On Management of the Health Content Lifecycle

Hamman W. Samuel[1], Osmar R. Zaïane[2]

Dept. of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[1]hwsamuel@cs.ualberta.ca, [2]zaiane@cs.ualberta.ca

***Abstract*-The Internet is an ideal tool for promoting public health goals of prolonging life, health, and improving the quality of life. There are many websites with health-related information where one can go to as an information source, for health advice, or self-diagnosis. However, these health websites require a more acute awareness of ethical issues due to potential life threatening risks from misuse of information. Providing disclaimers and accreditation logos only goes so far in covering potential legal conflicts, but fulfilling ethical obligations for non-maleficence requires more action on our part. As such, the content lifecycle of these websites requires greater emphasis on privacy, security, and trustworthiness. We propose and give a high-level description of a Health Content Management System (HCMS) that addresses both the managerial, as well as the ethical issues with health content. Surveys of existing health websites and content management systems demonstrate the need for the proposed system. Moreover, the novelty of the proposed HCMS is appraised and asserted in comparison with similar health framework concepts. Our contributions include survey results of more than 50 health websites, taxonomy of health websites' characteristics, discussion about legal versus ethical obligations, and a blueprint for typical and novel features for health websites. Moreover, this study presents a new approach to analysing health content via lifecycles.**

Keywords-Ethics; Trust; Medical; Websites; CMS; CMF; Review; Disclaimer; Liability

I. INTRODUCTION

Because of the proliferation of the Internet, the online webscape is ideal in promoting public health and its goals of prolonging life, health, and improving the quality of life. Among the myriad websites on the World Wide Web, there is a considerable number of websites that are dedicated to providing information on health-related topics. While websites like Facebook or New York Times may contain health-related topics, the scope of this study is focused on *health websites*, which are defined as websites that exclusively contain health content, such as WebMD.com. Moreover, within this category of websites with only health content, we further narrow down this study to health websites by content authorship. Owner-Engineered Content (OEC)is managed by the website owners and undergoes a content creation, editing and approval workflow, or content lifecycle. Typical examples include health blogs and health article websites, where a domain expert is part of the content management workflow. Also, certain websites that allow registered users, who are not domain experts, to share their views and thoughts, thereby contributing to content and creating User-Generated Content (UGC). These websites are termed as social media, and UGC follows a different lifecycle from OEC. This study identifies and differentiates between these lifecycles within the health context and in terms of content authorship.

Ultimately, health websites are significant in the light that health is a popularly searched topic on the Internet. The Pew Internet and American Life project surveys have shown a steady growth in popularity of the Internet as a resource for getting health information. In 2005, 8 in 10 Internet users had searched for health topics [4], while recent statistics show that 4 in 5 Internet users have sought healthcare information online [5].It should be noted that this survey focused on all internet resources, but at the same time, searches of health topics from most popular search engines usually had health websites at the top of their search results. Undoubtedly, there is a growing trend of using online media for getting health advice, information and self-diagnosis from specific niche websites like WebMD.com. But with this popularity comes life-changing and life-threatening situations because health-related websites contain information that may potentially be harmful. Moreover, disclaimers in terms of service are not preventing users from using health websites to get medical advice. This makes health websites distinct from other categories of websites.

The potentially critical nature of health websites leads to challenges in how health content lifecycles should be managed. A typical website requires knowledge in HTML and other web technologies, while its content lifecycle may be less stringent in terms of credibility, timeliness, and accuracy of information. In contrast, health websites contain life-impacting and sensitive information which, if misused or miscommunicated, could be life-threatening. As an example, a US-based website called ApricotsFromGod.info continues to exist and claims cures for cancer, even though Food and Drug Administration (FDA) has not approved their claims [2].According to web traffic statistics as of 2012, ApricotsFromGod.info had 1,482 visitors in the past year (according to Compete.com), who were exposed to the risk of taking incorrect and untested health advice. Because of the relative ease with which websites can be set up and created, there is the potential that many health websites may not be trustworthy.

The majority of websites may not have the same level of stringency as health websites. For instance, a user visits a computer gaming website which has some advice about a particular game's strategy. Contrast this with a patient taking some advice about a drug from a health website. The effects of the accuracy of advice given on these two websites are distinct, from non-critical to critical. Also, suppose a patient shares his/he rmedical conditions on a health social network with a select group

DOI:10.5963/PHF0202005

of users. A breach in security or privacy is more severe in this situation, compared to a typical social network with someone sharing personal stories about what they did last night, for instance. Consequently, health content management requires an awareness of potential ethical issues.

We propose a Health Content Management System (HCMS) with two-fold capabilities: 1) meeting the functional content lifecycle requirements typical of all websites; and 2) covering ethical aspects, such as privacy, security, and trust. In this study, high-level functional requirements of the proposed HCMS are outlined, and implementation pathways are investigated. Moreover, the novelty of the proposed HCMS is appraised and examined. Our contributions include survey results of more than 50 health websites, a taxonomy of health websites' characteristics, and a blueprint for typical and novel features for health websites. Moreover, this study presents an overarching ethics perspective on health content lifecycles.

The paper is organized as follows. In Sections 1 and 2 we provide background information on concepts used in this study, while in Section 3 we examine existing systems and procedures to ensure safe health information. In Section 4, we outline functional specifications for our proposed system, and in Section 5, an outline of implementation and evaluation of the proposed system is described.

II. BACKGROUND

A. Content Lifecycle

Content lifecycle, or more specifically web content lifecycle is a term used to describe various stages of managing information on a website. There are various paradigms and views about how many stages are involved in managing the content, the simplest being the two-stage life cycle. In the two-stage lifecycle proposed by McKeever [25], the first stage involves content being created, while the second stage involves this created content being delivered or published to the intended audience. Another possible view by McGovern et al. [27] involves three stages, in which content goes through a creation, editing, and publishing cycle. In the four-stage view, the content lifecycle involves creating, reviewing, managing, and delivering content. Rockley [26] adds an additional stage and defines content lifecycle as create, edit, manage, and publish.

