MEDIEQ: METADATOS Y SISTEMAS DE EXTRACCIÓN SEMÁNTICA DE INFORMACIÓN SANITARIA EN INTERNET Y SU APLICACIÓN EN ESTRATEGIAS DE CALIDAD

M.A. MAYER[1], A. LEIS[1], P. RUIZ[1], V. KARKALETSIS[2], K. STAMATAKIS[2]
[1]*Web Médica Acreditada. Colegio Oficial de Médicos de Barcelona. 08017-Barcelona. España.*
[2] *National Center for Scientific Research (NCSR) "Demokritos". Atenas. Grecia.*

Los lenguajes de metadatos propuestos por el World Wide Web Consortium (W3C), basados en XML/RDF, y los sistemas de extracción y análisis automatizados de contenidos son tecnologías de las que disponemos actualmente para ofrecer un Internet más eficiente y más útil. Por otro lado, 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 personas. Se hace necesario proveer a los usuarios de salud de herramientas para discernir la información correcta de aquella que no lo es. Los sistemas de acreditación y los portales temáticos que filtran las webs de mayor calidad deben incorporar, para aprovechar al máximo su potencial y ofrecer el mejor servicio, las tecnologías comentadas anteriormente. Describimos el proyecto europeo MedIEQ que complementa de forma práctica, tecnologías de metadatos y extracción automatizada de datos en el área de la información de salud en Internet.

1. Introducción

Es bien conocido que la información médica existente en la Red es cada vez más amplia y crece de forma exponencial día a día así como el interés por parte del público en general. Sirva como ejemplo el hecho de que si situamos en un buscador de amplia utilización como Google, la palabra salud, obtenemos 140.000.000 de enlaces relacionados. Esta información presenta contenidos y calidad muy variables, desde contenidos científicos contrastados y elaborados bajo premisas de Medicina Basada en la Evidencia, hasta aquella información que puede ser engañosa o peligrosa para la salud si se utiliza inadecuadamente por los pacientes o sus familiares. Desde la Unión Europea [1] y diversas organizaciones nacionales e internacionales están proponiendo soluciones y recomendaciones con el objetivo de garantizar, en lo posible, que dicha información presente unos mínimos criterios de fiabilidad. Los sellos de calidad y los portales que filtran la información tras el cumpliendo de una serie de criterios de calidad establecidos previamente, son algunas de las soluciones propuestas cada vez más conocidas por los usuarios de Intenet.[2-4] La información sobre salud está dirigida tanto al público en general como a los profesionales.

Además es necesario seguir desarrollando y aplicando 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 del propio sistema y de las webs acreditadas o filtradas. Disponemos de lenguajes de metatados y Web Semántica y análisis de contenidos mediante la utilización de vocabularios y clasificaciones específicamente creados para la descripción y etiquetado de contenidos. Cabe destacar el RDF-CL,[5] Dublin Core,[6] Friend of a Friend (FOAF),[7] HIDDEL [8] y otros estándares

desarrollados por grupos de trabajo del World Wide Web Consortium (W3C) [9] como el Content Label Incubator Group (WCL XG),[10] así como tesauros del ámbito de las ciencias de salud como el Unified Medical Language System (UMLS) de la National Library of Medicine [11].

El proyecto europeo MedIEQ (Quality Labeling of Medical Web Content using Multilingual Information Extraction) [12] 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, [13] WRAPIN y QUATRO,[14] 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. El objetivo es la mejora en 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) [15] 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. Escenario de trabajo: organismos de acreditación y filtrado de información web

2.1 Web Médica Acreditada: agencia 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: [16, 17]

- 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.

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. Algunas aplicaciones de web semántica son: 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 y la guía elaborada por la Unión Europea, e-Europa 2002: Criterios de calidad para sitios web relacionados con la salud.[1]

2.3 Sistemas de extracción multilingüe

Desde hace años se está investigando en el desarrollo de tecnologías de extracción de datos en textos.[18,19] Los textos pueden estar en páginas HTML web. Básicamente la extracción se basa en obtener en primer lugar unos datos aislados (palabras, cadenas de palabras), que en una segunda fase se integran para generar o traducirlos en nuevos datos en un formato que nos sea útil. En el caso que nos ocupa, determinadas expresiones halladas en la web nos indicarían la existencia de determinados datos asociados a las mismas, los cuales, en un segundo paso servirían para conocer y deducir su contenido automáticamente. Sirva como ejemplo: si encontramos expresiones en la web como "Envíenos su consulta", "El Dr. X responde" o "Servicio de consulta web" el sistema deduce, en base a las relaciones conceptuales que hemos definido previamente, que existe "Consulta virtual" en esta web. Posteriormente se asignará en el campo descriptivo de esta web como afirmativa la existencia de este servicio y se incluirá en el informe sobre la misma, en este caso, en un campo concreto del archivo XML/RDF. 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. 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 escenario propuesto, Web Médica Acreditada, se aplican las herramientas de la plataforma que permiten 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 (como se muestra en la figura 1):

- 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 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.

