración de información documental por medio de lenguajes de marcado. Se hacen notar dos líneas de problemas: de estándares de descripción y del mercado de software. Dentro de la primera se describe la sintaxis concreta de sgml, el subconjunto html y la extensión xml. Respecto a la segunda se tipifica la oferta en el mercado. Finalmente se considera el lugar del profesional de la información en este contexto, enfatizando su importancia para el desarrollo de la web semántica en general y, especialmente, en herramientas de apoyo a la misma, como las ontologías.*

Palabras Clave: *Estructuración de información, Información documental, Lenguaje de marcado, Organización del conocimiento, Web semántica.*

Title: Documents and markup language: concepts, problems and trends

Abstract: *This article presents a panorama of the documental information that is structured using markup languages. This subject presents two different issues: description standards and software market limitations. Within the first we find sgml concrete syntax, the html subset and xml. The second presents itself as a tipification of software products in the market. Having described these matters, the article reflects upon the role of the information professional in this context emphasizing his importance for the development of the Semantic Web in general and the tools that such Web needs, such as ontologies.*

Keywords: *Documental information, Knowledge organization, Markup language, Semantic web, Structured information.*

Bosch, Mela. *"Documentos y lenguaje de marcado: conceptos, problemas y tendencias". En: El profesional de la información, 2001, noviembre, v. 10, n. 11, pp. 4-9.*

Introducción: una gran solución y grandes problemas

El desarrollo de los lenguajes de marcado cambió el panorama de la estructuración de documentos, a pesar de que no era su objetivo inicial. En la década de los años 80 se planteó como una forma de asegurar la transferencia de datos manteniendo aspectos de presentación.

Sgml fue desarrollado en sus orígenes (1969) por **Charles Goldfarb** con el nombre de *general markup language* (*gml*), y tenía como objetivo facilitar el intercambio de documentos en *IBM*. Permite definir estructuras tipificadas de documentos a partir de las reglas que los rigen, las cuales son expresadas físicamente en los documentos por medio de marcas (señales o códigos convenidos). Fue adoptado rápidamente para la circulación de documentos por el *Departamento de Defensa* de EUA y por la *Oficina de Publicaciones Oficiales* de la *Comunidad Europea*, ambos clientes de *IBM* (**Bryan**, 1998). En una meteórica gestión (ya que ni siquiera fue previamente autorizado como estándar americano), *ISO* lo lanzó como la norma *ISO 8879* en 1986 con el nombre de "standard generalized markup language" (sgml).

Su desarrollo no se produjo de manera anecdótica, ya que la historia ha marcado dos de las características de sgml: está pensado para sistemas de oficina, es decir, para documentos corporativos, y está orientado al procesamiento de datos. Más concretamente dentro de este ámbito, al intercambio.

Si se observa la portada de la norma *ISO 8879* los descriptores indican claramente su contenido: *Data processing, Documentation, Logical structure, Programming (computers), Artificial languages, Programming languages* (*ISO*, 1986).

Sgml es, por lo tanto, un lenguaje artificial útil para representar la organización lógica de documentos a efectos de procesamiento y programación. En los años siguientes apareció como la gran solución para la transferencia de datos documentales con una arquitectura lógica determinada. Sin embargo, a la vez se generaron grandes problemas: de estándares de descripción y de mercado de software.

Mela Bosch, Profesora e investigadora de la Universidad Nacional de La Plata y profesora de la Universidad Nacional de Mar del Plata, Argentina.
euris@sinectis.com.ar

Fecha de recepción 13-09-01
Aceptación definitiva: 2-11-01

Antes de puntualizar sobre estos temas agregamos una reflexión: tal y como puede deducirse de lo que se ha comentado hasta ahora sobre su origen, y según nos hace notar **Winograd** (de forma pionera hace más de 18 años), el problema de la descripción y representación de las estructuras conceptuales en medios informáticos estuvo ligado y determinado por las necesidades de proceso. Era más importante lograr similitud con el lenguaje de las máquinas que con el de las personas: "*In designing a programming language there are two potentially conflicting sources of criteria or ganization (things can be classed as alike because they are implemented with the same underlying mechanism, or because they have similarity for the language user). In most programming languages, the balance lies to wards the implementation side*" (**Winograd**, 1983, p. 417).

«El desarrollo de los lenguajes de marcado cambió el panorama de la estructuración de documentos, a pesar de que no era su objetivo inicial»

Esto nos lleva a una discriminación conceptual más general: los lenguajes, a los que definimos como conjuntos de símbolos normalizados para la representación y comunicación, se dividen en: naturales, controlados y artificiales.

Los primeros son los que se crean en un contexto social y convencional por conjuntos de individuos. Incluyen tanto las lenguas de las comunidades humanas como las formas de comunicación animal (aún hay muchos que discuten su valor y otorgan sólo al ser humano la capacidad de poseer lenguaje). Por su parte, los controlados son porciones del lenguaje natural normalizados para fines específicos: entre éstos se encuentran los lenguajes documentales y las ontologías. Finalmente los artificiales son conjuntos de símbolos arbitrarios también para fines concretos, incluyéndose entre ellos: los de programación —destinados al procesamiento— y los lenguajes de descripción, como sgml, cuyo objetivo es la interpretación por parte de lenguajes de programación (**Isasi**, 1997).

Ésta es una mera simplificación expositiva que nos permitirá avanzar operativamente en el desarrollo de este trabajo; no ignoramos que existen formas combinadas complejas como los lenguajes de transcripción, los de restricciones y los lenguajes gráficos, entre otros.

Es importante no olvidar que con sgml nos encontramos ante un lenguaje de naturaleza artificial cuya meta es la descripción de información para facilitar su proceso; de donde surge el tema de qué es lo que se describe y para qué, produciéndose así un efecto de retroalimentación sobre el lenguaje. Este punto nos conduce al primero de los problemas que estamos analizando.

Problemas de estándares de descripción

Por lo que puede deducirse de todo lo comentado hasta ahora, el objetivo de sgml de permitir la transmisión de datos determina la forma de descripción. La ventaja es que se permite establecer grupos de códigos propios según la característica de la estructura documental. Esto da origen a varias líneas de cambios dentro del lenguaje de marcado:

—Sintaxis concreta.

—Subconjuntos.

—Extensiones.

1. Sintaxis concreta. En el campo de la lingüística se denomina sintaxis a las reglas que definen el significado de los códigos y a los nombres reservados usados por un lenguaje en particular. Al tratarse de un lenguaje de descripción, sgml necesita ajustarse al objeto de su descripción y, a la vez, mantener consistencia en códigos y denominaciones. La forma de lograrlo es por medio de dos tipos de sintaxis: la abstracta y la concreta.

La primera es usada para especificar cómo deben ser escritas tanto las declaraciones de sgml como las de tipo de documento. Por otra parte, nos encontramos con el conjunto de reglas utilizadas para definir cómo deben ser codificados documentos específicos. Una forma particular de sintaxis concreta, llamada en inglés *reference concrete syntax*, fue definida de manera explícita en la *ISO 8879*. Su objetivo es suministrar una referencia para que las sintaxis concretas eventuales la tomen como base. Se conoce como *international reference version* (*IRV*) y contiene los siguientes puntos:

—Códigos que deben ser ignorados (*shunned character number identification*).

—El conjunto básico de caracteres *baseset* definidos por la norma *ISO 646*.

—Códigos ascii homologados.

—La ampliación de ese conjunto *descset*.

—Códigos de caracteres de funciones (*function character identification*).

—Reglas de denominación para cuando se crean entidades o etiquetas (*naming rules declaration*).

—Conjunto de delimitadores alternativos de marcado (*delimiters declaration*).

—Convenciones de denominación de declaraciones (*reserved name use*).

—Las cantidades permitidas para el anidamiento de elementos y entidades.

Esta sintaxis es asumida automáticamente por los sistemas que procesan sgml. Para ello es necesario indicarlo de la siguiente forma: *syntax public "iso 8879-1986//syntax reference//en*"

A esta sintaxis pueden agregarse esquemas propios de codificación. La sintaxis concreta de referencia debe ser entonces conocida y comprendida por sus usuarios potenciales, sean humanos o máquinas. Los cambios en las opciones por defecto se deben hacer de acuerdo con ciertas reglas, declarando públicamente (es decir de forma explícita) sus variantes (**Bryan**, 1998).