While there are also many other views, this study adopts the five-stage lifecycle methodology: create, approve, publish, unpublish, and archive [28]. This methodology is selected because it best reflects the health content lifecycle, where content does get outdated and needs to be unpublished, archived, or deleted. Moreover, the five-stage approach includes an approval step, which is an important aspect for ensuring credibility of health content. Figure 1 shows an overview of the five-stage lifecycle.It should also be noted that not all steps might be followed in this lifecycle when it comes to content generated by registered site users, UGC,versus content created by the website owners, OEC. For instance, most social networks and forum discussions may have some partial moderation of content, but other steps such as editing or approving may not be followed because the site owners do not assume liability for content generated by users.

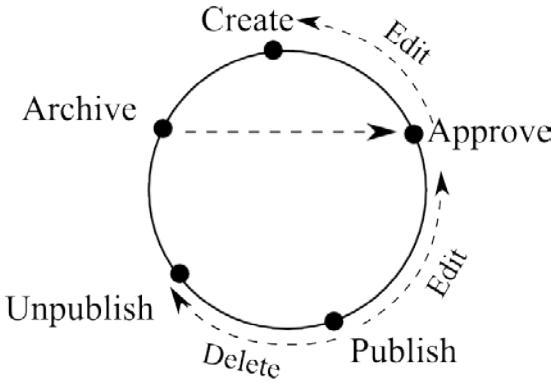

Figure 1 Five-stage Content Lifecycle

B. Content Management

The different lifecycle tasks are performed on content, such as text, images, audio, video, animation, etc. [6]. The mechanisms through which the various stages of the content lifecycle are achieved and realized is referred to as content management, while the system or sets of tools that perform content management are called content management systems. Thousands of CMS are available on the Internet, supporting different platforms, technologies, and licensing [13]. CMS provide an unobtrusive mechanism for creating websites. The basic and common functionalities required of most websites are present by default. CMS serve as blueprints for websites. In addition, most CMS can be expanded with plugins to include additional functions such as polls, galleries, surveys, forums, blogs, wikis, journals, to mention a few. A majority of CMS come with these additional features built-in. Essentially, CMS handle the presentation front-end of websites from a back-end interface that hides some of the complexities of web technologies.

DOI:10.5963/PHF0202005

In addition to CMS, there are also Content Management Frameworks (CMF), a term sometimes used interchangeably with web application frameworks. CMF are more streamlined and bare-bones versions of CMS that facilitate the development of websites that require more flexibility in terms of functionality. In essence, CMF are more elementary, customizable, and configurable than CMS [23, 19]. While this is generally the case, many CMF actually double-up as CMS, and are full-featured. A simple illustration of these relationships for CMF versus CMS in general can be given using set notation. Let $f \rightarrow x$ be functional attributes in x. Then, $f \rightarrow CMF \subseteq f \rightarrow CMS$.

C. Health Informatics Ethics

Ethics deals with decisions about right versus wrong, good versus bad. These normative and moral questions involve people and how they affect each other [29]. An ethical conflict is opposition between moral ideas and interests. In the field of medicine, ethical issues have been well-known due to the nature of the profession in dealing with lives. Consequently, health personnel are trained in ethics as part of their education. In addition, health personnel ought to adhere to codes of ethics and regulations which aim to reduce the risks of ethical conflicts. For instance, the Hippocratic Oath taken by medical practitioners puts the principle of doing no harm first [30, 31, 32]. However, as medicine has evolved, so have the ethics principles. For instance, the Nuremberg code for human medical test subjects was developed to avert future occurrences of medical experiments and atrocities conducted in World War II [33]. The introduction and adoption of information technologies into medical care created new ethical conflicts due to new stakeholders becoming involved. Instead of interactions between patients and physicians, computers involved system developers, information managers, and other non-domain stakeholders as part of the overall process of providing medical care [32].

Ethics issues in health informatics are referred to as health informatics ethics. Health informatics ethics or medical informatics ethics has resulted from various evolutions and overlaps in medicine, information and ethics. In the late 1940s, Norbet Wiener, considered the father of cybernetics, looked at the socio-ethical relationship between machines and humans, which can be considered as a prelude to the overlap between informatics and ethics [30, 33]. In the 1970s, Kostrewski and Oppenheim, and Robert Hauptman looked at various questions regarding ethics in information research. By the late 1990s, Severson had already identified core principles of information ethics encouraging respect for ownership and privacy of information,as well as non-maleficence [33].

As computing incorporated ethics, the medical profession was also embracing the use of technology in healthcare with services such as computer-based patient records, decision support systems, telemedicine, e-health, and so on [31, 34]. As Goodman-Miller pointed out, the applicability of the Hippocratic Oath could no longer be restricted to medical practitioners, but had to include developers of these computer-based medical systems [32]. As the stakeholders became more aware of the notion of health informatics ethics, solutions were proposed to handle ethics issues, such as regulations, standards and guidelines, honor codes, and codes of ethics [31].There are various codes of ethics that address how information related to healthcare topics needs to be managed. Codes are developed by health informatics organizations such as the International Medical Informatics Association (IMIA), or the European Federation for Medical Informatics, among others.

Looking at the various dominant codes from these bodies, health informatics ethics can be categorized into three areas: medical ethics, informatics ethics, and software engineering ethics. Medical ethics refers to ethical and normative expectations in general of health practitioners, while informatics ethics define ethical principles, such as privacy, when information is involved. Moreover, software engineering ethics covers ethics principles, such as disclosing any dangers or known defects in software, and technical guidelines for building systems. Medical ethics includes non-malfeasance, integrity, equality and justice, and beneficence, among others. Issues covering informatics ethics include privacy, openness, security, accountability, and so forth. Lastly, software engineering ethics refer to normative expectations of developers towards society, employees, profession, client, and self. These three categorizations of ethics principles are meant to mitigate harm to users, if followed, when health-related systems are being built and deployed.