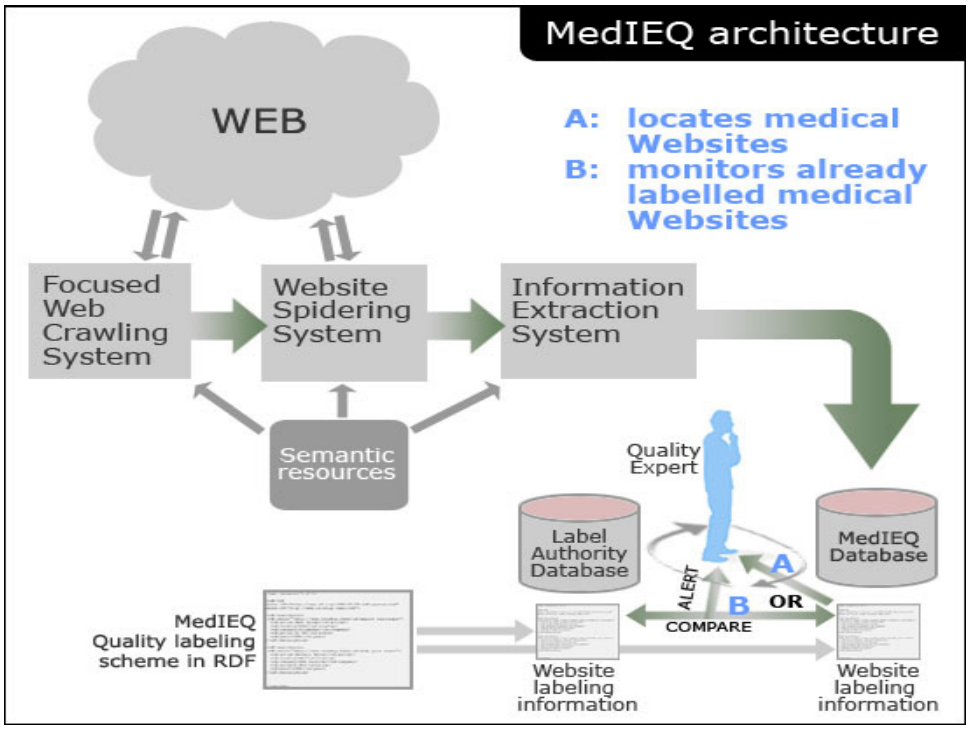

Figura 1: *Modelo de pie de figura a una distancia de una línea del texto.*

3. Conclusiones

El extraordinario y rápido crecimiento en el número y contenidos de las páginas web sanitarias y el interés manifiesto de los usuarios de Internet por obtener información sobre salud, hacen necesario aplicar mecanismos de control como los sellos de calidad y sistemas de acreditación o los portales que actúan como filtros. Para alcanzar este objetivo de una forma adecuada, debemos combinar dichos mecanismos de control con diversas herramientas disponibles actualmente, que pueden permitir optimizar las tareas de control 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 incluso a la creación de portales de información sanitaria que agrupen estos recursos de forma automática adaptándolos a los parámetros que definan las diferentes organizaciones implicadas.

Aunque es cierto que se requiere más experiencia en este campo y por otro, de momento es aplicable únicamente en algunos datos y descriptores que caracterizan las webs, MedIEQ puede ofrecer un esquema de trabajo, basado en la utilización de estos lenguajes estandarizados y los sistemas de extracción 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. Todo ello ha de redundar en proporcionar a los usuarios de mayor y mejor orientación sobre las

características de la información sanitaria en Internet y contribuir a su educación sanitaria y facilitar y simplificar algunas de las tareas de los sistemas de acreditación y constituir una línea de aplicación a desarrollar más ampliamente.

Agradecimientos

MedIEQ es un proyecto financiado por la Unión Europea, bajo el programa de acción en la comunidad en el campo de la Salud Pública (2003-2008) del Directorate General DG-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

[1] Comisión de las comunidades europeas. eEurope 2002: Criterios de calidad para los sitios web relacionados con la salud. Consultado en: 10-9-2006. Accesible en: http://europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_es_fin.pdf.

[2] Mayer MA, Leis A, Karkaletsis, Archer P, Stamatakis K, Perego A, Ruiz P. La visión integradora del proyecto QUATRO en la aplicación de la web semántica para garantizar un Internet de confianza. En: *Actas de IX Congreso Nacional de Informática de la Salud. Las TIC en la protección de la Salud*. Sociedad Española de Informática de la Salud, Madrid, 28-30 Marzo (2006).

[3] Eysenbach, G. Consumer health informatics. *BMJ*. **320 (4),** 1713-1716 (2000).

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

[5] RDF Content Labels: Schema Description. Consultado en: 14-9-2006. Accesible en: http://www.w3.org/2004/12/q/doc/content-labels-schema.htm.

[6] Dublin Core Metadata Initiative. Consultado en: 5-9-2006. Accesible en: http://es.dublincore.org.

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

[8] 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*, 169-173 (2001).

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

[10] W3C Incubator Group Report Draft 0.9.2; August 2006. Consultado en: 14-9-2006. Accesible en: http://www.w3.org/2005/Incubator/wcl/XGR-report.

[12] 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). *Medical and Care Compunetics 3.* IOS, Proc ICMCC Event, 183-190 (2006).

[13] 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,* **10 (6),** 515 (2002).

[14] Quality Assurance and Content Description (QUATRO). Consultado en: 1-9-2006. Accesible en: http://www.quatro-project.org.

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

[16] Web Médica Acreditada. Consultado 14-9-2006. Accesible en: http://wma.comb.es.

[17] 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), **Vol.I(1),** 1287-1292 (2005).

[18] Grishman R. Information extraction: techniques and challenges. *Lecture Notes In Computer Science*, **Vol. 1299,** 10-27 (1997).

[19] Olvera MD. Rendimiento de los sistemas de recuperación de información en la World Wide Web: revisión metodológica. *Rev Esp Doc Cient*, **23(1),** 63-77 (2000).