Existen tres formas de especificar las variedades de sintaxis concreta en un documento:

a. Por medio del uso de la opción *switches*: se indican puntualmente diferencias de la sintaxis concreta de referencia o de cualquier otra que se utilice. Permite establecer pares de caracteres que pueden ser cambiados. Por ejemplo, si se quiere indicar que en lugar de usar "[ ]" (códigos ascii 91 y 93, homologados *ISO 646*) se prefiere utilizar: "{ }" (cuya codificación es 123 y 125) se haría siguiendo el procedimiento que a continuación se expone:

syntax public "iso 8879-1986//syntax reference//en"

switches 91 123

93 125

b. Declarando públicamente que se usa una variante de sintaxis concreta:

syntax public "iso 8879-1986//syntax prueba//sp"

Con esto se informa que se emplea una sintaxis llamada *prueba* y que recoge el juego de caracteres en español. Ésta, de nueva creación, debe estar redefinida cubriendo todos los aspectos de la *IRV* mencionada anteriormente, asumiendo algunos puntos concretos y modificando un grupo particular que incluye, por ejemplo, entidades o elementos (**Bryan**, 1998). Con esto se elaboran sintaxis concretas orientadas a usos de algunas áreas temáticas y organizaciones.

c. Produciendo declaraciones de tipo de documento (*dtd*) que engloban las variantes de sintaxis respecto a la básica de sgml.

Las *dtds* están normalizadas siguiendo la *ISO 12082*, que determina un marco de trabajo general. Además, nos encontraremos con otras *dtds* que responden a iniciativas con fines específicos, como *Sgml initiative in health care* (*Hl7 health level-7 and sgml/xml*) o *National Center for Biotechnology Information* (*Ncbi*), de la *National Library of Medicine* y de los *National Institutes of Health* (**Cover**, 2000).

Un importante esfuerzo común es la *dtd* de *tei* (*text encoding initiative*) financiada por varias universidades y la UE, cuyo objetivo es construir una referencia universal para la codificación de textos (*Tei*, 2001).

Existen también otras *dtd*s valiosas para el profesional de la información como *ead* (*encoded archival description*) para documentos corporativos y de archivo, mantenida por la *Network Development and Marc Standards Office* de la *Library of Congress* (*LoC*) junto con la *Society of American Archivists.* Esta norma está en directa relación con la específica de marc: *marc dtd* (*machine readable cataloging document type definition*).

«El problema de la descripción y representación de las estructuras conceptuales en medios informáticos estuvo ligado y sobredeterminado por las necesidades de proceso»

2. Subconjuntos. Pueden ser más simplificados o más completos que las sintaxis concretas. Lo importante es que toman aspectos de la sintaxis abstracta. El más conocido es html, con versiones sucesivas que aparecen de forma continua y cuyo formato de intercambio se encuentra definido por la norma *ISO 9069*.

Existe también la *ISO 10744*, que desarrolla el *hypermedia/time-based structuring language* conocido como *HyTime.* Se trata de un subconjunto con una sintaxis propia que permite la representación en hipermedia vinculada a la evolución temporal. Incorpora técnicas que permiten asociar información adicional sin que haya sido almacenada como parte del documento fuente. *HyTime* puede referenciar cualquier texto, imagen, efecto sonoro o área espacial dentro de una publicación multimedia.

Un aspecto destacable es que hace posible que las direcciones de información se almacenen de manera independiente al sistema, lo que posibilita la gestión automatizada de ficheros. De esta forma se usa para verificar y cambiar direcciones cuando un documento se reestructura o es transferido a una nueva ubicación.

Esta norma define formas arquitectónicas de documentos, que en realidad no son más que reglas para crear y procesar los componentes de éstos. Se definen cuatro tipos de formas arquitectónicas en la *ISO/IEC*

10744:1997: de elementos, de atributos, de entidad de datos y de atributos de datos.

Así pues, se constituyen los *architectural form definition requirements* (*afdr*), es decir los requerimientos de arquitectura, cuyas especificaciones se pueden encontrar en el anexo 1 de *ISO/IEC 10744:1997*. Su uso se declara así:

<!afdr "iso/iec 10744:1997">

Existen, además, aplicaciones de *HyTime*, como *smdl* (*standard music description language*, *ISO/IEC DIS 10743:1995*).

3. Extensiones. La fundamental es xml, que puede entenderse como un dialecto de sgml (se suele usar el término dialecto porque contempla alteraciones estructurales respecto del lenguaje madre). Es una versión extremadamente simplificada de sgml (algo así como un sgml genérico) que admite el procesamiento en web de objetos, algo que no es posible hacer en html. Permite definir objetos como entornos de realidad virtual, movimiento, señales olfativas, etc.

En realidad es un metalenguaje, ya que permite a su vez crear subconjuntos. De hecho se está reescribiendo todo html en xml. En esencia lo que permite es una forma flexible, pero normalizada, de añadir etiquetas a los documentos. En html, que tal y como hemos visto es un subconjunto, las etiquetas son acotadas y tienen una semántica específica para la navegación hipertextual. Por su parte, sgml es excelente y completo, pero a la hora de implementar la navegación resulta complicado dado que los enlaces son "artesanales". Xml no sustituye a éste, que sí puede gestionar documentos xml, favoreciendo la funcionalidad de navegación y la manipulación de diferentes objetos no sólo textuales.

«La web semántica brindará contenido significativo a las páginas de la Red, creando un ambiente donde agentes de software puedan fácilmente efectuar tareas sofisticadas»

En su aspecto formal xml se nutre de reglas de producción lógicas para interpretar la sintaxis y permite su compilación. Define una regla básica y establece el valor de cada uno de los elementos.

En cuanto a sgml, los valores ya están establecidos en la sintaxis abstracta de sgml o en las *dtd*. Este lenguaje, al haber sido concebido para describir documentos de texto, estipula la forma de organización, o sea, no sólo la sintaxis sino la estructura. Por lo tanto no es posible poner un título de nivel 1 después de uno de nivel 3, por ejemplo, porque tiene restricciones para anidar etiquetas. Estas limitaciones, como ya indicamos con anterioridad, están en la *IRV*, de manera que hay que indicarlo cada vez que se modifican las restricciones y la sintaxis básica, o bien declarar una *dtd* con todos los cambios.

Xml no requiere una *dtd*, pero es necesario realizar otras especificaciones y declaraciones. Por ejemplo, se debe señalar quiénes son los usuarios: navegadores, etc. También en forma de declaraciones se deben indicar elementos, atributos, entidades, así como anotaciones para especificar datos externos.

Aunque se ha logrado normalizar la parte de xml correspondiente para libros, artículos y series con la norma *ISO 12083*, el problema es que en este momento nos encontramos en medio de una gran número y proliferación de variantes. Además, el subconjunto html se origina con extensiones que cambian según las ofertas de productos. Esto nos lleva a los problemas que analizaremos en el siguiente punto.

Problemas del mercado de software

Si bien es posible codificar textos sgml en cualquier editor de texto, la tarea artesanal es muy lenta y susceptible de que se produzcan errores. Se utilizan tres tipos de software para sgml: editores, analizadores (*parsers*) y visualizadores (*browsers*).

El editor es el que permite redactar *dtds* de documentos. La gama oscila desde los más sencillos freeware —software gratuito— hasta los más complejos que permiten convertir texto procedente de otro origen; otros integran hojas de estilo y entremezclan sus posibilidades con análisis y visualización.

El analizador o *parser* compara las reglas de edición con la sintaxis concreta o la *dtd*. Trabajan a la par que el editor —o sobre documentos una vez compuestos— para verificar o modificar el marcado. Algunos pueden trabajar con hojas de estilo *dsssl* (*document style semantics and specification language*, *ISO 10179*) y *spdl* (*standard page description language*, *ISO 10180*). Las primeras permiten agregar abundantes detalles tipográficos mientras que la segunda está más orientada a la impresión.

Finalmente los visualizadores o *browsers* permiten recorrer el documento en toda su arborescencia. El líder en el mercado es *SoftQuad*, con *Xmetal*, cuya versión de prueba puede ser descargada de la web.

El panorama de editores, analizadores y visualizadores es amplio y confuso. Se crean subconjuntos, sintaxis concretas y *dtd*s de acuerdo con la conveniencia de las empresas, no existiendo además compatibilidad total entre ellos. Los más populares resultan los edito-

res html que permiten diseñar páginas web, entre los cuales encontramos *FrontPage*, de *Microsoft*, y *Composer*, de *Netscape*. Tienen la ventaja de que no es necesario conocer los lenguajes de marcado, pero no suelen ser totalmente compatibles entre ellos y con otros editores.

En cuanto a xml estamos en plena efervescencia. Su importancia estriba en que permite encapsular instrucciones de programación, ya sea *VBScript* o *Java applets*, compilar controles *ActiveX* y hasta lenguajes antiguos como *Cobol*. También tiene capacidades de edición, análisis y visualización, pero como además cuenta con posibilidades de programación, requiere un lenguaje especial para modelar los datos. Es aquí donde aparece *document object model* (*dom*) con adaptaciones para *Explorer* y *Netscape*. *Microsoft* desarrolló *com* (*component object model*) para el desarrollo de aplicaciones con xml.