Health informatics ethics can be defined in terms of *principles* in the relevant codes of ethics, as well as *activities* are carried out by health informatics professionals to ensure ethics principles are adhered to [18]. These activities are outlined in technical guidelines such as the International Standards Organization (ISO) standards on health informatics set forth by its Technical Committee ISO/TC 215, which address security and safety of health software, health devices, and privacy issues with health information [35].The U.S. Food and Drug Administration (FDA)'s provides documents and guidelines for validating software and medical devices so that they are marked as safe and secure. Two pertinent documents by the FDA are "General principles of software validation" [36] and "Guidance for the content of premarket submissions for software contained in medical devices" [37]. In addition, the U.S. Department of Health and Human Services (HHS) references the "Health Insurance Portability and Accountability Act (HIPAA)" that deals with privacy and security of health data via the privacy and security rules which give individuals the rights of confidentiality and accuracy of their personal information [38].

DOI:10.5963/PHF0202005

III. EXISTING SYSTEMS

In this section we look at what is being done to ensure that health information is being safely used. Moreover, we examine existing health frameworks to evaluate the awareness of health informatics ethics within their content lifecycles.

A. Existing Online Safety Measures

It has become fundamentally easy for anyone with internet access to be able to develop a website or create a blog. This openness has led to an explosion of information. Health-related information has also been affected by this trend as many patients have gone online to share their health experiences with others. The availability of health information does have positive impact in building community and encouraging like-minded patients to make informed decisions about their health. However, this openness also has the potential to provide harmful information that may be outdated or incorrect [34].

The effects of bad health information seem intuitively obvious, in which a suggested cure leads to a worsening condition or unhealthy side-effects. However, bad information can also be indirectly harmful. For instance, subscribers of websites such as ApricotsFromGod.com or CancerAnswer.com may be foregoing other well-established medical treatments in lieu of unverified claims [2]. The term 'information', synonymous with 'content', needs to be clarified here in the e-health context. This is because there is a distinction between health websites that can be used for getting health advice and any other perceived health-related websites, such as one for a health organization. Both these websites contain information, but can be classified as quasi-critical or non-critical respectively. The quasi-critical nature of information in e-health comes from the likelihood of patients using it for self-diagnosis, treatment, or even challenge a physician's advice, which can be dangerous [17, 39]. Non-critical information, on the other hand, is for promotional purposes. For example, a non-critical website might be a web presence for a hospital, with details such as the names of the staff and the hospital location. Because this information is promotional and does not directly contain health advice, it can be seen as non-critical in the context of health content.

However, it should be noted that this distinction is often not so clear-cut. Lagu et al. carried out a study of web blogs supposedly written by health professionals. While these blogs did not contain health advice, there were endorsements made for products, drugs, and even other health professionals [39]. This sort of endorsement information could indirectly affect health decisions made by visitors to these blogs and should fall under the quasi-critical category.

Compliance with informatics ethics principles such as privacy, security, access, accountability, etc. need to be addressed within such websites [18]. At first glance, it seems this is being done. An arbitrary survey of popular health websites showed that many health websites display accreditation logos to supposedly show their compliance with ethical standards. However, a recent study found that only 66% of the websites that displayed the HONcode logo of the Health on the Net Foundation were actually in full compliance [8]. It is very likely that the remainder of these websites only showed the logo to create an illusion of credibility. Moreover, many health websites do not assume liability for content that registered users generate on their websites because the users are possibly not domain experts. These websites usually provide disclaimers in their terms of service and leave the burden of verification to the end-users. Despite these trend, statistics show that patients and even physicians are trusting online resources [17, 9, 22]. Consequently, there is an urgent need for handling ethical concerns effectively.

B. Health Frameworks

In the e-health domain, there have been attempts at building dedicated frameworks for healthcare information systems. For instance, Microsoft developed an architecture and design blueprint in 2006 that describes a Connected Health Framework (CHF), which was updated in 2009 [12]. Also, UK-based companies Genetics Ltd. and EIBS Ltd. have targeted their CMS, Contensis and Easy Site respectively to National Health Services (NHS) and healthcare organizations in the UK since early 2000 [11, 10]. Furthermore, a US-based company, Greystone.Net has been providing e-solutions and consultations for healthcare organizations since 1996. Their services include building websites for these organizations [3, 7].

C. Caveat Lector

Even though these existing health frameworks are available, certain factors need to be taken into consideration. Microsoft's CHF is not a programming or development environment, but a specification. In addition, the CHF is focused on patient medical records, but does not address healthcare in terms of quasi-critical information on health topics. This is markedly different from the uses of patient records. Furthermore, the CHF takes a more business-oriented perspective on healthcare, which is only a subset of the interactions that take place in e-health.

The existing CMS focused towards health, such as Contensis and EasySite are focusing solely on organizational content of the NHS and healthcare organizations. There is no health advice being shared on these websites. Consequently, these are not websites that patients would use to get health advice or information on a health topic. Furthermore, Greystone.Net only provides advice in selection of a CMS, and does not state any in-house developed CMS/CMF focused on quasi-critical health websites.

DOI:10.5963/PHF0202005

IV. FUNCTIONAL SPECIFICATIONS

Given these pending issues, we propose a solution through a domain-specific content management system for building health websites in the context of health content. We refer to this content management system as a Health Content Management System (HCMS). The features expected of this HCMS need to provide core functionality to manage content lifecycles, which is expected of any website in general. Also, the stakeholders need to be identified, and their interactions with each other determined in order to understand requirements. It is equally important to examine existing health websites to get a better understanding of features specific to health websites, in contrast with general websites. Accordingly, a survey of over 50 health websites was conducted. The sampled websites were selected randomly, but with attention to enough variance in terms of popularity and search rankings.

