Metadata element set in a Quality-Controlled Subject Gateway: a step to an health semantic Web

Benoit Thirion [a], Gaelle Loosli [b], Magaly Douyère[a] , Stéfan J. Darmoni [a, b]

[a] *L@sTICS, Rouen University Hospital & Medical School, France*
[b] *Perception, Information and System Lab, INSA Rouen & Rouen University, France*

Abstract

Background: Quality-controlled subject gateways are Internet services which apply a selected set of targeted measures to support systematic resource discovery. Considerable manual effort is used to process a selection of resources which meet quality criteria and to display a extensive description and indexing of these resources with standards-based metadata.Objective: Several metadata element sets are proposed to describe, index and qualify health resources to be included in a French quality-controlled health gateway called CISMeF. The main objectives were to enhance Internet health document retrieval and navigation, and to allow interoperability with other Internet services.Results: The Dublin Core metadata element set is used to describe and index all Internet health resources included in CISMeF. For teaching resources, some elements from IEEE1484 Learning Object Metadata are also used. For evidenced-base medicine resources, specific metadata are employed which assess the health content quality. The HIDDEL metadata set is used to enhance transparency, trust and quality of health information on the Internet. Conclusion: Comprehensive metadata element sets can be extremely useful to describe, index and assess health resources on the Internet in a quality-controlled subject gateway. Machine-readable metadata creates an Semantic Web which is more efficient for end-users as compared to the current Web.

Keywords:

Database; Information Storage and Retrieval; Internet; Medical Informatics; Quality Control

1. Introduction

Metadata is concise information concerning all types of data. On the Internet, metadata specifically refers to: 1) descriptive information about the Web resources used to improve information retrieval, 2) it refers to the content, structure and logistical information of all data including electronic resources, 3) it is used for data discovery and control of data, 4) it helps to enhance Internet health document retrieval and navigation. There is a need for an interoperable infrastructure for Digital Libraries, quality-controlled subject gateways, and other Web-based services that rely on cross-institutional and cross-border co-operation. Agreement on a metadata standard which serves as a starting point for information exchange in specific domains and provides a common ground for cross-domain interoperability, is a crucial element of this infrastructure. The main metadata standard from a cross-domain perspective is the Dublin Core, now recommended across Europe for use in many sectors as the gold standard of choice to ensure interoperability between resource discovery systems on the Internet.

Quality-controlled subject gateways were defined by Koch [1] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery. Considerable manual effort is used to process a selection of resources which meet quality criteria and to display a extensive description and indexing of these resources with standards-based metadata. Regular checking and updating ensure optimal collection management. The main goal is to provide a high quality of subject access through indexing resources using

controlled vocabularies and by offering a deep classification structure for advanced searching and browsing. The objective of CISMeF (French acronym for Catalog and Index of French-language health resources) [2-3] is to describe and index the main French- language health resources to assist health professionals and consumers in their search for electronic information available on the Internet. CISMeF is a quality-controlled subject gateway initiated by the Rouen University Hospital (RUH). Its Universal Resource Locator (URL) is http://www.chu-rouen.fr/cismef. CISMeF began in February 1995. In July 2002, the number of indexed resources totalled over 10,000, with an average of 50 new resources indexed each week. Each of the following phases proposed by Koch, which characterise a typical quality-controlled subject gateway, are implemented in CISMeF: (a) selection and collection development, based on the Net Scoring, list of 49 criteria to assess quality of health information (URL: http://www.chu-rouen.fr/netscoring) [4], (b) collection management, (c) intellectual creation of metadata (done by experts), (d) resource description (an extensive and documented metadata set), and (e) resource indexing (using a controlled vocabulary system).

The main objectives of this work were to enhance Internet health document retrieval and navigation, and to permit interoperability with other Internet services. To allow interoperability, gateways apply open standards. CISMeF uses two standard tools for organising information: the MeSH (Medical Subject Heading) thesaurus from the US National Library of Medicine (NLM), and several metadata element sets: (a) the Dublin Core metadata format [5] to describe and index all the health resources included in CISMeF, (b) some elements from IEEE1484 Learning Object Metadata for teaching resources [6], (c) specific metadata for evidenced-base medicine resources which also qualify the health content, and (d) the HIDDEL metadata set [7] will be used to enhance transparency, trust and quality of health information on the Internet in the EU-funded MedCIRCLE project.

