A metadata vocabulary for self- and third-party labeling of health web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL)

G. Eysenbach[1)], C. Köhler[1)], G. Yihune[1)], K. Lampe[2)], P. Cross[3)], D. Brickley[3) 4)]

1) Dept. of Clinical Social Medicine, Unit for Cybermedicine, University of Heidelberg, Germany, email: ey@yi.com
2) FinOHTA / STAKES Finnish Office for Health Technology Assessment, Finland
3) ILRT Institute for Learning and Research Technology, United Kingdom
4) W3C Metadata Interest Group, MIT, Boston

We describe HIDDEL (Health Information Disclosure, Description and Evaluation Language), formerly known as medPICS (platform for Internet content selection in medicine), a metadata vocabulary designed to enhance transparency, trust and quality of health information on the web. The vocabulary may be used (1) by webmasters to self-describe their contents and policies; (2) by info-mediaries (e.g. Healthfinder, NHS Direct/NeLH), e.g. third party evaluators, rating or portal services, to annotate other websites; (3) and by users, to describe their preferences. As an XML application it conforms to the W3C's RDF Specification. The metadata vocabulary is primarily intended to enable descriptions of whole health websites or health information providers. The vocabulary is designed to provide a computer-readable electronic "label" of a health website, telling users who is behind the website, how the website is sponsored, what the content, aim and target audience is, how the information was compiled, what risks the service bears, or what people say about the resource. Client-software can "read" this label automatically, compare it to the user's own set of preferences and needs, and alert and advise users.

Introduction

Meta-data on the World-Wide-Web is mostly used as additional information distributed together with the document it describes (i.e. using the META tag in HTML), but there are also technical possibilities to distribute meta-data by third parties independently from the document or information provider, by using the W3C PICS Standard (Platform for Internet Content Selection, http://www.w3.org/RDF/). PICS was essentially designed to allow third parties to distribute meta-data about other resources, primarily to allow filter software to block access of minors to pornographic and harmful material. The PICS standard is currently migrating towards becoming an application of the XML/RDF technology of the W3C (see http://www.w3.org/RDF/) [1]. The successor standard, RDF (the Resource Description Framework), grew out of work on expanding PICS (then called PICS-NG) to provide for more flexible descriptive capabilities (e.g. textual comments). RDF is a W3C Recommendation for Web "resource description" which includes labeling, classification, cataloguing, rating etc. RDF in turn adopts the W3C XML Recommendation as a new file format for ex-changing such data, replacing the PICS 1.1 format.
In this paper we present a self-description and third-party rating metadata vocabulary for networked health resources, expressed for example in PICS, XML or RDF.

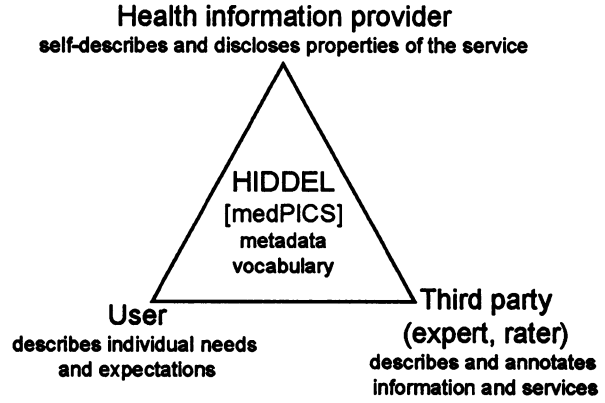

Figure 1. HIDDEL (formerly known as MedPICS) is designed to be applied by three different user groups

A number of other initiatives are working on developing, identifying and standardizing generic metadata vocabularies, for example the Dublin Core (DC) group [2; 3]. However, the DC vocabulary is primarily designed to be used to provide descriptive metainformation supplied by web-authors, but not from external evaluators, and does not provide disclosure elements required to allow consumers to make informed decisions when using a health resource. For example, there are no DC elements allowing the

1067—5027/01/$5.00

description of potential conflicts of interest or biases; disclosure of sponsors; purpose of the site, or description of internal quality management procedures. All these elements have been recognized as being important elements of an "ethical" website.[4; 5] Full disclosure is especially important in health care, thus, generic models such as Dublin Core have thus far not taken up the idea that disclosure could be provided as meta-data. Moreover, most of the DC elements are generally more suitable to be applied on a document level, rather than on a website or health information provider level. For example, some Dublin Core elements have been developed to describe the date of publishing or last update of a document, but there are no elements allowing description of the general policy for how often the content is reviewed and updated. Health consumers also should be able to check the level of evaluation the interactive application has undergone, and to tell what the purpose of the information on the website is (e.g. educational vs marketing) or whether there are any gross biases.