En esta babel en la que nos encontramos, xml empieza a ser la *lingua franca* que permitirá crear y compartir documentos. Pero aún estamos lejos de esa situación: las ofertas de software se multiplican, los productos son costosos y la capacitación que requieren es grande (**Conallen**, 2001).

La web semántica, un futuro

La pregunta para los profesionales de la información es saber qué lugar les cabe en este panorama de estructuración de la información documental. Su posición puede estar, sin duda, en el apoyo que puedan prestar a los desarrolladores, trabajando en la definición de *dtd*s y sintaxis concretas adaptadas a las necesidades de cada corpus de información. En este sentido, el estudio y capacitación en este contexto es indispensable. Pero otro punto donde serán de inestimable valor es en el desarrollo de la web semántica.

Con el uso de los lenguajes de marcado, los sistemas informáticos en la www realizan tareas de rutina y también otras más expertas, siempre y cuando se trate de trabajar dentro de las páginas web. Sin embargo no se tiene aún un procedimiento fiable para procesar la semántica. Por esta razón se producen avances en lo que se está llamando web semántica.

En los párrafos que siguen haremos una reseña comentada del artículo de **Berners-Lee**, **Hendler** y **Lassila** titulado "The semantic web, a new form of web content that is meaningful to computers will unleash a revolution of new possibilities", publicado en *Scientific American* en mayo de 2001, desde la perspectiva de interés para el profesional de la información.

Según estos autores, la web semántica brindará contenido significativo a las páginas de la Red, creando un ambiente donde agentes de software, moviéndose de una página a otra, puedan fácilmente efectuar tareas sofisticadas para usuarios. Según ellos no será una red separada sino una extensión de la actual en la que la información tenga un significado definido, permitiendo que máquinas y personas puedan trabajar en cooperación. Estos autores nos dicen que para que la web semántica sea operativa los ordenadores deben tener acceso estructurado a colecciones de información y conjuntos de reglas de inferencia que ellos pueden usar para conducir al razonamiento automatizado.

En el documento mencionado se indica que los sistemas de representación de conocimiento (tal y como se denomina esta tecnología) han sido centralizados, requiriendo que todos compartan exactamente la misma definición de conceptos. Por otra parte limitan el tipo de preguntas que pueden formularse y, para evitar tales problemas, estos modelos de representación han tenido un propio y limitado conjunto de reglas para hacer inferencias sobre sus datos.

En cambio los autores dicen que el desafío de la web semántica está en proveer un lenguaje que exprese tanto los datos como una lógica para darles sentido de diferente origen y forma, valiéndose para ello de tres tecnologías: xml, *resource description framework* (*rdf*) y las ontologías.

—Xml, sobre el que ya hablamos, es importante en su doble carácter de metalenguaje de descripción que permite crear etiquetas, tanto propias como estándar, y a la vez como lenguaje de especificación de puntos de ejecución de porciones de software o programas completos que pueden hacer uso de ellas.

—*Rdf* (marco de descripción de recursos) es el complemento de xml. El documento que estamos reseñando nos indica acertadamente el etiquetado genérico xml, que permite agregar marcas arbitrarias a los documentos, las cuales no proveen semántica. Su significado podría ser expresado por el *rdf*, que son conjuntos de tríos —constituido cada uno de ellos por el sujeto, verbo y predicado de una frase elemental— y que pueden escribirse usando etiquetas de xml.

En síntesis, tal como plantean estos autores, *rdf* es un lenguaje de metadatos genérico que ofrece una manera natural para describir la mayoría de los objetos procesados. Cada sujeto y cada predicado son identificados por el *universal resource identifier* (*uri*) —recordemos que url es el tipo más difundido de *uri*—.

Volviendo a lo que se reseña en el artículo que nos ocupa y que estamos comentando, los verbos podrán ser identificados también por *uri*s, permitiendo definir otros nuevos simplemente definiendo un *uri* para ellos. De esta forma, los autores nos hacen notar que el lenguaje natural puede usar un mismo término para signi-

ficar algo distinto, según el contexto, y utilizar un *uri* diferente para cada concepto específico. Los tríos de *rdf* forman redes de información entre objetos conexos, y dado que éstos usan *uri*s para codificar datos en un documento, se asegura que los conceptos no son simplemente palabras sino que las vinculan a una definición única que todos pueden encontrar en el web.

Los uri aún no se encuentran muy difundidos y no lo estarán, creemos, hasta que una norma ISO los generalice, algo que, esperamos, ocurra en breve. Sin embargo, coincidimos con **Berners-Lee** en que esto no evita la superposición, pues dos recursos, por ejemplo dos bases de datos en línea, pueden usar diferentes identificadores para lo que de hecho es el mismo concepto.

Los sistemas informáticos deben disponer de un mecanismo que permita descubrir tales significados comunes para cualquier base de datos a la que se conecten. En este caso interesa encontrar una forma de representar semántica común en estructuras diferentes.

Berners-Lee, **Hendler** y **Lassila** confían en que la solución a este problema se puede encontrar en el tercer componente básico de la web semántica:

—Las ontologías. Este término tiene su origen en la filosofía, y la definición más general en el contexto de los sistemas informáticos es el de la especificación de una conceptualización.

Cada uno de los conceptos es expresado en una red terminológica que define sus atributos. Por ejemplo tipos de datos y sus comportamientos tales como relaciones con otros conceptos. Además, tienen una forma de establecer el alcance de estos atributos y comportamientos por medio de reglas, permitiendo que la ontología deduzca, o por lo menos proponga, a qué clase o categoría puede pertenecer cada nuevo concepto que se ingresa.

Sintéticamente una ontología tiene una taxonomía de conceptos —que define clases de objetos o conceptos y relaciones entre ellos— junto a un conjunto de reglas de inferencia. Según estos autores, puede facilitar el funcionamiento de la web para mejorar la exactitud de la recuperación, ya que el programa adquiere capacidad para realizar la búsqueda sólo en las páginas que se refieren al concepto preciso y, a la vez, funciona también cuando la respuesta no radica en una única página. Además, con las ontologías comienzan a surgir soluciones a los problemas terminológicos en la web. El significado de términos o de códigos xml usados en una página puede ser definido desde ésta a una ontología.

Dejamos ahora la reseña del documento sobre la web semántica y retomamos la reflexión inicial sobre los lenguajes naturales, artificiales y controlados. En este contexto las ontologías se sitúan dentro de los lenguajes controlados. Es fácil ver la vinculación entre ellas y los lenguajes documentales, pues muchos de sus fines son similares.

Es importante destacar que la inteligencia artificial, después de años de lucha con el procesamiento del lenguaje natural, parece haber optado como solución global el uso de los lenguajes controlados.

La incógnita es si los profesionales de la información seremos capaces de ampliar nuestras miras desde la tradicional representación de información concebida para recuperación en un sistema automatizado y asumiremos trabajos para enfocar la representación de conocimiento en la web.

Bibliografía

Berners-Lee, T.; **Hendler, J.**; **Lassila, O**. "The semantic web, a new form of web content that is meaningful to computers will unleash a revolution of new possibilities". En: *Scientific American*, 2001, mayo. Consultado en: 2001.
http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

Bryan, M. *Sgml: an author's guide*. New York: Addison-Wesley, 1998.

Conallen, J. *Building web applications with UML*. Boston: Addison-Wesley, 2000.

Cover, R. *Sgml/xml applications in cross-domain and multi-disciplinary enterprises*. *Oasis*. Consultado en: 2001.
http://www.oasis.org

Cover, R. Publicly available software for sgml/xml/dsssl. Consultado en: 2001.
http://www.oasis.org

Ead (*Encoded archival description*)**.** Consultado en: 2001.
http://www.loc.gov/ead/

International Standards Organization. IS0 8879 standard generalized markup language. Ginebra: 1986.

*International Standards Organization. ISO/IEC 10744 hypermedia/time-based structuring language***.** Ginebra: 1997.

Isasi, P.; **Martínez, P.**; **Borrajo, D.** *Lenguajes, gramáticas y autómatas*. New York: Addison-Wesley, 1997.

Jacquesson, A.; **Rivier, A.** *Bibliothèques et documents numériques*. Paris: Electre, 1999.

Marc dtd, machine readable cataloging document type definition. Consultado en: 2001.
http://www.loc.gov/marc/marcdtd/marcdtdback.html

TEI guidelines for electronic text encoding and interchange: interchange, 2001. Consultado en: 2001.
http://etext.lib.virginia.edu/TEI.html

Winograd. T. *Language as a cognitive process.* New York: Addison-Wesley, 1983.