2. Methods

Description of the Dublin Core: The Dublin Core Metadata Initiative (DCMI) is a project from the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA). The DCMI is a metadata 15-element set intended to facilitate the discovery of electronic resources (URL: http://dublincore.org). Originally conceived for an author-generated description of Web resources, the DCMI is now used by museums, libraries, government agencies, and commercial organizations alike. The construction of an interdisciplinary, international consensus around a core element set is the central feature of the DCMI which benefits from active participation and promotion in over 20 countries in North America, Europe, Australia, and Asia. The DCMI is intended to be used by non-cataloguers as well as resource description specialists. CISMeF covers three main health areas: evidence-based medicine, education of health care professionals and students, and patient education. Specific metadata are used for these domain.

Description of IEEE 1484 Learning Objects Metadata (LOM) for teaching resources: The IEEE 1484 Learning Object Metadata (LOM) (URL: http://ltsc.ieee.org/doc/wg12/LOM_WD6_4.pdf) version 6.4 contains over 60 elements in the following nine categories: General,. Lifecycle, Meta-metadata, Technical Educational, Rights, Relation, Annotation, Classification. LOM metadata includes the 15 DCMI elements (see URL: http://www.ischool.washington.edu/sasutton/IEEE1484.html).

Description of EBM (Evidence-Based-Medicine) metadata element set: CISMeF uses two specific metadata elements for EBM resources and more broadly ‘sensitive’ information. Sensitive information is defined as information found in documents published on the Internet, which could be used in a medical decision: These two metadata elements are: (a) indication of level of evidence which we proposed to be the main criterion chosen for the quality of the health information content [8] and (b) the method used to calculate the level of evidence as more than twenty are currently used in the literature. CISMeF explicitly indicates if level of evidence is mentioned for each indexed 'sensitive' document. Furthermore, this criterion is easily

searchable using the Doc'CISMeF search tool [3].

Description of "Health Information, Disclosure, Description and Evaluation Language" (HIDDEL): HIDDEL is a standard vocabulary/metadata language developed in the MEDCIRCLE project (URL: http://www.medcircle.info) [7]. HIDDEL is designed to be used by 1) information providers to describe and disclose properties of e-health services (self-rating) and 2) third-parties, e.g. by subject gateways, to express third-party opinions about health information providers.

3. Results

Using DCMI in CISMeF: The fifteen Dublin Core elements are optional and repeatable. Resources included in CISMeF are described by 11 of 15 items taken from version 1.1 of the DCMI (URL: http://dublincore.org/documents/dces/). These are: author or creator, date, description, format, identifier, language, publisher, resource type, rights, subject and keywords, and title. CISMeF does not use the 4 other DCMI items (contributor, coverage, relation, source) [9] because they were not necessary to describe health resources to be included in CISMeF. To capture more information for each health resource indexed in CISMeF, another element set was developed locally to meet specific search and retrieval needs. The following eight fields were added in the data and metadata and are specific to CISMeF: institution, city, province or state, country, target or audience, type of access, cost and sponsorship. Some of these fields (e.g. cost) are also present in LOM.

From 1995 to 1999, CISMEF used only static HTML. As CISMeF uses the MeSH to index resources, each HTML page is based on a MeSH term and includes Dublin Core metadata. In November 2002, CISMeF used 6,853 MeSH terms (34% of the MeSH thesaurus). Since 2000, CISMeF also includes a database and a search tool which generates an HTML (or XML or RDF) page for every indexed resource.

Using LOM in CISMeF: CISMeF is one of the search tool of the French Medical Virtual University (FMVU) Consortium which was created to test various tools and methods required to build a virtual university (URL: http://www.umvf.org). To describe and index teaching resources, this consortium decided to use in its search tools only the 11 elements of the LOM Educational category because they are the most specific. Also, a feasibility study showed that: the CISMeF team spends an average of 30 minutes to describe and index a teaching resource with the Dublin Core set and needs 30 minutes more for the LOM Educational subset.