A. Typical Health Websites Profile

Among the over 50 websites surveyed, the majority of health websites included user blogs, discussion forums or community groups, health articles, and outbound links to other websites. In addition, various websites provided health tools such as medical terms dictionaries, symptom checker, doctor search, and health trackers with personal diaries and journals.It should be noted that the health websites surveyed did not include health discussions on more general-domain social media such as Facebook, Twitter, LinkedIn, YouTube, etc.

Overall, the 3 categories of websites were identified: 1) General Topic (GT), 2) Specialized Topic (ST), 3) Specialized Demographics (SD). Health websites can be for a wide range of general health topics, diseases, or drugs. Some are more specialized for particular diseases or topics, such as websites for Cancer, HIV patients. For instance, PatientsLikeMe.com is for patients diagnosed with life-changing diseases such as epilepsy, Parkinson's, fibromyalgia, among others. Websites can also be targeting certain demographic groups such as men, women, seniors, or kids.

In terms of content, 3 methods of how content is generated were distinguished: 1) Owner Engineered Content (OEC) generated by the website administrators/staff, 2) User-Generated Content (UGC) generated by registered users, and 3) External Aggregated Content (EAC) from other websites.It was also observed that most of the surveyed websites provided disclaimers regarding health advice with sentences such as "The Site Does Not Provide Medical Advice". This raises an interesting conflict of ethical and legal interests. Though these health websites seem to be legally free from liability by saying they do not provide medical advice, statistics show that people who visit these websites are in fact getting health-related advice. Nevertheless, legalistic "hands washing" does not equate to ethical obligations, as ethically the site owners ought to do more to ensure safety and trustworthiness of their content.

Figure 2 shows the frequency of occurrence of the communication types, categorizations, and content types. Based on the survey statistics, 51% of the health websites surveyed address general topics. These topics are generated for the most part via OEC, which accounts for 46% of the websites surveyed. In addition, 56% of content was for B2X communication.

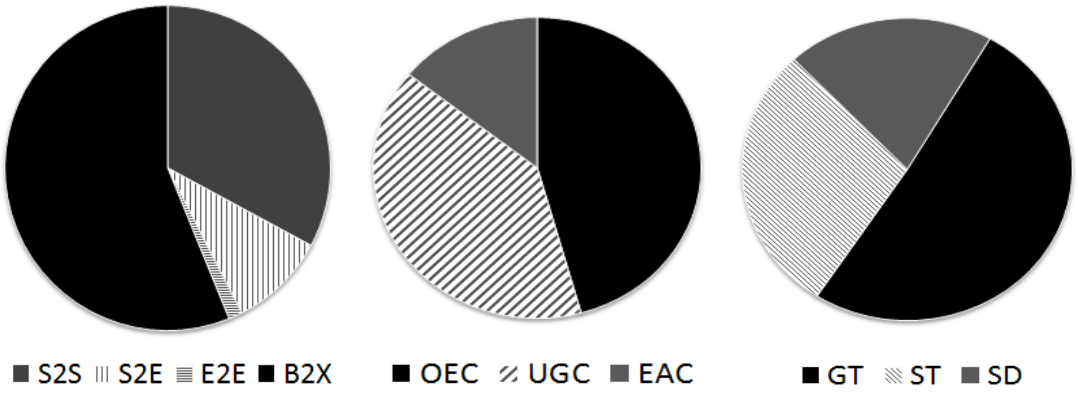

Figure 2 Relative frequency of occurrence of common properties

B. Stakeholders

Stakeholders identified were in the context of health advice. From the survey of the health websites, there were two main users identified in this scenario: subjects and experts. 'Subjects' refers to patients or anyone seeking health information online. On the other hand, 'experts' implies physicians, users with experience in healthcare, and domain experts in health. In terms of communication, while it is normally thought that patients receive advice from physicians, it is also common for patients to take advice from others who have gone through similar situations. Three possible scenarios of how two-way communication happens between these stakeholders were identified: 1) Subject-to-Subject (S2S), 2) Subject-to-Expert (S2E), and 3) Expert-to-Expert (E2E). A fourth mode of one-way communication was also identified: Broadcast-to-any (B2X), in which the subject or expert releases information for general consumption.

DOI:10.5963/PHF0202005

C. Lifecycle Stages

In theory, there are critical parts of the content lifecycle that need to be highlighted, especially when looking at health content. As depicted in Figure 1, after content is created, it may go through various iterative reviews and edits before it gets approved. After being eventually approved, the content would be published, but may at some point need editing or become expired, which requires re-approval. On the other hand, published content may also be unpublished, either via deleting it permanently, or hiding it from view, or archival. Aside from deletion, which ends the content's lifecycle, content may be re-instated, which would require approval again. This five-stage lifecycle becomes ideal for a health website because approval is a critical step, which ensures that content is accurate, up-to-date, and free of obvious misinformation. Incorporating this five-step process in the editorial policies of a health website would be one approach to ensuring content accuracy. However, these features can equally be automated and built into the HCMS. For instance, a threshold value of expiry time for articles can be set which would require re-approval of the articles to ensure no outdated information is provided. Similarly, more stringent workflows can be built for the publishing step, so that effective moderation is available when user-generated content is in question.

These scenarios are ideal for Owner-Engineered Content (OEC) because the website owners have a vested interest, as well as legal and ethical obligations to ensure their content is accurate and trustworthy. Because the owners are marketing the content on their website, content credibility is in their best interests. However, in practice the lifecycle stages quickly become blurry. In the first case, users who are generating the content might not be the website owners. This is a common trend among various health websites with discussion forums. The content on websites like SteadyHealth.com and PatientsLikeMe.com is mostly generated by registered users who are usually not medical experts and merely share their opinions and experiences. The process of edits and approvals is no longer applicable because of the volume of users and also the ratio of medical peer reviewers to non-medical users. It takes a lot of work hours and additional staffing costs to be able to moderate UGC which is being created faster due to its opinionated nature. Instead of having an iterative editing process, some websites rely on content moderation to filter out offensive words and arguments, but this leaves a lot to be desired to ensure content quality. As a solution, the majority of health websites have opted to post disclaimers so that they are not liable for UGC.