Se cuestiona el papel de los editores

LA MITAD DEL COMITÉ DE REDACCIÓN de la revista *Machine learning* —una de las mejores en temas de inteligencia artificial, publicada desde hace ya 15 años por *Kluwer Academic Publishers*— la abandonó el pasado 8 de octubre de 2001 como protesta porque la revista era demasiado cara (la suscripción anual costaba unas 79.000 pesetas —475 euros—) y porque el editor ponía obstáculos a la libre circulación de los artículos —por ejemplo, no deja que los autores cuelguen en webs particulares los artículos ya publicados—.

Los que se han marchado han ofrecido sus servicios a una revista con menos solera, que sólo tiene 18 meses de vida, pero directamente competidora. Se publica gratuitamente por el web: *Journal of machine learning research*, de la que posteriormente la editorial *MIT Press* realiza una edición impresa de pago.

Tratando de contrarrestar esta acción *Kluwer* ha reducido el coste de la suscripción a 23.000 PTA (138 euros) y también incrementará en un 18% el número anual de páginas. **Michael I. Jordan**, profesor de informática y estadística en la *University of California*, en Berkeley, líder de ese grupo de 40 disidentes, opina que esto la editorial lo tenía que haber hecho al menos hace ya 6 años.

Varios miembros del comité editorial de la revista *Logic programming* también dimitieron en 1999 para unirse a *Theory and practice of logic programming*, una revista publicada por *Cambridge University Press*, editorial que fomenta la difusión electrónica de artículos.

La Red va imponiendo su modelo

Éstos y otros muchos casos van sumando hacia una nueva concepción de la transmisión del conocimiento.

El pasado 1 de septiembre se publicó la iniciativa de la *Public Library of Science* (*PLoS*) de boicotear las revistas que no se ofrezcan gratuitamente por web a los 6 meses de haberse publicado.

Pero ¿es que realmente la editoriales obtienen grandes beneficios?

Depende de cada revista y de cada editorial. No se pueden poner todas en el mismo saco. Algunos títulos son "negocio" pero otros son semideficitarios, y las editoriales los mantienen por prestigio o por la imagen de poder ofertar todo un conjunto más o menos coherente. De igual forma ocurre con las mismas editoriales: las hay grandes, las hay abusivas y explotadoras, pero también las hay pequeñitas (hasta unipersonales), de cuyos escasos beneficios apenas pueden malvivir unos pocos...

Sin embargo está resultando difícil desplazar el sistema actual. Hace unos años se empezó a hablar de la publicación de revistas gratuitas por internet. Curiosamente, no sabemos por qué razón algunas voces, bastantes, decían entonces que las revistas "sin comité científico o de redacción" no tendrían credibilidad. Pero la verdad es que desde que existe la Red no hay ninguna razón que impida publicar electrónicamente revistas con artículos revisados por un comité científico que se reúne virtualmente.

Evidentemente más tarde aparecieron revistas-e con artículos revisados publicadas por consorcios universitarios, y todo hacía pensar que la cosa prosperaría rápidamente.

Dejando a un lado la aceptación o no del medio electrónico frente al papel, muchos pensamos que se acercaba deprisa el fin de los editores de revistas científicas. Pero hoy vemos con cierta sorpresa que no ha sido así, y que a pesar de que sigue la "espiral de revistas" —por razones varias— los editores siguen ahí con su sistema tradicional todavía muy implantado, con unos cauces establecidos que nadie se atreve a emular.

Las universidades, por ejemplo, "propietarias" de muchos de los autores y de sus artículos, no acaban de ver claro realizar la función de editar revistas y se ven obligadas a pagar por parte de sus propios artículos, simplemente porque los editores se los venden reempaquetados, con un valor añadido y una marca de calidad que si

se organizaran bien ellas también podrían aportar.

Una propuesta concreta

Con el fin de que los trabajos científicos se puedan difundir instantáneamente —actualmente hay artículos impresos que tardan más de 1 año en publicarse— y gratis por todo el mundo, un grupo de personas creó a principios de este año *The Public Library of Science*, una organización no lucrativa constituida legalmente, que piensa publicar revistas online con artículos, revisiones, estados de la técnica, ensayos, etc.

Las revistas "tendrán comités editoriales rigurosos y se editarán con una gran calidad", pretendiendo ser las mejores y más difundidas en su campo.

Los costes del comité de redacción, edición, diseño y publicación serán cubiertos básicamente por los propios autores, los cuales pagarán 300 US$ (55.000 pesetas) por cada artículo publicado. Los autores que no puedan abonar dicha cantidad podrán recibir una subvención.

El grupo promotor está buscando financiación de organizaciones no lucrativas, empresas comerciales, etc., para la puesta en marcha y los primeros años de funcionamiento.

Se han empezado a crear ya los primeros comités editoriales en biología y medicina, y *PLoS* espera comenzar a publicar a principios de 2002.

Si los autores tienen que pagar esa cantidad nosotros lo vemos un poco difícil.

http://chronicle.com/free/2001/10/2001101801t.htm

http://www.publiclibraryofscience.org/plosjournals.htm

Tomàs Baiget
baiget@sarenet.es

Recursos informativos en las redes

Papi: una propuesta de *RedIris* para el acceso ubicuo a recursos de información

Por **Rodrigo Castro-Rojo** y **Diego R. López**

DESDE LA APARICIÓN, hace ya algunas décadas, de los protocolos internet (en especial del ip) ha sido común asociar la idea de una dirección (o un conjunto de direcciones) ip con una persona o, al menos, con una determinada institución. Sin embargo, en los últimos años se han manifestado dos tendencias que hacen cada vez más difícil establecer esta asociación.

Rodrigo Castro-Rojo y Diego R. López

Por un lado, la creciente movilidad de los usuarios y la penetración de internet hacen posible que las conexiones se produzcan prácticamente desde cualquier lugar: otras redes corporativas, conexiones residenciales, ciber-cafés, etc. En cada uno de estos lugares, el usuario utiliza una dirección ip completamente diferente y perteneciente, en cada caso, a conjuntos ip heterogéneos.

A esto se suma una tendencia tecnológica que ha sido constante: la aparición de dispositivos que hacen cada vez más difícil que un servidor pueda determinar la ip real de la que proviene una petición. El uso de mecanismos de *enmascaramiento*[1] permite ocultar parte de las redes de una organización, mejorando su seguridad y optimizando los accesos al resto de internet. En el mismo sentido, el uso de caches web[6] se ha hecho muy común, en particular en la red académica de la que *RedIris* es responsable.

Como consecuencia, el acceso a los recursos web por parte de las organizaciones afiliadas a *RedIris* ha ido presentando más problemas.

Por otro lado, se han establecido cada vez más contratos de acceso a internet entre estas instituciones y las compañías proveedoras de contenidos[5, 3]. Durante la primavera del año 2000 *RedIris* organizó una reunión con representantes de diferentes organizaciones a ambos lados de estos problemas de conec-

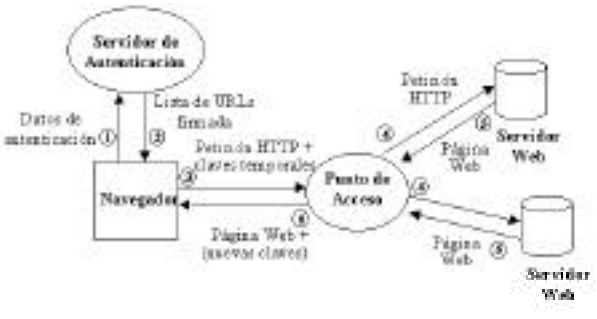


Figura 1. Arquitectura del sistema Papi, mostrando las interacciones entre el navegador del usuario y los componentes del sistema

tividad. El resultado fue una lista de requisitos para lograr una solución, así como el compromiso adquirido por *RedIris* de trabajar en un sistema de acceso que los cumpliera. Este fue el punto de partida del proyecto *Papi* (*Punto de acceso a proveedores de información*) en el que actualmente colabora *RedIris* con varias organizaciones afiliadas y con proveedores de contenidos.

2. Requisitos del sistema

Las exigencias que este nuevo modelo de acceso debía cumplir fueron compiladas por *RedIris* a partir de las peticiones de las dos clases de actores: clientes (instituciones afiliadas a *RedIris*) y proveedores (compañías de publicación). La lista de la tabla 1 es el resultado de la recopilación de las necesidades, en cierta medida contradictorias[4], de ambas partes.