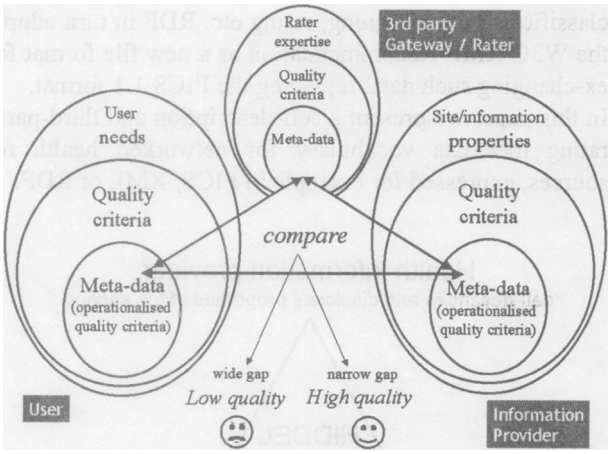

Figure 2. *HIDDEL/medPICS allows health information providers to disclose and describe properties of their site or service. Users use the same vocabulary and the same scales to describe their needs or expectations, and third parties use the same vocabulary to describe, annotate, comment on, or evaluate sites or services. This enables the user to make informed choices and to select sites which meet his/her own needs. "Quality" is defined as the gap between consumer needs/expectations and the meta-data supplied by the information provider and/or third parties.*

In summary, there is a need to expand the meta-data concept and to shift thinking from using meta-data elements only to enhance information retrieval, towards a model where meta-data elements may also describe disclosure information, site policies on the host side, and statements of third parties about other sites. The same vocabulary and framework can then be used on the client side to code user preferences, allowing automatic selection of suitable information meeting personal quality criteria. To fill this gap, we proposed (as early as in 1997) a medical core metadata vocabulary set based on the W3C PICS (Platform for Internet Content Selection) standard, which we called medPICS [6; 7]. MedPICS 0.3 already contained elements that are evaluative, e.g. elements typically assigned by third parties, as well as self-description which have disclosure character.

The central idea behind promoting the widespread use of MedPICS is that users may express their personal preferences (e.g. "the website should not be sponsored by a pharmaceutical company and should target consumers") in a computer-readable metadata format, and the client-software could then automatically and continuously compare these preferences with the respective metadata given by the health information provider (and possibly by third parties, e.g. portal sites) whenever a user accesses a site (Fig. 3), or this meta-data can be indexed to allow direct searches for sites meeting a given users preferences.

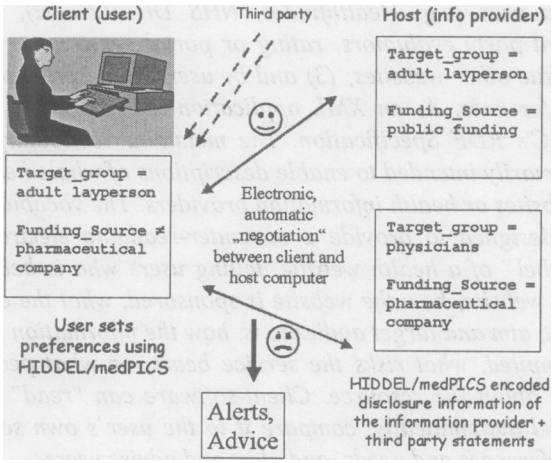

Figure 3. *Widespread use of HIDDEL/medPICS would allow users to maximize the perceived quality of services by automatically selecting sites which comply to individual user needs, or get advice if they don't.*

Model formulation process

An overview of the development process to date is given in Figure 4. The vocabulary is still under discussion and presented here in its version 0.6. The development of the original medPICS vocabulary 0.3 has been described elsewhere [6]. Briefly, it was based on a systematic literature review of the then available articles, reports and documents pertaining to quality of health information on the web, such as evaluation guidelines of OMNI and

BHIA; as well as FDA hearings pertaining to cross-border sales of drugs.