Another interesting case arises with OEC where the site owner expresses personal opinions and experiences, such as personal blogs and websites like ApricotsFromGod.info. Content in this case is self-created and self-edited, and the lifecycle stages are not followed. This content, even though being maintained by the owner of the website, is by nature more like UGC because it is personal and undergoes no peer reviews. In essence, self-review does not count as a review. Lagu et al. noted that even health professionals who have personal blogs do not often follow ethics practices such as privacy. In their survey, various health practitioners inadvertently revealed information about their patients when blogging [39].

For UGC, the content lifecycles in practice involve either partial reviewing or community reviewing. Partial reviewing involves health domain experts as moderators who edit and approve content. For instance, the WebMD.com discussion boards feature staff that occasionally inspect conversations and discussions. On the other hand, community reviewing involves feedback from other users. This feedback can be ratings given by other users on the content posted. For example, in the PatientsLikeMe.com community, users can receive "helpful marks" from other users to show that their discussions and posts were useful. These sort of ratings can be seen as partial reviews of the content that somewhat help other users gauge the credibility of content. Ratings are also available on personal blogs where readers can provide star ratings or helpful/not helpful votes.

D. Novel Features

Features listed so far rely on finding commonalities across the health websites surveyed. However, there are also features that are not common for a significant majority of health websites. These novel features include any innovative functionality that is either not common, or not yet provided in existing health websites.

1) Extensible Adaptability:

The HCMS will be based on an extensible component-based architecture that allows creation of new functionalities via new components. Also, the HCMS will be extensible to different platform technologies such as desktop computers, mobile devices, and cloud infrastructures to support the ubiquitous computing philosophy. This is a novel feature and will allow health advice to be accessed in various ways.

2) Community Knowledge:

The HCMS will be able to leverage existing content from other sources, categorized into 2 groups: health content and health literature. For health content, external feeds from other health websites will be used to show related information and links. In the absence of feeds, web content extraction components will be available to directly query web pages and retrieve content in some structured manner. The selection of source is very important because without some credibility associated with the source, the content being aggregated would lead to liability and risks. Hospital-affiliated websites that provide information are obvious candidates for these feeds, for instance Mayo Clinic [40] and Health Link Alberta [41] and other similar websites. With health literature, information about health articles related to user discussion topics will be shown using Medical Subject

DOI:10.5963/PHF0202005

Headings (MeSH) lookup via reliable health knowledge sources such as PubMed [14].

3) Assisted Advising:

A novel feature in HCMF will be components to allow securely importing and maintaining Personal Health Records (PHRs) for patients. Patients will be able to opt to privately share these records with experts in S2E interactions. In turn, experts will be better informed through PHRs before giving advice. There are various websites dedicated to PHRs, such as Microsoft HealthVault and HelloHealth.com, but this is a unique feature to incorporate PHRs with health advice sharing scenarios.

4) Federated Communication:

The survey results of existing websites showed that the B2X communication mode is most popular, while E2E communication is the least common. The HCMS will facilitate all communications between the stakeholders. Physician experts will be able to communicate with each other in a private and restricted environment. Expert wikis will also be available for shared knowledge that the experts may wish to share with everyone at some point. Patient subjects will be able to communicate with each other to share knowledge and real-world experiences. Subjects will also be able to communicate with experts to get health advice.This also adds value and quality to the UGC lifecycle because the interaction allows medical experts to review the content and provide feedback. Admittedly, this scenario is a departure from traditional evidence-based medical practices. However, the alternative is to let users interact with non-medical experts who might not be in the best position to give feedback.

The survey results showed that real-time chatting via instant messaging is also a rare feature. Federated communication will also allow message exchange via both Static Asynchronous Messaging (SAM) and Dynamic Synchronous Messaging (DSM). SAM will be available via mail messages and post responses. Mail messages are private messages exchanged between users on the website, while post responses include comments made to blog posts, microblog posts, and forum threads. DSM will be available via real-time instant messages with logging. A simplistic view of the federated communication model is shown in Figure 3. Directionality of communication is specified to give a relative denotation of the stakeholder that can initiate the communication.

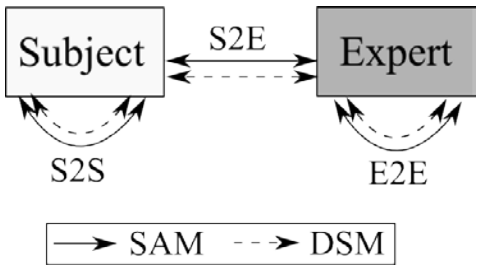

Figure 3 Health advice sharing among stakeholders

5) Privacy-Preserving Data Mining (PPDM):

A novel feature is to incorporate multi-faceted data mining functionality within the HCMS. Various data mining tools will be built into the HCMS via its EA feature. Within data mining tasks, the need for confidentiality of data is a big issue, and privacy preservation is a critical task [1, 15]. Consequently, privacy preservation components will be added to ensure sensitive data are kept confidential.

There are different aspects of data mining in relation to the World Wide Web, such as web content mining, web usage mining, and social data mining. Web content mining deals with content search and retrieval while web usage mining involves analyzing user access [24]. Social data mining and social network analysis (SNA) deals with patterns of interactions between people [16]. The HCMS PPDM feature will cover these different aspects of web mining, and serve two sets of users: internal and external.

Internal users, or the users of the website will be able to get intelligent statistics and analysis based on the content and interactions. Very few existing health websites match content with related drugs and other similar content. A recommender component in HCMS will provide users automatically categorized predictions of related topics based on what they type. Also, text summarizations of discussion treads, articles, and other long-length content will be available via machine learning techniques.

External users, or users not involved in content generation, such as researchers will be able to retrieve anonymized data from the HCMS databases for use in their data mining tasks. Web services will be built in the HCMS that will facilitate data retrieval protocols over secure connections. In addition, anonymization components will be built into the HCMS to ensure that data privacy is preserved.