Esto implica garantizar la privacidad del usuario con respecto a sus patrones de acceso. Únicamente la organización a la que pertenece puede identificarlo a partir de los datos empleados para acceder a los recursos. Por su parte, los proveedores de contenidos pueden acumular estadísticas anónimas, sin que les sea posible trazar el comportamiento individual de los usuarios.

Tabla 1

1. El sistema no debe permitir la entrada de usuarios no autorizados a ningún servicio.

2. El acceso debe ser permitido o denegado con independencia de la dirección ip que sea origen de las peticiones.

3. Los procedimientos de autenticación y de control de acceso estarán lo suficientemente delimitados como para ser gestionados por organizaciones independientes.

4. Cada usuario debe ser autenticado por un servidor operado por la organización a la que pertenece. Cada institución debe ser capaz de manejar su propio servidor de autenticación. Una vez se ha identificado correctamente debe tener acceso a todos los servicios para los que está autorizado durante un período limitado de tiempo.

5. Los mecanismos de autenticación utilizados por el servidor deben ser todo lo flexibles que sea posible, de manera que cada organización pueda usar su propio esquema de autenticación.

6. Los proveedores de información tendrán la capacidad de definir reglas de control de acceso basadas en los atributos públicos de una petición: organización a la que pertenece el usuario, tiempo de acceso solicitado, etc.

7. Debe garantizar la movilidad de los usuarios.

8. Los mecanismos implementados por el modelo han de ser totalmente compatibles con otros procedimientos de control de acceso empleados por los proveedores de información.

9. Se podrá acceder a los servidores de los proveedores utilizando los navegadores más comunes (*Netscape*, *IE*, *Lynx*) y desde cualquier sistema operativo.

10. Los procedimientos para el control de acceso serán transparentes para el usuario.

11. Cuando un usuario acceda a un recurso usando el sistema, su identidad no debe poder ser deducida de la información requerida para permitirle el acceso. Por otro lado esta información debe permitir diferenciar las entradas realizadas por usuarios distintos.

3. La arquitectura del sistema *Papi*

No es normal que se ofrezca a usuarios individuales el acceso a los servidores de los proveedores de contenidos. La práctica común es permitirlo a un cierto número de ellos dentro de una organización o, como resulta más habitual, a todos los que poseen acceso a su red. Las instituciones cliente suelen disponer de bases de datos internas y de sistemas de autenticación propios, basados en datos que no pueden ser entregados a terceros. Por ejemplo: es muy frecuente permitir el uso de cierta información solamente a personas que pertenezcan a determinados departamentos o que tengan un rango concreto dentro de la organización.

Sin embargo los proveedores de información necesitan mantener el control sobre quien accede a sus servidores, definiendo los filtros adecuados para garantizar que se cumplen los acuerdos establecidos con sus clientes. Es también una práctica común que estos filtros recolecten información para proporcionar estadísticas acerca del uso de sus datos.

Estas consideraciones han llevado a dividir el modelo en dos elementos independientes: el servidor de autenticación (*As*) y el punto de acceso (*PoA*). Esta estructura, mostrada en la figura 1, hace que el sistema de acceso sea mucho más flexible y que resulte posible integrarlo en entornos muy diversos. No es necesario establecer una correspondencia uno a uno entre *As*s y *PoA*s: un *PoA* puede manejar peticiones provenientes de una cantidad indeterminada de *As*s y dirigirlas hacia cualquier número de servidores web.

El protocolo empleado por *Papi* tiene dos fases: autenticación y control de acceso. La primera comienza en el momento en que el usuario se conecta al servidor de autenticación para obtener un conjunto de nuevas claves temporales. Durante el período de vigencia de éstas el usuario no necesita volver a contactar con dicho servidor.

Dentro de la segunda, el punto de acceso verifica las claves temporales asociadas con la información solicitada. Cada vez que el usuario trata de acceder a un sitio controlado por el *PoA*, las claves temporales son enviadas automáticamente por el navegador (sin intervención del usuario), dado que son almacenadas como *cookies* http.

1. El servidor de autenticación (*As*)**.** Su propósito es ofrecer al usuario un único punto para que se identifique, y proporcionarle (de manera completamente transparente) las claves temporales que le permitirán acceder a los servicios para los que esté autorizado. La estructura funcional del servidor incluye los siguientes elementos:

—El módulo de autenticación implementa el mecanismo que emplee la organización en la que el *As* se encuentre. El servidor está diseñado para que cada organización pueda usar su propio módulo de autenticación. La distribución actual de *Papi* incluye módulos basados en una base de datos específica tanto en servidores *pop* (*post office protocol*) como *ldap* (*lightweight directory access protocol*).

Figura 2. Un ejemplo de página de autenticación

—El módulo de gestión de sitios genera, a partir de la información aportada por el usuario y de las reglas definidas por la organización, una lista de los sitios a los que está autorizado para acceder y el tiempo durante el que puede hacerlo.

—La interfaz del servidor recibe la petición de autenticación desde el navegador del usuario y la pasa al módulo de autenticación. Si es identificado correctamente, los sitios y períodos de tiempo para este usuario son extraídos del módulo de gestión de sitios. Cada elemento de la lista se encripta usando una clave privada que representa al *As* dentro de la estructura *Papi*, y se envía de vuelta hacia el navegador del usuario como una lista de enlaces incluida dentro de una página html que contiene los resultados de la autenticación. Si el procedimiento falla se envía un mensaje de error.

Figura 3. Un ejemplo de página conteniendo la lista de sitios autorizados

Un ejemplo de página empleada por el servidor de autenticación para recoger los datos del usuario puede verse en la figura 2. En este caso se emplean los clásicos elementos "Usuario" y "Contraseña" pero el sistema puede utilizar cualquier otro (y en cualquier número). Una vez el usuario se ha autenticado correctamente, la lista de sitios autorizados es enviada a su navegador como se muestra en la figura 3.

2. El punto de acceso (*PoA*). Este elemento se encarga de controlar la entrada a un conjunto de sitios web para una determinada organización. El proveedor de información (o el propietario de los servidores web) es el responsable de gestionarlo. Los puntos de acceso *Papi* pueden ser adaptados a cualquier servidor web, con independencia de su implementación.

Es más, el acceso a un determinado servidor puede ser ofrecido a través de diferentes *PoA*s, a la vez que es posible que un *PoA* controle el acceso a más de un sitio web. Otra propiedad muy importante de este sistema es su completa compatibilidad con otros mecanismos de control, dado que no impone ninguna restricción en cuanto al número de procedimientos que puedan usarse con este propósito. En otros términos: el control de acceso basado en la tecnología *Papi* es ortogonal a procedimientos como la protección basada en passwords, filtros basados en direcciones ip, controles derivados de conexiones *TLS*, etc.

Un *PoA Papi* consta de dos clases de elementos fundamentales:

—Un módulo de generación de claves, a cargo de generar y enviar hacia el navegador del usuario las claves temporales necesarias para acceder a los recursos cuyo acceso controla el *PoA*. Son creadas de acuerdo con las peticiones enviadas por los servidores de autenticación reconocidos por el *PoA* y según las reglas locales.

—Uno o varios módulos de control de acceso. Cada vez que se recibe una petición para acceder a una parte del recurso protegido por el módulo en cuestión, éste se encarga de verificar las claves temporales que el navegador del usuario envía junto con la petición. Si las claves son correctas deja pasar la petición y sus resultados son devueltos al usuario.

4. Estado actual del sistema *Papi*

Ahora mismo existe una implementación de este modelo, disponible bajo licencia *GPL*[2], que está siendo evaluada y puesta a punto dentro de un proyecto piloto en el que participan el *Centro Informático Científico de Andalucía* (*Cica*), el *Consejo Superior de Investigaciones Científicas* (*Csic*), la *Universidad Autónoma de Madrid* (*UAM*), la *Universidad de Sevilla* y la *Universitat Oberta de Catalunya* (*UOC*) como entidades usuarias de información, junto a *Silverplatter* como proveedor de contenidos:

http://www.rediris.es/app/papi/

La tecnología *Papi* ha despertado el interés de otras redes académicas internacionales que están acometiendo proyectos similares con el objetivo de solucionar los problemas que detallábamos al comienzo de este artículo. En concreto, el equipo de desarrollo de *Papi* trabaja en estrecho contacto con los de *Sparta* (la nueva generación del sistema de autenticación centralizado *Athens*, empleado por la *UK Education & Research Networking Association, Ukerna*), en el Reino Unido, y *Shibboleth* (un proyecto de autenticación cruzada entre universidades auspiciado por *Internet2* en EUA).