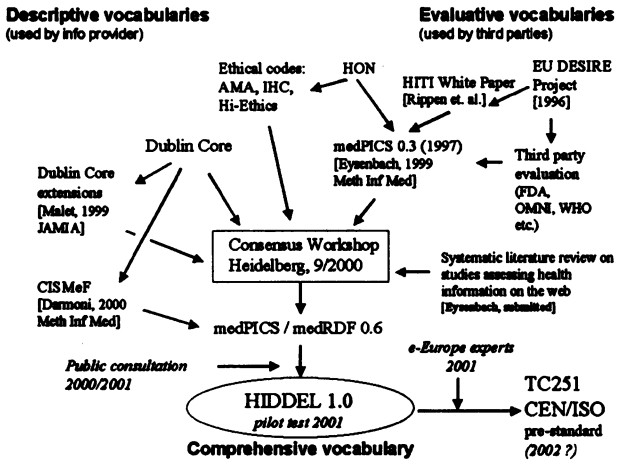

Figure 4. *Overview of the HIDDEL/medPICS development process thus far*

A further important step in developing the vocabulary was to systematically review the literature for studies which already operationalised some of the quality criteria. This review was conducted in May 2000 [5]

An invitational workshop held in September of 2000 brought together 80 medical researchers, health professionals, consumer representatives, representatives of government agencies, inter-governmental and non-governmental medical associations, librarians and computer scientists participants from 20 different countries to discuss quality issues and requirements for health information (see Workshop Proceedings at http://www.jmir.org/2000/3/suppl2/). The focus of the workshop was to attempt an operationalisation and implementation of proposed ethical guidelines for publishing health information on the web [8-11]. Breakout groups of the workshop tried to establish which "quality criteria" could be operationalised. This involves for example the formulation of a clear question asked to the information provider or external evaluator, and the definition of the possible values a element can take, including the definition of controlled vocabularies. MedPICS is now preferably called HIDDEL to account for the fact that it is a technology independent vocabulary (not tied to the PICS technology).

Model description

The vocabulary consists of about 50 so-called basic elements, which may be presented as a taxonomy (tree structure, see. Fig. 5). Examples for basic elements are Infoprovider_feedback_email_technical (containing the email address for technical questions) or Infoprovider_feedback_email_content (for content questions).

Each basic element can in principle be combined with one of 9 sub-elements, for example CHECKED (What is the result of the evaluator verifying the element?), CHECKED_REASON (Explanation for CHECKED), PRESENT (Basic element is present on the site), HOW (Description of how users can identify the basic element on the page(s) themselves).

An information provider can for example combine the basic element Infoprovider_feedback_email_content with the subelement HOW and describe in this element "The feedback email is provided at the bottom of each page". A third-party evaluator could use the subelements CHECKED to confirm this, or to express disagreement.

For many (basic) elements, a controlled vocabulary exists to complement the free text value, in this case, elements end with "_CV".

Elements can have additional attributes, giving additional information about the respective element, for example who entered them and when.

In XML, the health information provider could publish on his webpage the following statements indicating his address and the email which should be used for giving feedback:

```
<infoprovider>
  <feedback>
    <address>
       Bergheimer Str. 58, Heidelberg
    </address>
    <email> feedback@mysite.com </email>
  </feedback>
</infoprovider>
```

Rating services (gateways) evaluating some of these statements could publish a statement like the following:

```
<infoprovider>
  <feedback>
    <adress>Bergheimer Str. 58, Heidelberg
     <checked by="medcertain team">
        Positive</checked>
    </address>
    <email> feedback@mysite.com
     <checked by="medcertain team"
       creator="Gunther" date="22.1.2001">
        Negative
          <reason creator="Gunther">
          Email was bounced! </reason>
     </checked>
    </email>
  </feedback>
</infoprovider>
```

