EL PROYECTO EUROPEO MEDIEQ (QUALITY LABELLING OF MEDICAL WEB CONTENT USING MULTILINGUAL INFORMATION EXTRACTION): LA WEB SEMANTICA AL SERVICIO DE LOS USUARIOS DE SALUD

THE EUROPEN PROJECT MEDIEQ (QUALITY LABELLING OF MEDICAL WEB CONTENT USING MULTILINGUAL INFORMATION EXTRACTION): WEB SEMANTIC AT HEALTH USERS SERVICE

Mayer Pujadas, Miquel Ángel. Director de Web Mèdica Acreditada. Colegio Oficial de Médicos de Barcelona, Passeig de la Bosanova 47 08017 Barcelona, mmayer.wma@comb.es; **Leis Machín, Angela**. Directora Adjunta. Colegio Oficial de Médicos de Barcelona, mleis.wma@comb.es; **Karkaletsis, Vangelis**. Coordinador de MedIEQ. National Center for Scientific Research NCSR "Demokritos", Grecia, vangelis@iit.demokritos.gr; **Villarroel, Dagmar**. Ärztliches Zentrum für Qualität in der Medizin, Alemania, Villarroelgonzales@azq.de.

Resumen: La calidad de la información médica en Internet es muy variable y posee un gran potencial para beneficiar o para dañar a un gran número de personase y es por ello que se hace necesario proveer a los usuarios de salud de herramientas para discernir la información correcta de aquella que no lo es. Se han propuesto diversas soluciones como los sistemas de acreditación y los portales temáticos que filtran las webs de mayor calidad. Estas herramientas deben incorporar, para conseguir una máxima eficiencia y un uso más amplio por los navegantes en Internet, lenguajes estandarizados como los propuestos por la W3C, que se basan en metadatos, así como sistemas de extracción y análisis automatizados de contenidos que los estructuran de una forma eficiente. En este trabajo se describe el proyecto europeo MedIEQ como un sistema que integra de forma práctica tecnologías de metadatos y extracción automatizada de información en el área de la información sanitaria en Internet.

Palabras clave: calidad, información sanitaria, web semántica, Internet, extracción información, estándares.

Abstract. Quality of Internet health information is essential because it has the potential to benefit or harm a large number of people and it is therefore essential to provide consumers with some tools to aid them in assessing the nature of the information they are accessing and how they should use it without jeopardizing their relationship with their doctor. Organizations around the world are working on establishing standards of quality in the accreditation of health-related web content. For the full success of these initiatives, they must be equipped with technologies that enable the automation of the rating process and allow the continuous monitoring of labeled web sites alerting the labeling agency. In this paper we describe the European project MedIEQ that integrates the efforts of relevant organizations on medical quality labelling, multilingual information retrieval and extraction and semantic resources on the Internet in the medical information field.

Keywords: quality, health information, semantic web, Internet, information extraction, standards.

1. Introducción

La información médica existente en la Red es cada vez más amplia y crece de forma exponencial día a día. Esta información presenta contenidos y calidad muy variables, desde contenidos científicos claramente contrastados hasta aquella información que puede ser engañosa o peligrosa para la salud si se utiliza inadecuadamente por los pacientes y sus familiares. Se han propuesto diferentes soluciones para garantizar que dicha información presenta unas garantías mínimas de confianza entre las que destacan las guías de recomendaciones y los sellos de calidad. Desde la Unión Europea y desde diversas organizaciones, se conceden sellos de confianza a aquellas webs que se han sometido a procesos de revisión para una mejora de sus contenidos o se crean portales que realizan un filtrado previo de los recursos presentes en Internet cumpliendo con una serie de criterios de calidad establecidos previamente. Estos recursos se orientados tanto para al público en general como a los profesionales. (1-3)

Por otro lado, es necesario desarrollar y aplicar herramientas que optimicen el trabajo de las agencias de calidad que realizan la revisión y descripción de webs médicas acreditadas, ofreciendo a los usuarios información clara y comprensible sobre las características de dichas webs. Se hace imprescindible integrar y estandarizar herramientas como los lenguajes de metatados y análisis de contenidos mediante la utilización de vocabularios y clasificaciones para su descripción como Dublin Core, (4) Friend of a Friend (FOAF), (5) HIDDEL (6) y otros estándares desarrollados por el World Wide Web Consortium (W3C) (7) así como tesauros del ámbito médico como el Medical Subject Headings (MeSH) o el Unified Medical Language System (UMLS) de la National Library of Medicine (8) y que se están aplicando en otros campos del conocimiento.