Para terminar es importante resaltar que *Papi* proporciona un nuevo marco para armonizar los requisitos y necesidades tanto de los proveedores como de los consumidores de información cuando el acceso a ella se realiza en el marco de una relación contractual. La arquitectura y los protocolos que utiliza permiten superar las limitaciones que presentan las tecnologías empleadas hasta ahora para el control de acceso, garantizando tanto la independencia de los actores como la privacidad de los usuarios. Además todos los procedimientos se mantienen transparentes, de forma que su empleo no impone costes de aprendizaje y puede ser fácilmente implantado en cualquier organización.

5. Notas

1. **Danzig**, **P**.; **Swartz**, **K**. **L**. "Transparent, scalable, fail-safe web caching". *Network appliance technical report*. Consultado en: 11-09-01.
http://www.netapp.com/tech_library/3033.html

2. *Free Software Foundation.* "The GNU general public license". Consultado en: 11-09-01.
http://www.fsf.org/copyleft/gpl.html

3. **Fuchs**, **I**. "Remote authentication and authorization for *Jstor*". En: *Jstornews*, 1998, otoño, n. 2, 3.

4. **Giavarra**, **E**. "Licenses, contracts and intellectual property rights". En: *Jornadas sobre recursos electrónicos. Sociedad Española de Documentación e Información Científica (Sedic)*, 2000.

5. **Lynch**, **C**. "A white paper on authentication and access management issues in cross-organizational use of networked information resources". *Coalition for Networked Information*. Consultado en: 11-09-01.
http://www.cni.org/projects/authentication/authentication-wp.html

6. **Pearson**, **O**. "Squid (a user's guide)". Consultado en: 11-09-01.
http://squid–docs.sourceforge.net/latest/html/book1.htm

***Rodrigo Castro-Rojo** y **Diego R. López,** RedIris, Serrano 142,*
28006 Madrid.
rodrigo.castro@rediris.es
diego.lopez@rediris.es

La *Biblioteca Virtual Miguel de Cervantes*

Por **Vicente Sabido**

LA *BIBLIOTECA VIRTUAL Miguel de Cervantes Saavedra* es un ambicioso proyecto de digitalización documental de los clásicos hispánicos que, con un catálogo de más de 1.600 autores, pretende (en el plazo de tres años) poner a disposición del público de habla española un total de 30.000 obras.

Es una iniciativa de la *Universidad de Alicante* y del *Banco Santander Central Hispano,* encontrándose abierto a universidades, instituciones, empresas y particulares que deseen colaborar con objeto de ofrecer gratuitamente por internet lo más esencial de la literatura y el pensamiento hispánicos.

Así, la *Biblioteca* publica textos electrónicos clásicos, tesis doctorales, trabajos inéditos de creación e investigación e incluye numerosos enlaces a portales de bibliotecas españolas y latinoamericanas así como a recursos bibliográficos y bibliotecarios de ámbito mundial. Este proyecto tiene paralelos en otros idiomas: el *Artl* de la *Universidad de Chicago*, el *Electronic Text Center* de la *Universidad de Virginia* y el *Proyecto Gutenberg*[1], entre los más conocidos.

Además del catálogo general, que va ser objeto principal de este artículo, existe otro, de obras con copyright más riguroso, sólo accesible con login y contraseña a usuarios que disfruten del estatuto de investigador. La cuestión de los derechos de autor hace que la *Biblioteca Virtual* sólo ofrezca al público en general, en principio, obras literarias o de pensamiento anteriores a 1930 —algo similar ocurre, en el ámbito anglosajón, con el citado *Proyecto Gutenberg*— de modo que los textos

Tabla 1.

Escritores de la literatura española de la Edad de Oro, renacentista y barroca.

Baltasar del Alcázar, Mateo Alemán, Pedro Calderón de la Barca, Fray Bartolomé de las Casas, Miguel de Cervantes, Juan del Encina, Alonso de Ercilla, Vicente Espinel, Fray Luis de Granada, Baltasar Gracián, Fray Antonio de Guevara, León Hebreo, Juan Huarte de San Juan, Hurtado de Mendoza, Lope de Vega, Fray Luis de León, Padre Mariana, Arias Montano, Agustín Moreto, Francisco de Rojas Zorrilla, Fernando de Rojas, Diego de Saavedra Fajardo, San Juan de la Cruz, Santa Teresa de Jesús, Antonio de Solís, Pedro Soto de Rojas, Juan de Timoneda, Tirso de Molina, Bernardo de Valbuena, Juan de Valdés, Garcilaso de la Vega, Conde de Villamediana, Juan Luis Vives.

Tabla 2.

Autores del siglo XVIII.

—Poetas:

Nicasio Álvarez de Cienfuegos, Eugenio Gerardo Lobo, Francisco Gregorio de Salas, José Iglesias de la Casa, Tomás de Iriarte, Nicolás Fernández de Moratín, Juan Meléndez Valdés, José Antonio Porcel, Félix María de Samaniego, Cándido María Trigueros, José Vargas Ponce.

—Dramaturgos:

Luciano Comella, Vicente García de la Huerta, Leandro Fernández de Moratín, Ramón de la Cruz.

—Ensayistas y críticos:

León de Arroyal, Antonio de Capmany, Juan Pablo Forner, Gaspar Melchor de Jovellanos, Agustín Montiano y Luyando, Padre Feijoo, Padre Masdeu, Pablo de Olavide, Juan Sempere y Guarinos.

—Novelistas:

José Cadalso, Padre Isla, Diego de Torres Villarroel.

Finalmente hay que citar al periodista Francisco Mariano Nifo y al príncipe de los bibliógrafos españoles Bartolomé José Gallardo.

Tabla 3.

Autores del XIX.

—Dramaturgos:

Manuel Bretón de los Herreros, Duque de Rivas, José Echegaray, Antonio García Gutiérrez, Eugenio Hartzenbusch, Adelardo López de Ayala, Francisco Martínez de la Rosa, Ventura de la Vega, Tamayo y Baus, José Zorrilla.

—Poetas:

Gustavo Adolfo Bécquer, Padre Arolas, Manuel de Cabanyes, Rosalía de Castro, José de Espronceda, José María Gabriel y Galán, Alberto Lista, José Mor de Fuentes, Gaspar Núñez de Arce, Nicomedes Pastor Díaz, José Quintana, Salvador Rueda.

—Costumbristas románticos:

Serafín Estébanez Calderón, Mariano José de Larra, Ramón de Mesonero Romanos.

—Novelistas:

Pedro Antonio de Alarcón, Wenceslao Ayguals de Izco, Fernán Caballero, Leopoldo Alas, "Clarín", Patricio de la Escosura, Manuel Fernández y González, Gertrudis Gómez de Avellaneda, Padre Coloma, Emilia Pardo Bazán, José María de Pereda, Benito Pérez Galdós, Felipe Trigo, Juan Valera.

—Ensayistas:

Jaime Balmes, Juan Donoso Cortés, Ángel Ganivet, Marcelino Menéndez Pelayo, Manuel Milá y Fontanals.

Otras figuras culturales dignas de mención son el bibliógrafo Cayetano Alberto de la Barrera, Antonio Alcalá Galiano, José María Carnerero, Ramón López Soler, el Abate Marchena, José Joaquín de Mora, Eugenio de Ochoa y Silverio Lanza.

—Autores hispanoamericanos, entre otros:

Hilario Ascasubi, Andrés Bello, Alberto Blest Gana, Estanislao del Campo, Evaristo Carriego, Rubén Darío, Garcilaso de la Vega el Inca, Paul Groussac, Nicolás Guillén, Ricardo Güiraldes, Vicente Huidobro, Jorge Isaacs, Gabriela Mistral, Pablo Neruda, Amado Nervo, Ricardo Palma, José Enrique Rodó, Juan Ruiz de Alarcón, Arturo Uslar Pietri.

tos electrónicos pertenecen casi siempre a escritores situados, aproximadamente, entre los cronistas de Indias y el Modernismo.

Hay pocos autores medievales españoles representados en la *Biblioteca Cervantes Virtual*. Tal vez la causa sea el poco interés que para el gran público tienen estas figuras. De la gran literatura española de la Edad de Oro, renacentista y barroca, la biblioteca alberga entre otros a los escritores que vemos en la tabla 1.

Del siglo XVIII hay una muy interesante representación, por lo difícil que resulta a veces encontrar (me lo dicta mi experiencia docente) ediciones de ese siglo, reservado hasta hace poco a expertos. Entre otros, obras de los autores indicados en la tabla 2; por último comentar que el siglo XIX es el más nutrido (tabla 3).

Hay que decir que los textos digitalizados incluyen en ocasiones un facsímil de la portada de la edición original, lo que nos permite conocer lugar, editor y año de publicación. Además, podemos indicar que en las obras se moderniza sistemáticamente la ortografía y la puntuación.