It is noted that internal users will require real-time PPDM, and specific algorithms for real-time data mining will be required to provide dependable results [21].

DOI:10.5963/PHF0202005

6) Trust Metrics:

Although there are dedicated websites for ratings of doctors and practitioners, such as RateMDs.com, RateMyMD.ca, DoctorScoreCard.com, DoctorRate.com, and HealthGrades.com, nearly all health websites lacked incorporation of some trust metrics that give the user an indication of how trustworthy the content is. A novel feature in the HCMS will be indicators for trust for content based on a community-based ranking system combined with other analytical results.This will be important to providing a mechanism for UGC lifecycle management in situations where experts opt not to interact with subjects. For instance, content that receives a lot of negative votes from other users could be unpublished or marked for editing, thereby helping moderators hone in on critical content.

E. Taxonomy of Health Websites

The results above allow for a generalization of high-level properties of typical health websites. We propose a possible taxonomy shown in Figure 4 to summarize these properties. The taxonomy also includes novel features that would be available in HCMS.

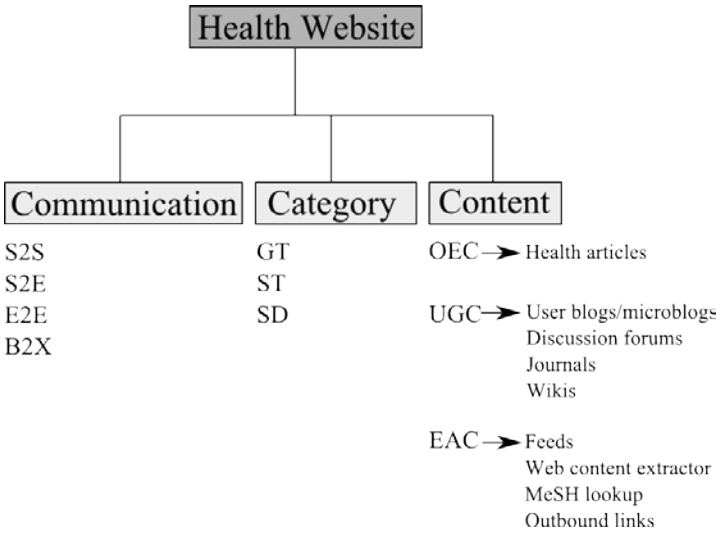

Figure 4 Taxonomy of health websites

V. IMPLEMENTATION AND EVALUATION

In this section, a high-level view of how the system will be implemented is provided, as well as an outline of evaluation through implementation and pilot study.

A. Information-Centric View of Health Websites

The core of health advice sharing websites is the information or content. In formulating a design for the HCMS, we propose an Information Block(IB) for capturing information requirements of different types of health websites. An IB can be seen as a template that encapsulates the content and navigation to content on a website. The notion of content used here is from the end-user's perspective, while the exact representation and rendering of the content is delegated to the presentation aspects. The following specification describes the IB. Different types of websites require different vocabularies and content arrangements. For instance, a forum has communities, sub-communities, and threaded discussions, while a blog has pages, posts, and categories. Also, a web page has a page title and content, as well as the page's position in the site map. Furthermore, social networks have notions of walls and news feeds. All these arrangement and vocabulary contracts are hard met by a single web framework.

However, the terms "post", "thread", "comment", and so on have structural commonalities, even though different vocabulary is used to refer to them. We identify three aspects that are common to the navigation items and content, namely *caption*, *description*, and *metadata*. The caption is usually a heading for an article topic, forum post, blog post, message, etc. The description is content matter of the IB, which may be HTML or textual. Rendering of HTML is not the responsibility of an IB, and is delegated to the client-side. Finally, metadata refers to any additional information about the IB, such as authorship, timestamp, etc.

We claim that all textual data on health websites can fit this generalization. A navigation menu item may have only the caption, while a web page would have both a caption, represented by the page title, and a description, represented by the content. The metadata can contain other information such as timestamp or authorship. An IB can be grouped and arranged in hierarchies to depict levels and clustering common in most navigational representations. Also, different vocabulary words can be used to describe content, e.g. post, thread, comment, message, etc.

DOI:10.5963/PHF0202005

It follows that a blog's content can be represented as an IB, as well as a threaded discussion post. In addition, the hierarchy of the communities and sub-communities can be represented by the IB as well. An IB not only represents the readable content on a web page, but also the menu navigation and respective hierarchies and groupings. Consequently, IBs can serve as the building blocks for the generic HCMS framework to create different forms of websites with their own vocabularies about content.

Given an IB as the unit content to the website, we can define the term 'quasi-critical information blocks' or QIB as independent entities in the content that have some related ethical concerns. QIB can be a web page, a blog post, an image, video, a forum post and so on. QIBs can also be combined together to re-form a new QIB. For instance, a group of images, text and video can all be present in one blog post, which can be a QIB. QIB can also have meta-data attached, such as the title of the blog, author, or date created. The distinguishing characteristic is the need to address ethical concerns because there is a likelihood of a QIB having potentially misleading information.

B. System Architecture

A high-level view of the overall concept and architecture of the HCMS is briefly introduced. Figure 5 shows a relationship between different hardware abstraction levels and software layers. This depiction is meant to compartmentalize and summarize where functional requirements will be implemented within the hardware-software context.