Recently, DCMI proposed a new section DC.education to develop and promote a set of basic principles for the development and application of modular interoperable metadata for dissemination to the global education and training communities. DC.education will map with several elements of the LOM Educational subset. The CISMeF metadata element set will use DC.education as soon as it will be finally approved.

The use of HIDDEL in CISMeF: CISMeF is a member of the MedCIRCLE project which is a collaboration of trusted European health subject gateways, medical associations, accreditation, certification, or rating services, which share the common goal of evaluating, describing, or indexing health information. The MedCIRCLE project is funded by the European Union under the Action Plan for Safer Use of the Internet (URL: http://www.europa.eu.int/information_society/programmes/iap/index_en.htm). This project began in March 2002 and will last 18 months. As a quality-controlled subject gateway, CISMeF uses HIDDEL only as a third-party. Some elements of the HIDDEL are similar to Dublin Core (e.g. HIDDEL.Identity and DC.Author). Most of the HIDDEL elements are common with the Net Scoring previously used by CISMeF whereas some are already present in the CISMeF database (e.g. HIDDEL.policies). CISMeF focuses this rating on the main French publishers of health resources (national agencies, medical societies, universities and hospitals) which are included in the CISMeF database. In CISMeF, each publisher will have a MedCIRCLE seal with a link to the MedCIRCLE central repository where HIDDEL metadata elements are

displayed. CISMeF will apply full transitivity from these publishers: each document from one MedCIRCLE rated publisher which is indexed in CISMeF will also receive the MedCIRCLE seal of the publisher with the same link to MedCIRCLE central repository. Thanks to this transitivity, over 5,000 resources will have a MedCIRCLE seal in CISMeF in August 2003 (50% of the CISMeF resources). The HIDDEL language is integrated in the CISMeF database and in the CISMeF pages via RDF into HTML.

This metadata element set will be useful for cross-searching distributed and heterogenous subject gateways. We have successfully tested the interoperability of the CISMeF metadata element set with the FMVU e-learning platform using the XML version of CISMeF resource pages. These metadata elements were manually written and updated by the CISMeF team from 1995 to 1999 and currently automatically created and updated from the CISMeF database. Till 2002, the CISMeF metadata element set were mostly human-readable and not so easily machine-processable. From August 2002, we have used Resource Description Framework (RDF) into HTML (URL: http://www.w3.org/RDF) to become easily machine- processable and therefore to fulfil one of the main goal of this metadata element set: to become interoperable with other Internet services. RDF is a language for encoding knowledge on Web pages to be used by electronic agents searching information. Developed by the World Wide Web Consortium (W3C), it is a major building block of the Semantic Web initiative. The Semantic Web is the abstract representation of data on the Web, based on the RDF standards and can be defined as an extension of the current Web in which information is given well-defined meaning to facilitate the interchange of computers and people [10].

4. Discussion

Several main tools could be targeted for the retrieval of health information on the Internet in ascending order: *level 1*: search engine, general or more specialised searches, such as MedHunt-Ch (URL: http://www.hon.ch/); *level 2*: catalogue and index without thesaurus, such as MedWebPlus-Us (URL: http://www.medwebplus.com/) and HealthWeb -Us (URL: http://healthweb.org/); *level 3*: catalogue and index with thesaurus, such as the Unified Medical Language System (UMLS) metathesaurus and MeSH thesaurus. The latter thesaurus is used in the following Health catalogues: DDRT (Diseases, Disorders and Related Topics) from the Karolinska Institute Library, Sweden (URL: http://www.mic.ki.se/Diseases/), CliniWeb [11] (URL: http://www.ohsu.edu/cliniweb/), Oregon Health Sciences University-USA, OMNI (Organizing Medical Networked Information-UK) (URL: http://omni.ac.uk/) [12] and HON (Health on the Net-Ch) from Switzerland [13]; *level 4*: catalogue and index with thesaurus, metadata, and description of sites. To our knowledge, CISMeF and Healthinsite-Au (URL: http://www.healthinsite.gov.au/) [14] have now reached level 4.