Otros proyectos similares:

Bimicesa utiliza los estándares más difundidos: html, jpg, gif, wav, avi, etc., y por el momento no ofrece los textos completos en formato comprimido zip para ser descargados por ftp, pero no descarta la posibilidad de hacerlo en el futuro.

1. *Artfl*, *The Project for American and French Research on the Treasury of the French Language*, es una empresa cooperativa de *Analyses et Traitements Informatiques du Lexique Français*, del *Cnrs* y del *Electronic Text Services* de la *Universidad de Chicago*.

http://humanities.uchicago.edu/orgs/ARTFL/

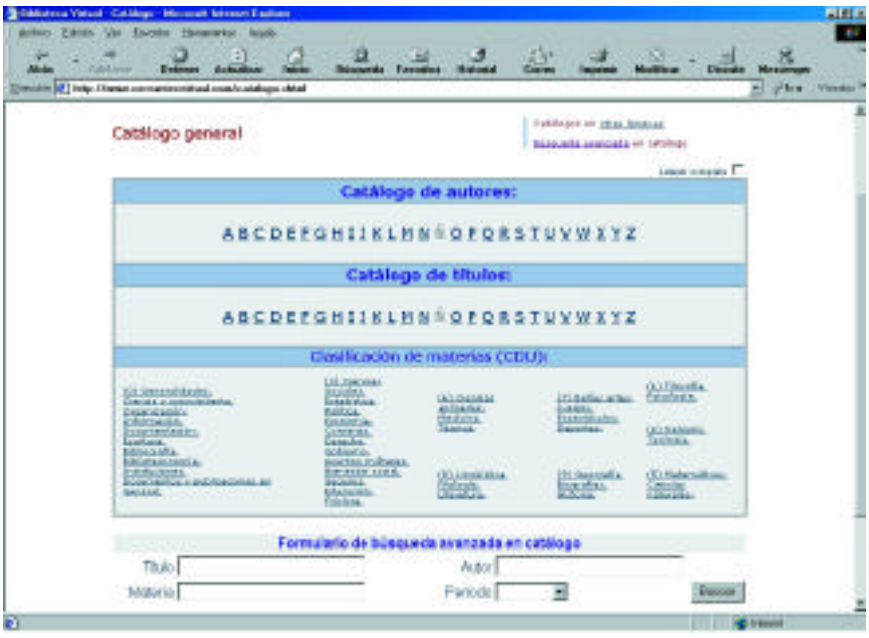
2. El *Electronic Text Center* de la *Universidad de Virginia* combina la consulta en línea de miles de archivos (unos 45.000 en inglés, francés, alemán, latín, chino, japonés y otras lenguas) codificados en sgml y xml con un servicio bibliotecario que ofrece hardware y software adecuados a la creación y análisis de textos.

http://etext.lib.virginia.edu/

3. El *Proyecto Gutenberg*, dirigido por Michael Hart, ha puesto ya en el dominio público más de 2.000 textos de la literatura universal, especialmente la de habla inglesa. Sale adelante gracias al entusiasmo de Hart y de cientos de voluntarios en todo el mundo que trabajan de forma desinteresada en esta biblioteca electrónica gratuita.

http://promo.net/pg/

Al estar compuestas en formato html, el lector no pierde todos esos caracteres españoles (eñes, vocales acentuadas o con diéresis, etc.) que resultan fundamentales para la cabal comprensión del texto. Por otra parte, la lectura puede efectuarse en línea, cosa a veces dificultosa y poco útil, o bajarse el texto completo para leerlo posteriormente offline. También puede emplearse algún programa para capturar automáticamente la obra en su totalidad, ya que con frecuencia cada capítulo es, materialmente, una página web distinta. Soy del parecer que, a excepción de textos muy breves, la lectura en pantalla se hace muy penosa, siendo preferible imprimir la obra para leerla al modo tradicional. Contra los que piensan que a veces es mejor imprimir un libro electrónico a comprarlo, diré que el gasto de tiempo y de material suele al menos encarecer al doble el precio del libro (siempre, claro, que éste pueda adquirirse con facilidad).



Catálogo general

Clásicos universales

Por disponer *Bimicesa* de traducciones a lenguas españolas de **Alfieri**, **Apuleyo**, **Aristóteles**, **Balzac**, **Baudelaire**, **Bossuet**, **Byron**, **Camoens**, **Chejov**, **Condillac**, **Dante**, **Descartes**, **Dickens**, **Diderot**, **Dumas**, **Erasmo**, **Fénelon**, **Goethe**, **Goldoni**, **Hume**, **Kant** y **Voltaire**, éstos figuran en

el catálogo principal. Hay además nutridos catálogos de obras en catalán, gallego, inglés, portugués, alemán, francés e italiano, gracias a los enlaces con otras bibliotecas electrónicas españolas y extranjeras.

Un ejemplo lo encontramos con la entrada **Shakespeare, William** que permite localizar la obra completa del autor en su lengua original, incluso algunas piezas dramáticas traducidas al español, tal como *A buen fin no hay mal principio*, en la versión ya clásica de **Luis Astrana Marín**. Con acertado criterio la *Biblioteca* ha elegido la traducción del *Hamlet* que publicó en 1798 **Leandro Fernández de Moratín**, que intenta salvar las numerosas escabrosidades de la obra y ofrecer un texto adecuado al gusto ilustrado del escritor madrileño y sus lectores.

Por supuesto podemos consultar los originales ingleses. Por ejemplo, *As you like it*, al cual se accede mediante un vínculo específico. El problema es que el enlace ha sido cambiado en el host y descubrimos, tras la correspondiente inquisición, que debería ser actualizado al que vemos a continuación en segundo lugar.

Ejemplo de una lista de autores

www.tech.mit.edu/Shakespeare/Comedy/asyoulikeit/asyoulikeit.html

www.tech.mit.edu/Shakespeare/asyoulikeit/index.html

Esta labor de reparar enlaces rotos no siempre puede ser realizada de manera automática, por lo que a veces es posible encontrarse con estos errores.

En la misma línea, en la entrada **Shaw, Bernard,** encontramos un vínculo roto a *Man and Superman*; otro, operativo, al libro *Misalliance*, ofrecido en este caso por el *Proyecto Gutenberg*, en ascii simple (7 bits de cifrado), formato suficiente para textos en inglés sin eñes, tildes ni diéresis; otros dos a *An unsocial socialist* y a *The perfect wagnerite*, también del *Gutenberg*. Un nuevo vínculo nos conduce a *Pygmalion*, éste del *Proyecto Bartleby*, de la *Universidad de Columbia* en Nueva York; otro, a una versión en catalán de la misma obra (en html) y a una breve secuencia de vídeo de un momento de la representación de este *Pigmalió* por *Dagoll Dagom* (*Grup de teatre*). Es posible visualizarlo también a través de dos versiones: una para módems (de menos resolución en pantalla) y otra para conexiones más rápidas de tipo *Adsl*. Como puede verse, las posibilidades son variadas.

Lista de títulos por materia

Otras secciones de la *Biblioteca*, también en desarrollo, son:

1. Biblioteca de autor: en ella encontramos, además de clásicos imprescindibles como **Leopoldo Alas**, **Fray Bartolomé de las Casas**, **Miguel de Cervantes Saavedra**, **Garcilaso de la Vega**, **Garcilaso de la Vega**, **El Inca**, **Luis de Góngora y Argote**, **San Juan de la Cruz**, **Sor Juana Inés de la Cruz**, **Fray Luis de León**, **Francisco López de Gómara**, **José Martí**, **Tirso de Molina**, **Francis-**

co de Quevedo, **Santa Teresa de Jesús**, **Juan Valera y Lope de Vega**, a un grupo de escritores en su mayoría hispanoamericanos del siglo XX como **Roberto Arlt**, **Alfredo Bryce Echenique**, **Mario Benedetti**, **Dulce María Eloinaz**, **Oliverio Girondo**, **Ángel González**, **Nicolás Guillén**, **Nancy Morejón**, **Gonzalo Rojas**, **Alfonsina Storni** y **Arturo Uslar Pietri**.

Podemos conocer sus biografías, así como una selección de sus obras a veces recitadas por ellos mismos lo que, añadido a fotografías y vídeos, hacen de esta sección una auténtica galería multimedia. Las fotos se visualizan en el navegador de internet que utilicemos, pero el sonido y el vídeo requieren del *Reproductor de Windows Media* que puede descargarse gratuitamente desde un vínculo de la *Biblioteca*.