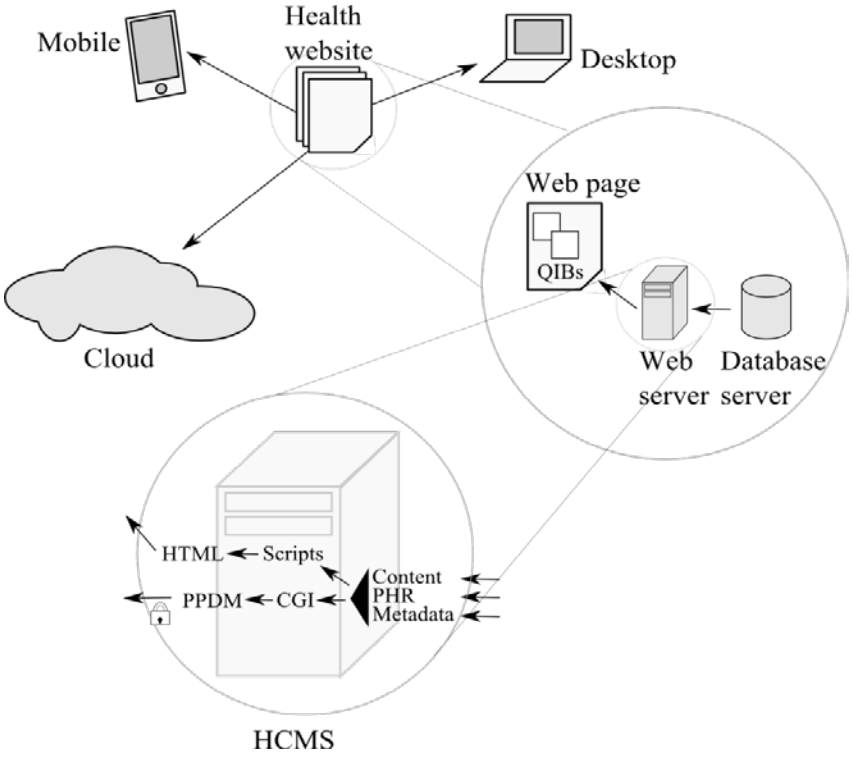

Figure 5 HCMS architecture

C. System Implementation

An investigation of existing CMFs was carried out using an online database of CMSs, CMSMatrix.org (CMSM). This was done to verify if there are CMFs that can be customized to provide the functions of an HCMS. The CMSM database contains various feature sets for describing each CMS.

The survey of CMFs using the CMSM database involved 3 steps. Firstly, features from the database that were relevant to the HCMS were selected. Next, CMFs were randomly selected based on their availability in the CMSM database. However, only CMFs that double-up as CMSs were used because the feature set of pure CMFs is overly limited. For instance, pure CMFs do not have blog or forum components, and these need to be re-written from scratch, which is not necessary in this case. Finally, a Match Percentage (MP) was computed for random selections of CMFs. MP is calculated as follows.

$$MP = \frac{\sum 1 \mid f\ CMSM_i = K_i}{n(K)} \times 100\%$$

Since $f \rightarrow x$ gives the features of x, the results of $K = f \rightarrow$ CMSM $\cap f \rightarrow HCMS$ are shown in Figure 6.

DOI:10.5963/PHF0202005

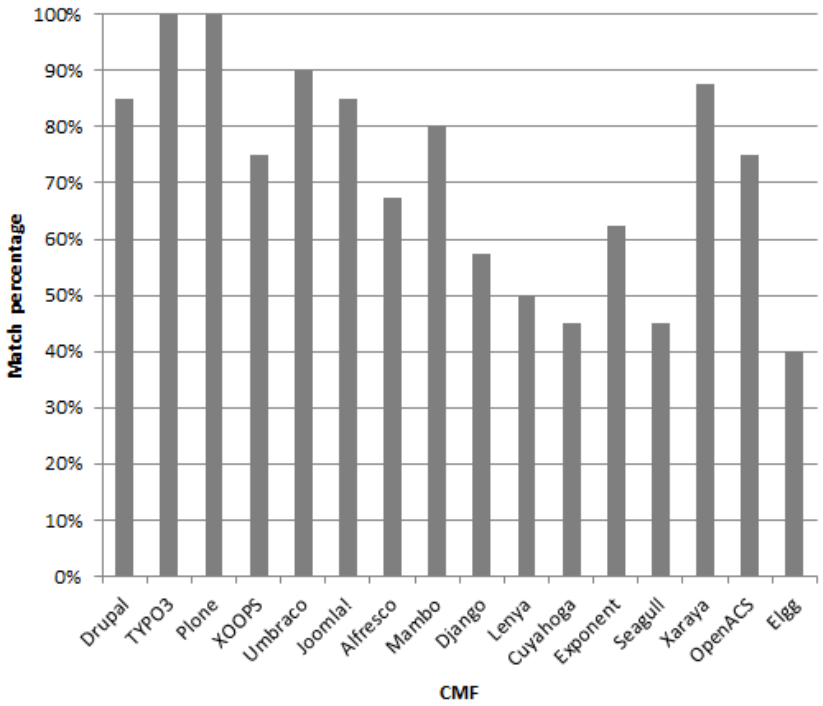

Figure 6Match percentage of popular CMF features with HCMS feature set

The results from Figure 6 seem very promising at first glance, as TYPO3 and Plone get 100% matches. However, the issue is that a 100% match does not necessarily mean that all the features of HCMS are readily available in TYPO3 or Plone. This is because of the way the result of $K = f \rightarrow CMSM \cap f \rightarrow HCMS$ is gotten. It is known that $n(K) \leq min(n(f \rightarrow CMSM), n(f \rightarrow HCMS))$. However, we note that $n(K) < n(HCMS)$. That is, the features within CMFs are actually less than the feature set of HCMS.

For example, features supporting federated communication are not available in TYPO3 and Plone. Setting up E2E communication requires private forums and restricted access, while S2S and S2E communications need their own separate space. This type of functionality will require installing multiple instances of TYPO3 or Plone, which will not be ideal for data mining and may lead to data redundancies. In addition, stringent privacy requirements for PHR data are not readily available. A more general illustration of the relationships between feature sets of CMFs, CMSs, and HCMS is given in Figure 7.