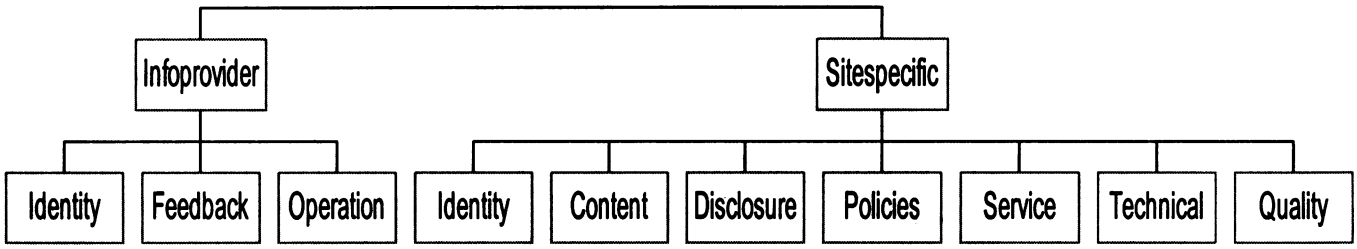

Figure 5. *First two levels of the tree structure of HIDDEL/MedPICS basic elements*

Pilot testing within MedCERTAIN

The vocabulary is currently going through a public consultation phase and will partly be tested in the context of the MedCERTAIN (MedPICS Certification and Rating of Trustworthy Health Information on the Net) project [12]. MedCERTAIN will help to implement this infrastructure, educate consumers and provide a certain level of monitoring to assure that meta-data is not abused for marketing purposes. One aim of the MedCERTAIN project is to make implementation for health information providers easy, even if they lack technical experience: They just have to fill in a form at MedCERTAIN and the metadata wizard will create a set of metadata for them. A so-called MedCERTAIN level 1 "transparency mark" is awarded, if certain metadata fields are submitted. In a level-2 evaluation, meta-data will be verified, and a level-3 evaluation means that the content has been checked, approved, recommended or certified by third parties such as medical experts or societies (this information can also be harvested from trusted gateways, as described below). The MedCERTAIN project will generate further data on the comprehensiveness, applicability and reliability of the vocabulary.

Discussion

The different environments for which we envisage HIDDEL/MedPICS to be used are (1) self-rating by information providers using XML: (2) self-rating by information providers for level-1 certification in the MedCERTAIN context [12]; (3) publication of evaluations, descriptions and annotations by third-party rating services as XML/RDF; (4) evaluation of websites by experts / expert communities using the HIDDEL/MedPICS vocabulary in the MedCERTAIN context; (5) consumers expressing their information preferences using browser add-ons.

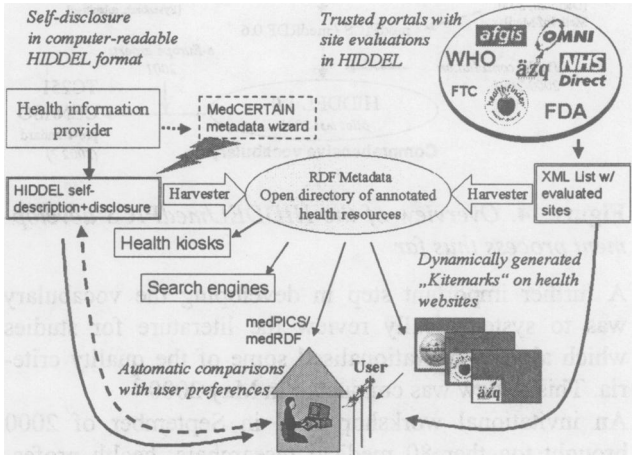

Figure 6. *Harvesting options.*

Harvesting opinions

HIDDEL/MedPICS descriptions may either be published by health information providers themselves (Fig. 6, left top) or may be published by portals or gateways e.g. as XML annotated lists (Fig. 6, right top). Harvesting software may aggregate this meta-information to compile an "open directory" of site descriptions, and the meta-information can be fed into search engines.

HIDDEL/MedPICS labels as benchmarking indicators for measuring achievements in public (e-)health

The U.S. Department of Health and Human Services (HHS) has articulated the need for website "labels" in its report Wired for Health and Well-Being [13], though it not necessarily referred to meta-data. The HHS also has made it a national health objective to increase the number of health Web sites that disclose certain information [4; 14] With sites using electronic HIDDEL/MedPICS labels, progress in the area can be easily measured and quantified. For example, automatic search engines may determine how many websites disclose the information demanded in the Healthy People 2010 document [14] as HIDDEL/MedPICS meta-data.