2. Hemeroteca: incluye revistas culturales y científicas como por ejemplo: *Hispania*, publicación oficial de la *Association of Teachers of Spanish and Portuguese of America*, encontrándose accesibles en formato html los cuatro números que aparecieron en 1995. Es de suponer que la disponibilidad de éste y otros muchos títulos irá creciendo con el tiempo, pues lo que existe es a veces algo escaso. La dificultad para acceder a muchas publicaciones periódicas de ámbito hispánico hace muy interesante la oferta de *Bimicesa*. Hay, no obstante, revistas mucho más completas como por ejemplo los *Anales galdosianos*.

3. Tesis doctorales: proceden de muchos departamentos universitarios de habla española y actualmente esta sección se encuentra en desarrollo. Las obras, difíciles a veces de publicar en volumen, se ofrecen en formato pdf. Como medida complementaria de apoyo a su lectura, el visualizador necesario para este formato está disponible para descarga desde un enlace de esta sección.

4. Miscelánea: incluye una serie de obras correspondientes a estudios críticos, actas de congresos sobre temas hispánicos, homenajes a hispanistas y otras obras colectivas. También en formato html.

No acaban ahí las posibilidades. Además de múltiples enlaces a portales de interés, como por ejemplo a bibliotecas de todo el mundo, a proyectos de edición electrónica y a instituciones relacionadas con la lengua y la cultura en español, la *Cervantes Virtual* ofrece una biblioteca de la historia; otra de imágenes; de signos (textos expresados mediante signos corporales para sordomudos a través de vídeos); de voces (textos recitados por rapsodas, asimismo en vídeo), etc.

En resumen, un proyecto (que ya es en parte realidad) de acercamiento a los textos clásicos hispánicos que con frecuencia puede resultar imprescindible para estudiosos y para el público interesado en la materia, pues la *Cervantes Virtual* ofrece muchos libros que, por su antigüedad o su rareza, son difíciles de encontrar incluso en bibliotecas públicas o universitarias. Materialmente, la página del portal está bien confeccionada y sus menús conducen con facilidad a todos los apartados y opciones. Como contrapunto comentar la disponibilidad de un servicio semanal gratuito de novedades incorporadas a *Bimicesa* que sólo requiere de una dirección de correo electrónico para recibir periódicamente las nuevas incorporaciones.

http://www.cervantesvirtual.com

Notas

1. **Sabido**, **Vicente**. "*El* Proyecto Gutenberg". En: *Information world en español*, 1996, octubre, n. 48, pp. 16-18.

Vicente Sabido. *Departamento de Filología Española, Universidad de Granada.*
vsabido@ugr.es

SOCIEDAD ESPAÑOLA
DE DOCUMENTACIÓN
E INFORMACIÓN CIENTÍFICA

Sociedad Española de Documentación e Información Científica

Creada en 1975, con más de 1300 asociados en toda España.

Es un foro y lugar de encuentro para los profesionales de la gestión de la información y el conocimiento, que ofrece:

- Cursos y seminarios de formación.
- Grupos de trabajo.
- Organización de Jornadas y Conferencias.
- Bolsa de trabajo.
- Representación nacional e internacional.
- Acreditación profesional.
- Clip -Boletín trimestral de la Asociación.
- Revista Española de Información Científica.

Si te preocupa la mejora y el reconocimiento social de tu profesión y piensas que se puede hacer algo, asóciate a SEDIC y colabora en nuestras actividades.

Sede: C/ Santa Engracia, 17, 3º
28010 Madrid
Tel: 91 593 40 59
Fax: 91 593 41 28
e-mail: sedic@sarenet.es
http://www.sedic.es

Publicidad
Dialog

color

mismo que octubre, p. 19

Publicidad
Zylab

Publicidad
Academic Press

b/n

fotolito nuevo

Publicidad
Swets

b/n

mismo que octubre, p. 43

BOLETÍN DE SUSCRIPCIÓN

El Profesional de la información

TARIFAS 2001

Para empresas, organismos o instituciones:

☐ Deseamos suscribirnos a EL PROFESIONAL DE LA INFORMACIÓN durante un año al precio de 17.000 pesetas +4% IVA / Eur 103 / Dfl 226 / US$ 186*. A los pagos realizados desde países distintos de España se les aplicará el IVA o VAT local.

Para suscriptores individuales (no escribir ningún nombre de institución; indicar sólo la dirección particular):

☐ Deseo suscribirme a EL PROFESIONAL DE LA INFORMACIÓN durante un año al precio especial de 9.990 pesetas +4% IVA / Eur 60 / Dfl 137 / US$ 135. A los pagos realizados desde países distintos de España se les aplicará el IVA o VAT local.

* El valor expresado en dólares incluye el correo aéreo a América.

❑ Carguen el importe a mi tarjeta ❑ MASTER CARD ❑ VISA ❑ AMERICAN EXPRESS ❑ DINERS

nº..... Caducidad: .....

❑ Se adjunta a este boletín un cheque nominal (en florines holandeses) a nombre de **Swets & Zeitlinger** .

❑ Pago por transferencia en pesetas a la cuenta nº 0049 4700 33 2511824235 a nombre de **Swets & Zeitlinger** , del Banco Santander Central Hispano, Pº de Gracia, 5. 08007 Barcelona, Tel.: 934 011 100. Adjunto fotocopia del resguardo.

❑ Por favor, envíenme una factura, especificando el número de pedido.

POR FAVOR, ESCRIBA EN MAYÚSCULAS

Nombre:.....

Institución:.....

Dpto:..... NIF institucional:.....

Dirección:.....

Código postal:..... Ciudad: ..... País: .....

Teléfono: ..... Fax: .....

Correo-e: .....

Envíe este boletín a:

Swets & Zeitlinger
PO Box 825
2160 SZ Lisse (Holanda)
orders@swets.nl

Información para los autores

Todos los profesionales que lo deseen pueden remitir a la redacción de la revista **El profesional de la información** sus colaboraciones en forma de:

- notas breves

- trabajos más amplios sobre temas de fondo para la sección "Artículos".

El texto ha de enviarse en formato electrónico. Aparte, los materiales gráficos, en papel o en ficheros gif, jpeg o tiff con unos anchos de entre 12 y 5,7 cm y una resolución de 300 ppp.

El tamaño ideal de un estudio para la sección "Artículos" es de 5.000 palabras (si el trabajo incluye materiales gráficos, el número de palabras ha de ser menor). En casos excepcionales pueden publicarse artículos de mayor extensión. Los trabajos de esta sección son aprobados según el sistema tradicional "peer-review": al menos dos expertos en el tema, del *Consejo Asesor* de la revista y/o externos, deben dar el visto bueno antes de su publicación.

Los textos deben enviarse sin formatos especiales (títulos, secciones, subsecciones, pies de página, sangrías, tabulaciones, colores, etc.).

Los trabajos para la sección "Artículos" deben incluir: a) título en castellano, b) resumen en castellano de 100-150 palabras, c) 5-10 palabras clave en castellano, d) título en inglés, e) resumen en inglés de 100-150 palabras, f) 5-10 palabras clave en inglés, g) texto completo en castellano y h) nombre de los autores, lugar de trabajo y dirección de correo electrónico. **Han de ser inéditos**.

Se valorará especialmente que los trabajos sean concisos y precisos. Se ruega a los autores que eviten una excesiva retórica.

Las referencias bibliográficas han de prepararse de acuerdo con el siguiente esquema:

Artículos de una publicación periódica:

Apellido, **Nombre**; **Apellido2**, **Nombre2**. "Título del artículo". En: *Título de la publicación periódica*, año, mes, v. [número del volumen], n. [número del ejemplar], pp. [página comienzo]-[página final].

Ponencia presentada en un congreso:

Apellido, **Nombre**; **Apellido2**, **Nombre2**. "Título de la ponencia". En: *Nombre del congreso*, año, pp. [página comienzo]-[página final].

Monografías:

Apellido, **Nombre**; **Apellido2**, **Nombre2**. *Título del trabajo*. Lugar de publicación: editor, fecha. Isbn [número].

Capítulo de una monografía:

Apellido, **Nombre**; **Apellido2**, **Nombre2**. "Título del capítulo". En: **Apellido**, **Nombre**; **Apellido2**, **Nombre2**. *Título de la monografía*. Lugar de publicación: editor, fecha, pp. [página comienzo capítulo]-[página final capítulo]. Isbn [número].

Recurso en línea:

Apellido, **Nombre**; **Apellido2**, **Nombre2**. Título del recurso. Consultado en: día-mes-año.
dirección

Las contribuciones se pueden enviar a la redacción de la revista o a cualquiera de los miembros del consejo de redacción.

El hecho de que un trabajo sea publicado en EPI no implica que la redacción se adhiera a las opiniones expresadas en él.

Redacción EPI
Apartado 32.280
08080 Barcelona

epi@sarenet.es

La redacción se reserva el derecho de adaptar los textos al estilo gramatical y literario de la revista.