El proyecto europeo MedIEQ (Quality Labeling of Medical Web Content using Multilingual Information Extraction) (9) desarrolla y amplía el trabajo realizado en anteriores proyectos europeos en el campo de la e-Salud y la aplicación de metadatos como: MedCERTAIN, MedCIRCLE, (10) WRAPIN y QUATRO, (11) centrándose en temas de calidad de webs médicas y mostrando el estado actual en la aplicación de tecnologías de rastreo y análisis de contenidos web y extracción multilingüe de la información, y aprovechando la utilización de recursos semánticos y sellos de calidad de dichas webs. Todo esto permitirá mejorar la monitorización de las webs médicas acreditadas así como su identificación y clasificación en áreas temáticas, basándose en ambos casos en metadatos y utilizando siete idiomas diferentes (checo, griego, español, inglés, alemán, finlandés, catalán). Estos metadatos están expresados mediante el estándar RDF/XML (Resource Description Framework) (12) lo que permite la integración con herramientas como los motores de búsqueda, que de esta forma serán capaces de "entenderse" con los usuarios al utilizar palabras clave con contenido semántico en el proceso de búsqueda y recuperación de esta información. Además se integran tecnologías de extracción automatizada de contenidos que permitan la simplificación de tareas de revisión y control así como la creación de nuevos recursos de información relacionados. En la primera parte del artículo se presenta el escenario de aplicación de las diferentes herramientas descritas, es decir, en este caso un sistema de acreditación de webs médicas, Web Médica Acreditada. Posteriormente se describirá en qué consiste los lenguajes de metadatos en los que se basan las anotaciones y descripciones que utiliza el programa de acreditación y finalmente los sistemas de extracción de información así como la integración de todas estas herramientas y que caracteriza al proyecto europeo MedIEQ.

2. Elementos y herramientas integradas en la propuesta de MedIEQ

2.1 Web Médica Acreditada: programa de evaluación de contenidos sanitarios en Internet

Este programa de acreditación se inició en 1999 y fue creado por el Colegio Oficial de Médicos de Barcelona con el objetivo de orientar en el buen uso de los servicios e información de webs de contenido médico. El proceso de acreditación de WMA incluye un Comité Permanente y una Comisión Delegada que decide la acreditación en función de la adaptación a las recomendaciones de WMA. El equipo que trabaja en WMA es multidisciplinar y está formado por médicos, abogados, comité deontológico, informáticos y diseñadores web. Se basa en la aplicación del Código de Conducta creado por WMA a través de la revisión activa de las webs que se incluyen en el programa de acreditación. Una vez acreditada la web se concede un sello de acreditación (un código HTML) que certifica esta acreditación y que contiene información sobre la misma. El código de conducta contiene los siguientes criterios: (13, 14)

- Identificación: autoría, institución y responsables de la web
- Contenidos: actualización y fuentes de información de los contenidos
- Confidencialidad: las medidas de confidencialidad seguidas por la web y los datos de los usuarios
- Control y validación: utilización de forma adecuada del sello de calidad concedido
- Publicidad y fuentes de información
- Consulta virtual (Documento de la Comisión Deontológico)
- Incumplimiento y responsabilidades: detección de problemas en los servicios ofrecidos por la web