Client-side tools for downstream filtering

Another application scenario is the development of appropriate client-side tools which will enable users to set their own preferences, enabling client and host to "negotiate" (possibly also integrating third-party opinions from other sites), and to let the browser display warnings and alerts if a host does not comply to the personal preferences or "quality criteria" of the user (Fig. 3).

Conclusion

In summary, we think that widespread implementation of this vocabulary may enhance trust among consumers into networked health resources, enable e-health providers to publish disclosure information in a standardized way, assure interoperability of rating services and gateways, and promote an infrastructure which enables consumers to make informed decisions. This "quality" management approach is – other than many "seal of approval" approaches - perhaps the first approach which is able to take into account that different user preferences and needs inevitably lead to different notions of what actually constitutes "quality".

A link to the most recent version of HIDDEL/MedPICS as well as additional documentation can be found at http://www.medcertain.org/metadata/index.htm

Acknowledgements

Partly funded by the European Union, Action Plan for Safer Use of the Internet, MedCERTAIN project.

References

1. Swick RR., Brickley D. PICS Rating Vocabularies in XML/RDF. 27-3-2000. URL: http://www.w3.org/TR/rdf-pics

2. Weibel, S., Kunze, J., Lagoze, C., and Wolf, M. Dublin Core Metadata for Resource Discovery. Internet RFC 2413. RFC2413 . 2000. URL: http://www.ietf.org/rfc/rfc2413.txt

3. Malet G, Munoz F, Appleyard R, Hersh W. A model for enhancing Internet medical document retrieval with "medical core metadata.". *J Am Med Inform Assoc* 1999;6:163-172.

4. Baur C, Deering MJ. Proposed Frameworks to Improve the Quality of Health Web Sites: Review. *MedGenMed* 2000;2: URL: http://www.medscape.com/Medscape/GeneralMedicine/journal/2000/v02.n05/mgm0926.baur/pnt-mgm0926.baur.html

5. Eysenbach G, Sa ER, Kuss O, Diepgen TL. A framework for evaluating e-health: Systematic review of empirical studies assessing the quality of health information and services for patients on the Internet. *submitted* 2001;

6. Eysenbach G, Diepgen TL. Labeling and filtering of medical information on the Internet. *Methods Inf Med* 1999;**38**:80-88. URL: http://www.yi.com/home/EysenbachGunther/publications/1999/Eysenbach1999e_MethInfMed.pdf

7. Eysenbach G, Diepgen TL. Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. *BMJ* 1998;**317**:1496-1500. URL: http://www.bmj.com/cgi/content/full/317/7171/1496

8. e-Health Ethics Initiative. e-Health Code of Ethics. *J Med Internet Res* 2000;2:e9. URL: http://www.jmir.org/2000/2/e9/

9. Boyer C, Selby M, Scherrer JR, Appel RD. The Health On the Net Code of Conduct for medical and health Websites. *Comput Biol Med* 1998;**28**:603-610.

10. Hi-Ethics Group. Health Internet Ethics: Ethical Principles For Offering Internet Health Services to Consumers. HIETHICS . 2000. 3-12-2000. URL: http://www.hiethics.org/Principles/

11. Winker MA, Flanagin A, Chi-Lum B, et al. Guidelines for Medical and Health Information Sites on the Internet. *JAMA* 2000;**283**:1600-1606. URL: http://jama.ama-assn.org/issues/v283n12/full/jsc00054.html

12. Eysenbach G, Yihune G, Lampe K, Cross P, Brickley D. MedCERTAIN: Quality Management, Certification and Rating of Health Information on the Net. *Proc AMIA Symp* 2000;230-234.

13. Science Panel on Interactive Communication and Health. Wired for health and well-being; the emergence of interactive health communication. Eng, T. R. and Gustafson, D. H. 1999. Washington, DC, US Department of Health and Human Services, US Government Printing Office.

14. Healthy People 2010: Understanding and Improving Health (Second Edition). 017-001-00547-9. 2000. Washington,DC, Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services. URL: http://www.health.gov/healthypeople/document/