CISMeF uses DCMI differently according to the "browse" or "search" strategy chosen by the end-user. The choice of the Dublin Core was prompted by its institutional origin and its notoriety in the academic world. Several other health sites are now successfully using the Dublin Core, including the NLM (see a comprehensive list of health sites using DCMI at the following http://www.chu-rouen.fr/documed/dc.html).

The use of metadata is one main criterion in accurately assessing the quality of health information on the Internet [4]. In order to use metadata, it is necessary to properly structure information. The quality of metadata description may reflect the quality of online information. The following search on Medline using Pubmed shows that metadata is a new field of research: using the following request “metadata” in all the fields of the Medline database (2002-11-14) because metadata is not (yet) a MeSH keyword, we found 64 references mostly published in the last three years, such as the latest reference which proposed specific metadata to describe video resources [15]. Metadata allow to structure information in the same way throughout several databases. Therefore, it should be easier to use these databases based on a common metadata set.

This metadata element set is useful for cross-searching distributed and heterogeneous subject

gateways and for the creation of meta-catalogs or meta-gateways, such as Renardus [16]. The aim of the EU-funded Renardus project (URL: http://www.renardus.org) is to provide users with integrated access through a single interface to high-quality Internet resources permitting to search and browse records from existing distributed subject gateways across Europe. In the near future the interoperability of CISMeF metadata element set with the Renardus consortium will be tested.

This concept of meta-gateway could be applied in the medical field with the creation of a health meta-gateway including the following health gateways sharing the same thesaurus (MeSH): CISMeF, CliniWeb, DDRT, HON, MedWebPlus, and OMNI. These gateways should share the same metadata element set. Dublin Core and HIDDEL could be the minimum common element set as the MedCIRCLE project expects formal standardization of this vocabulary in collaboration with standardization organizations and committees (TC251/CEN/ISO). HIDDEL will play a decisive role in demonstrating and ensuring interoperability of rating services and will enable harvesting and dissemination of third-party ratings. In addition to CISMeF, Mallet et al. [17] and Boulos et al. [18] previously proposed a specific health metadata element set. It is essential to find a common health metadata set used at least by the main health gateways to become interoperable**.** Fortunately, the CEN/ TC251 (European Standardization of Health Informatics) has developed a current project "Metaknow - Metadata for medical knowledge resources" with the aim of establishing a small set of medically relevant metadata items suitable to apply to medical knowledge (URL: http://www.centc251.org/WGII/N-01/WGII-N00-05.pdf).

To complement the interdisciplinary nature of this work, our approach will be to reach out to other health professionals that have terminologies already in place. Experts in the field of nursing language should be included among the metadata creators. The two CISMeF specific metadata elements for EBM resources should also be extended to those recently proposed by Sakai [19]. In contrast, even using the LOM Educational subset alone was time consuming (doubling the time to describe and index a teaching resource). Instead of using the entire LOM metadata set, a project is underway to test the DC.Education subset in the future. Metadata also allows to coherently structure any institutional Web site as shown by Davenport et coll. [20] with the National Institute of Environmental Health Sciences Web site. Information professionals such as librarians have a main role in that task including training. These professionals should be part of every editorial board of institutional and academic Web sites.

Architectural characteristics and technical tools supporting implementation of the Dublin Core metadata standards should be considered, with the objective of contributing to the global development of specifications and identify the crucial elements that need to be in place to support the deployment of the metadata standards in a way that semantic relationships can be expressed and implemented in a machine-readable way, thereby supporting the vision of the Semantic Web by providing practical elements for its implementation, i.e. as proposed in health by HealthCyberMap (URL: http://healthcybermap.semanticweb.org/) using UMLS [16] or CISMeF using a semi-formal ontology based on MeSH [3].

5. Conclusion

To help healthcare professionals and health consumers to more easily locate high-quality health information on the Internet, catalogues must use standard tools especially metadata to describe, index and qualify Internet health resources. Machine-readable metadata builds a Semantic Web that will be more useful for end-users than the current Web and will require a concerted use of metadata.

6. References

[1] Koch T. Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review* 2000: 24 (1) piii: 24-34.