Desde Web Médica Acreditada se revisan los contenidos de la web y se estudia su adaptación a las recomendaciones de calidad. Se realiza un informe que se envía al responsable de la web para que, si es el caso, se realicen las adaptaciones correspondientes y poder así completar el proceso de acreditación. Una vez completado el proceso de acreditación se envía un código HTML para que aparezca el sello de acreditación en la web. A este sello se le asocia un archivo en formato XML/RDF que describe las características básicas de dicha web. Posteriormente se realiza una revisión anual de la web acreditada. Dicha información en RDF queda almacenada en la base de datos de WMA. En la figura 1 se presenta el interfaz de trabajo para la gestión de todos los datos referentes al proceso de acreditación de las webs que se hallan incluidas en el proceso de acreditación. En el menú lateral aparece el apartado RDF de este gestor es el generador automático del archivo XML/RDF, este archivo quedará situado en la base de datos del sistema de acreditación y conectado a su vez con el repositorio (base de datos) de MedIEQ para su posterior revisión.

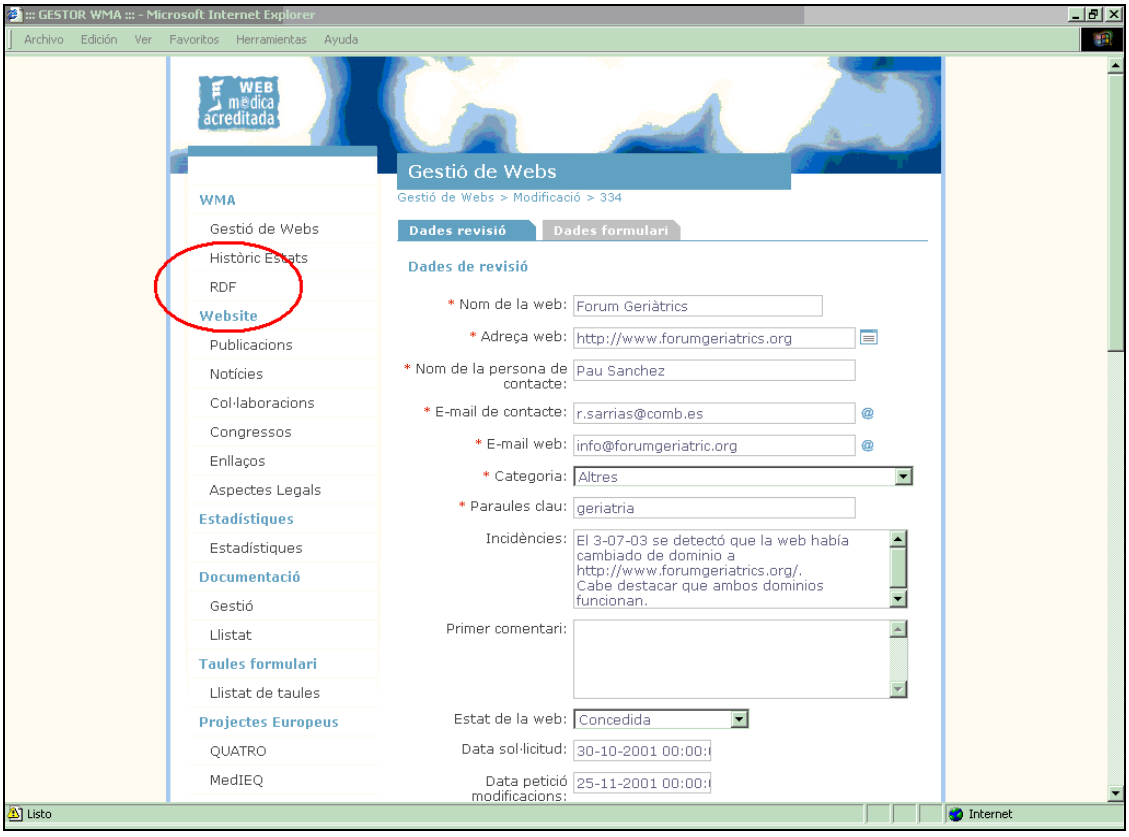

Figura 1. Interfaz de trabajo de Web Mèdica Acreditada para la gestión de webs acreditadas y generación del archivo RDF (señalado en rojo) que se depositará en la base de datos de este sistema de evaluación y en el repositorio de MedIEQ.

Podemos observar en las siguientes líneas de código el aspecto que presenta el archivo RDF generado de forma automática desde la Intranet de WMA de la web Forumgeriatrics.org (código interno 334) y que es descrita con metadatos:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:label="http://www.w3.org/2004/12/q/contentlabel#" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:wma="http://wma.comb.es/rdf/vocabularyv01#"
xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:quatro="http://purl.org/quatro/elements/1.0/"
xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#" xmlns:wn="http://xmlns.com/wordnet/1.6/"
<rdf:Description rdf:about="">
<dc:Title>Web Médica Acreditada</dc:Title>
<dc:description xml:lang="en">This document defines the WMA seal for one or more web
sites.</dc:description>
<dc:creator>
<label:Ruleset rdf:ID="Ruleset">
        <label:hasHostRestrictions>
                <label:Hosts>
                        <label:hostRestriction>forumgeriatrics.org</label:hostRestriction>
                </label:Hosts>
        </label:hasHostRestrictions>
<label:hasDefaultLabel rdf:resource="#label_1" />
        <dcterms:issued>2004-03-02</dcterms:issued>
        <quatro:validUntil>2006-12-02</quatro:validUntil>
</label:Ruleset>
        <label:ContentLabel rdf:ID="label_1">
                <foaf:homepage>http://www.forumgeriatrics.org/</foaf:homepage>
                <wma:wmacode>334</wma:wmacode>
                <dcterms:dateAccepted>2001-11-25</dcterms:dateAccepted>
                <quatro:dateReAccepted>2005-06-14</quatro:dateReAccepted>
                <wma:emailok>1</wma:emailok>
                <wma:emailcontact>r.sarrias@comb.es</wma:emailcontact>
                <wma:healthprof>1</wma:healthprof>
                <quatro:gk rdf:resource="quatro:WAI-A" />
                <label:hasClassification rdf:resource="http://wma.comb.es/rdf/vocabularyv01#ss" />
                <wma:webname>Forum Geriàtrics</wma:webname>
                <wma:generalupdate>1</wma:generalupdate>
                <wma:authors>1</wma:authors>
                <wma:healthsource>1</wma:healthsource>
                <wma:intlinks>1</wma:intlinks>
                <wma:extlinks>1</wma:extlinks>
                <label:hasClassification rdf:resource="http://wma.comb.es/rdf/vocabularyv01#fam" />
                <dc:audience xml:lang="en">Physician/GP</dc:audience>
                <quatro:gf>ES</quatro:gf>
                <dc:MESH>Geriatrics Assesment</dc:MESH>
        </label:ContentLabel>
 </rdf:RDF>
```

2.2 *Web semántica y lenguajes de metadatos: Dublin Core, HIDDEL, FOAF*

Debemos entender la web semántica como una extensión del concepto actual de web, basada en diferentes lenguajes de metadatos y que permiten una mayor estructuración de la información, elaborando relaciones entre los recursos y los contenidos con la finalidad de mejorar la interoperabilidad entre personas y máquinas. La web semántica aplicada a las iniciativas que están realizando la revisión de los contenidos y la descripción de las características de