[2] Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J and Thirion B. CISMeF: a structured Health resource guide. *Methods Inf Med* 2000: 39 (1) piii: 30-5

[3] Darmoni SJ, Thirion B, Leroy JP, Douyere M, Lacoste B, Godard C, Rigolle I, Brisou M, Videau S, Goupy E, Piot J, Quere M, Ouazir S and Abdulrab H. A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet. *Med Inform Internet Med* 2001: 26 (2) piii: 165-78

[4] Centrale Santé. Net Scoring : criteria to assess the quality of Health Internet information. 19 Sep 2001 [Web document, accessed 3 August 2002]. Available from Internet: <http://www.chu-rouen.fr/netscoring>.

[5] Weibel SL and Koch T. The Dublin Core Metadata Initiative. *D-Lib Magazine* 2000. Available from Internet: <http://www.dlib.org/dlib/december00/weibel/12weibel.html>.

[6] IEEE1484 – IEEE Learning Technology Standards Committee (LTSC). Available from Internet: <http://ltsc.ieee.org/>.

[7] Darmoni SJ, Haugh MC, Lukacs B and Boissel JP. Quality of health information about depression on internet : Level of evidence should be gold standard. *Br Med J* 2001: 322 piii: 1366.

[8] Eysenbach G, Yihune G, Lampe K, Cross P and 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 Symp* 2001 piii: 169-73.

[9] Darmoni SJ, Thirion B, Leroy JP and Douyere M. The use of Dublin Core metadata in a structured health resource guide on the Internet. *Bull Med Libr Assoc,* 2001: 89 (3) piii: 297-301.

[10] Berners-Lee T, Hendler J and Ora L. The Semantic Web. Scientific American. May 2001. Available from Internet: <http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2>.

[11] Hersh WR, Brown KE, Donohoe LC, Campbell EM and Horacek AE. CliniWeb: managing clinical information on the World Wide Web, *JAMIA* 1996: 3 (4) piii: 273-80.

[12] Norman F. Organising Medical Networks' information (OMNI). *Med Inf* 1998: 98 (23) piii: 43-51.

[13] Boyer C, Baujard O, Baujard V, Aurel S, Selby M and Appel RD. Health On the Net automated database of Health and medical information, *Int J Med Inf* 1997: 47 (1-2) piii: 27-9.

[14] Deacon P, Smith JB and Tow S. Using metadata to create navigation paths in the HealthInsite Internet gateway. *Health Info Libr J.* 2001: 18 (1) piii: 20-9.

[15] Shotton DM, Rodriguez A, Guil N and Trelles O. A metadata classification schema for semantic content analysis of videos. *J Micros* 2002: 205(Pt 1) piii: 33-42.

[16] Neuroth H and Koch T. Metadata Mapping and Application Profiles. Approaches to providing the Cross-searching of Heterogeneous Resources in the EU Project Renardus. DC-2001 (International Conference on Dublin Core and Metadata Applications. Tokyo). October 2001, pp. 122-29. Available from Internet: <http://www.nii.ac.jp/dc2001/proceedings/abst-21.html>.

[17] Malet G, Munoz F, Appleyard R and Hersh W. A model for enhancing Internet medical document retrieval with "medical core metadata". *J Am Med Inform Assoc* 1999: 6 (2) piii: 163-72.

[18] Boulos M, Roudsari A and Carson E. Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set. *Med Sci Monit,* 2002: 8 (7) piii: 24-36.

[19] Sakai Y. Metadata for Evidence Based Medicine Resources. DC-2001 (International Conference on Dublin Core and Metadata Applications. Tokyo). October 2001, pp. 81-5. Available from Internet: <http://www.nii.ac.jp/dc2001/proceedings/abst-12.html>.

[20] Davenport Robertson W, Leadem EM, Dube J and Greenberg J. Design and Implementation of the National Institute of Environmental Health Sciences Dublin Core Metadata Schema. DC-2001 (International Conference on Dublin Core and Metadata Applications. Tokyo). October 2001, pp. 193-99. Available from Internet: <http://www.nii.ac.jp/dc2001/proceedings/abst-29.html>.

7. Adress for correspondence

Darmoni Stéfan J., Computer and Networks Department, 1 rue de Germont 76031 Rouen Cedex, France
Tel: +33.232.88.88.29; Fax: +33.232.88.88.32
E-mail: Stefan.Darmoni@chu-rouen.fr