las webs de contenido sanitario, puede constituir una interesante aportación que dote de un mejor conocimiento a los usuarios sobre el tipo de información a la que están accediendo; permitiendo además que esta información pueda ser utilizada por motores de búsqueda "que entenderán" mejor lo que los usuarios realmente están buscando y obtendrán una información más elaborada, descriptiva y detallada del contenido de las webs objeto de búsqueda. Las aplicaciones de la web semántica son diversas como FOAF (Friend of a Friend) que se utiliza para la descripción de personas y organizaciones, el RSS (RDF Site Summary) que se aplica en las comunidades de noticias diversas utilizando lectores específicos. En el campo sanitario, se han utilizado lenguajes específicos como HIDDEL (Health Information Disclosure, Description and Evaluation Language) con diferente éxito en su aceptación y distribución.

Actualmente el proyecto MedIEQ está desarrollando un estándar en lenguaje RDF (Resource Description Framework) basado en la experiencia de Web Médica Acreditada y en las recomendaciones de otras organizaciones de acreditación y calidad de referencia como Health on the Net Foundation (14) y la guía elaborada por la Unión Europea, e-Europa 2002: Criterios de calidad para sitios web relacionados con la salud.

2.3 *Sistemas de extracción multilingüe*

Los sistemas de extracción de datos en las webs han tenido un gran desarrollo en las aplicaciones relacionadas con los motores de búsqueda en Internet. La optimización de resultados y los criterios de búsqueda utilizados se constituyen como los elementos básicos para la obtención de la información acorde con las necesidades de cada búsqueda. (15) Actualmente los motores de búsqueda como Google se basan en términos no semánticos y carentes de significado. En el proyecto MedIEQ los sistemas de extracción de información en diferentes idiomas, se basarán en descriptores estandarizados y conceptos semánticos del RDF schema que se propone. Al compartir definiciones comunes (RDF schema de metadatos) entre humanos y máquinas (buscadores) se puede garantizar una coincidencia entre los conceptos humanos y los términos de significado semántico con un entendimiento real entre ellos mejorando el resultado de las búsquedas y descripciones de los contenidos web.

2.4 *La propuesta de MedIEQ*

En el caso que nos ocupa, Web Médica Acreditada, la aplicación de la herramientas de la plataforma permite la monitorización continuada de las webs acreditadas, para ello a través de los sistemas de extracción de información puede seleccionar y comparar la información extraída de estas webs acreditadas con la información que se encuentra en la base de datos de este sistema de acreditación para detectar diferencias que indiquen a los responsables del sistema que deben revisarse de nuevo. Teniendo en cuenta el proceso de acreditación de WMA el sistema actúa de la siguiente forma (ver figura 2):

- Cada vez que se recibe una nueva solicitud de acreditación para WMA, el sistema es llamado para realizar una primera recolección de datos significativos de la web solicitante. El tipo de datos puede variar dependiendo de la web y son almacenados en una base de datos específica para realizar una primera valoración por el comité de revisión de WMA.
- Una vez el responsable de la web realiza los cambios sugeridos por los revisores de WMA, el sistema automático comprueba que los cambios han sido realizados. El informe de esta revisión vuelve a ser almacenada en una base de datos en formato RDF utilizando los campos o descriptores escogidos y para su posterior comprobación por el comité revisor de WMA, que decide si esta web puede obtener el sello de acreditación o no.
- Una vez la web obtiene el sello de WMA, el sistema de extracción será llamado periódicamente para examinar si los cambios que presenta, en base a los criterios de acreditación. Dependiendo de los cambios detectados el sistema alerta a WMA, facilitando de esta forma el proceso de revisión.
- Una vez generado el archivo RDF debe quedar en el repositorio de datos de los archivos RDF de MedIEQ para su posterior reevaluación o en las bases de datos de los sistemas de acreditación que deben estar conectados con este repositorio (MedIEQ Database) que debe actualizarse periódicamente. Esto permitirá la comparación entre el contenido que se encuentre en la base de datos y los datos obtenidos en el momento

de la revisión por el sistema de extracción múltiple. Si existen diferencias se notificarán al sistema de evaluación pertinente.

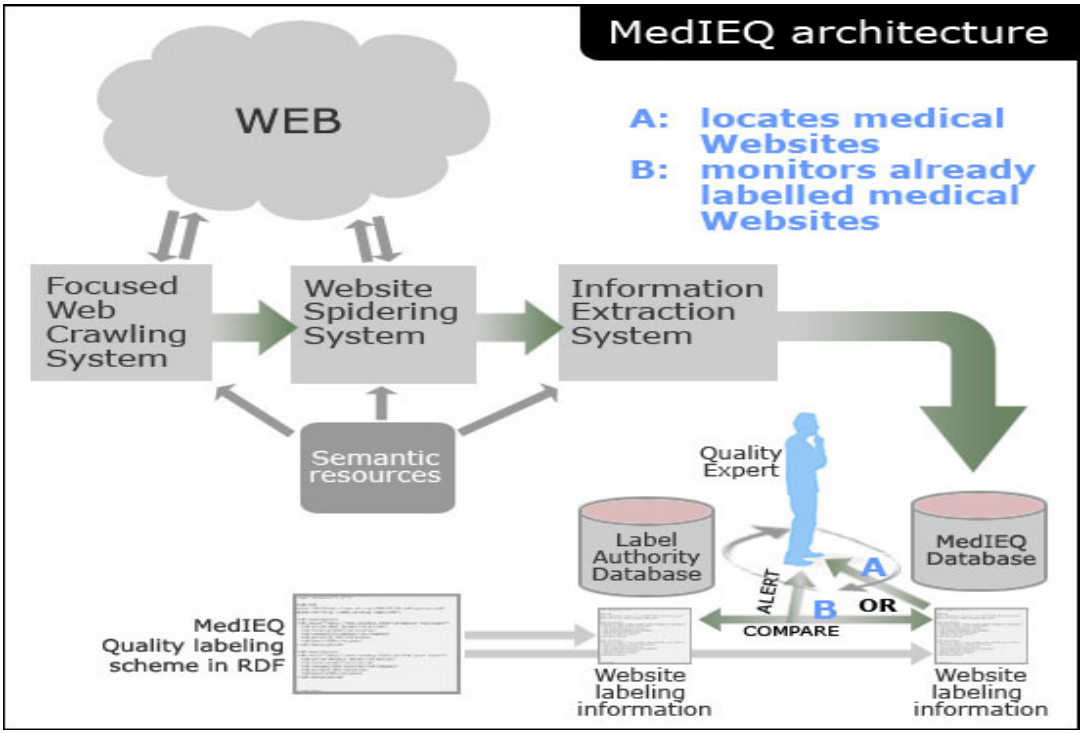

Figura 2. Esquema general de trabajo de MedIEQ utilizando recursos de metadatos y sistemas de extracción de datos web.

3. Conclusiones

El número de webs de contenido sanitario continúa creciendo así como el interés que despierta entre el público en general y los pacientes. Ante este crecimiento es necesario aplicar mecanismos de control como los sellos de calidad y sistemas de acreditación o los portales que actúan como filtros. Pero esto no es suficiente, ya que actualmente disponemos de herramientas que pueden permitir optimizar la búsqueda de esta información así como estandarizar, para su mejor comprensión, un lenguaje de metadatos asociado a estos recursos que los describa adecuadamente. Uno de los problemas principales de los sistemas de acreditación es la monitorización de estas webs acreditadas ya que requiere un gran esfuerzo. La aplicación de sistemas de extracción de información en estas webs y su asociación con estos lenguajes de descripción estandarizados pueden mejorar estas tareas de mantenimiento y contribuir a la creación de portales de información sanitaria que agrupen estos recursos de forma automática.

MedIEQ ofrecerá un esquema de trabajo que basado en:

- la utilización de lenguajes estandarizados para la descripción de contenidos de webs médicas que permitan una interoperabilidad inexistente actualmente basados en metadatos XML/RDF,
- la aplicación de sistemas de extracción automática de contenidos al servicio de diferentes plataformas de trabajo como las que caracterizan a los sistemas de acreditación y revisión de webs médicas para su selección, en vías de garantizar el cumplimiento de un mínimo de criterios de calidad, informando de esta manera a los sistemas de acreditación sobre la necesidad de revisión o actualización de la descripción.
- la facilitación de que los usuarios obtengan una mayor y mejor orientación sobre las características de la información sanitaria en Internet y contribuir a su educación sanitaria ya que se utilizan definiciones estandarizadas,

- la mejor visibilidad en los motores de búsqueda utilizando tecnologías de metadatos que permitan la estandarización de las terminologías utilizadas como palabras clave de búsqueda con una mejor comprensión hombre y máquina al asociar a estas web estos descriptores en metadatos que pueden hallarse en el repositorio de MedIEQ, en la base de datos del sistema de acreditación o en la propia web acreditada.

Agradecimientos

MedIEQ es un proyecto financiado por la Unión Europea (Ref. 2005107) bajo el programa de acción en la comunidad en el campo de la Salud Pública (2003-2008) del Directorate General SANCO. Participan el National Center for Scientific Research "Demokritos" (NCSR), Grecia (coordinador); el Institute of Informatics and Telecommunications, Software and Knowledge Engineering Laboratory (I-sieve Ltd.), Grecia; la Universidad Nacional a Distancia (UNED), España; Web Médica Acreditada (WMA) del Colegio Oficial de Médicos de Barcelona, España; la Agency for Quality in Medicine (AQuMED), Alemania; la University of Economics in Prague (UEP), República de Checoslovaquia; la Helsinki University of Technology (TKK), Finlandia; Geneva University Hospitals, Service of Medical Informatics (HUG), Suiza.

Referencias bibliográficas

1. **Eysenbach, G**. "Consumer health informatics". En: *BMJ*, 2000, v. 320, n. 4, pp. 1713-1716.

2. **Díaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW**. „Patients´ use of the Internet for medical Information". En: *J Gen Intern Med*, 2002, v. 17, n. 4, pp. 180-185.

3. Analysis of 9th HON Survey of Health and Medical Internet Users Winter 2004-2005. Consultado en: 5-9-2006. *http://www.hon.ch/Survey/Survey2005/res.html.*

4. Dublin Core Metadata Initiative. Consultado en: 5-9-2006. *http://es.dublincore.org.*

5. The Friend Of a Friend (FOAF project). Consultado en: 5-9-2006. *http://www.foaf-project.org.*

6. **Eysenbach G, Kohler C, Yihune G, Lampe K, Cross P, Brickley D**. "A metadata vocabulary for self- and third-party labeling of health web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL)." *Proc AMIA Annu Fall Symp JAMIA Suppl*, 2001, pp. 169-173.

7. World Wide Web Consortium (W3C). Consultado en: 2-9-2006. *http://www.w3.org.*

8. **National Library of Medicine**. Unified Medical Language System (UMLS). Consultado en: 1-9-2006. http://www.nlm.nih.gov/research/umls/umlsmain.html.

9. **Mayer MA, Karkaletsis V, Stamatakis K, Leis A, Villarroel D, Thomeczek C et al**. "MedIEQ – Quality Labelling of Medical Web Content Using Multilingual Information Extraction." En: L. Bos et al. (Eds). Press *Medical and Care Compunetics 3*. IOS, Proc ICMCC Event 2006, pp. 183-190.

10. **Kohler C, Darmoni SD, Mayer MA, Roth-Berghofer T, Fiene M, Eysenbach G.** "MedCIRCLE - The Collaboration for Internet Rating, Certification, Labelling, and Evaluation of Health Information". Technology and Health Care, Special Issue: Quality e-Health. *Technol Health Care.* 2002, v. 10, n. 6, pp. 515.

11. Quality Assurance and Content Description (QUATRO). Consultado en: 1-9-2006. *http://www.quatro-project.org.*

12. **Manola F, Miller E**. "RDF Primer. W3C Recommendation 10 February 2004". Consultado en: 31-8-2006. *http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.*

13. **Mayer MA, Leis A, Sarrias R, Ruíz P**. "Web Mèdica Acreditada Guidelines: realiability and quality of health information on Spanish-Language websites". En: R. *Engelbrecht et al (Eds).Connecting Medical Informatics and Bioinformatics. Proceedings of the 19th International Congress of the European Federation for Medical Informatics (CD-ROM)*, 2005, v. I, n. 1, pp.1287-1292.

14. Web Médica Acreditada (WMA). Consultado en: 8-9-2006. *http://wma.comb.es.*

15. Health on the Net Code. Consultado en: 30-8-2006**.** ***http://www.hon.org*****.**

16. **Olvera MD.** "Rendimiento de los sistemas de recuperación de información en la World Wide Web: revisión metodológica". *Rev Esp Doc Cient*, 2000, v. 23, n. 1, pp. 63-77.