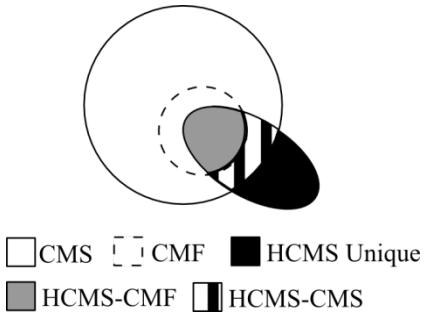

Figure 7Feature set comparison of CMF, CMS, and HCMS

It should be noted that the areas in Figure 7 are generalizations and not to scale. First, it is observed that $f \rightarrow HCMS \cap f \rightarrow CMF < f \rightarrow CMF$. This is significant because it shows that the HCMS does not require all the features of CMF. Furthermore, only a small subset of features of CMS is needed. It is also noted that HCMS has its own unique features, viz $(f \rightarrow HCMS) - (f \rightarrow CMS)$. Moreover, both CMFs and CMSs have features that are redundant to the HCMS.

Ultimately, CMFs are useful for managing routine tasks like rendering, content arrangement, and generic privacy and security concerns. However, CMFs need to be considerably extended to incorporate QIB requirements. Although HCMS has basic features similar to some CMFs, but the novel features are more than the CMFs provide. Consequently, we cannot translate the best selected CMFs directly into HCMS. The CMFs survey is a good starting point, but individual scrutiny of each CMF is needed.

D. Evaluation Plan

The health websites taxonomy in Figure 4 will be used for future evaluation of HCMS once implementation is completed.

DOI:10.5963/PHF0202005

Evaluation will involve two-steps: 1) using the HCMS to generate any particular type of website based on the taxonomy; and 2) monitoring the website generated with HCMS. For instance, would the HCMS be able to build a health social network (HSN) that has S2S and S2E features? Or could it generate a simpler health website with health articles? The second step in evaluation would be to actually monitor the websites created with HCMS. A pilot study will determine how users interact with HCMS websites, and will involve getting feedback from a sample of users, patients, and experts on the HCMS websites.

1) Anticipated Challenges:

A challenge in evaluation is getting enough user involvement. There are many mature health websites already with a lot of content. In contrast, a new website created with HCMS would have little or no content initially. This is a 'chicken-and-egg' problem. Experts and users are needed to generate content, but they may not be forthcoming. The community knowledge feature is expected to help, but the challenge in getting EAC is the choice of external websites. Questions of intellectual rights arise, as well as which website's content to trust. Whereas lack of content is an issue, in the presence of legacy content, there are new issues of how to import that content into the HCMS-generated website. These challenges need to be addressed in future work.

2) Sample Instantiations:

The following are simplified demonstration examples of the configuration of the IB to generate different types of health websites: article-based, blogs, wikis, forums, and communities. For each IB and QIB, appropriate privacy settings can be used to ensure authenticated and authorized access. Possible privacy parameters include the following.

- All: Anyone is allowed access.
- Guest: Only non-registered users are allowed access.
- Logged: Only users who are logged into the system are allowed access.
- Group: A specific group is allowed access. Typical groups include "friends".
- Owner: Only the IB author is allowed access.
- Admin: Only the website administrator/owner is allowed access.

Privacy scope per IB can be presented as a tuple that maps to the five-stage content lifecycle, {C, A, P, U, R} tuple, for Create, Approve, Publish, Unpublish, and Archive operations respectively. Each element in the tuple has a value equal to the privacy parameters denoted above. QIBs will have a more stringent setting in this tuple. In this notation, the publishing step implies who can read the published content after the approval step.

Article-based Health Websites

Article-based health websites will normally have a *menu* and *articles*, which defines the needed vocabularies. In addition, users may wish to *comment* on articles. For the first two vocabularies, it is normally the administrator who manages the content lifecycle. However, comments are either for registered users or for general public usage. Consequently, the privacy scope for each vocabulary is stated as follows.

- Menu: C{Admin}, A{Admin}, P{All}, U{Admin}, R{Admin}
- Article: C{Admin}, A{Admin}, P{All}, U{Admin}, R{Admin}
- Comment: C{All | Registered}, A{Admin}, P{All}, U{Admin}, R{Admin}

Health Blogs

For blogs, *posts* are usually arranged by *category*. In addition, users can *comment* on posts. Again, comments may be made by any user or by only registered users, depending on the admin's preference.

- Category: C{Admin}, A{Admin}, P{All}, U{Admin}, R{Admin}
- Post: C{Admin}, A{Admin}, P{All}, U{Admin}, R{Admin}
- Comment: C{All | Registered}, A{Admin}, P{All}, U{Admin}, R{Admin}

Health Wikis

The unique aspect of wikis is that many users can have editing privileges. Wikiposts can be arranged by category. For each wiki, all users may be allowed to contribute and create content. In a wiki, all users are involved in the content lifecycle.

- Category: C{Admin}, A{Admin}, P{All}, U{Admin}, R{Admin}
- Wiki Post: C{All}, A{All}, P{All}, U{Admin}, R{Admin}

Health Forums

Health forum content is grouped by forums, which can be hierarchical with sub-forum. Within each of these, content is grouped by topics, responses to which are grouped viathreads. Some forums may be by subscription only, which restricts to

DOI:10.5963/PHF0202005

user groups.

- Forum: C{Admin}, A{Admin}, P{All | Group}, U{Admin}, R{Admin}
- Topic: C{Registered | Group}, A{Admin | Group}, P{All | Group}, U{Admin | Owner}, R{Admin | Owner}
- Thread: C{Registered | Group}, A{Admin | Group}, P{All | Group}, U{Admin | Owner}, R{Admin | Owner}

Health Communities and Networks

In most social networks, content is organized on a user *wall*. Users can *post* content on the wall, as well as their *status*. Post and status content can have *comments* made. A major distinction of social networks is the presence of the "Friends" user group. In addition, the user has controlled over their own content, and usually decides viewing privileges, which translates to the approval step.

- Wall: C{Owner}, A{Owner}, P{All | Registered | Group}, U{Owner}, R{Owner}
- Post: C{All | Registered | Group}, A{Owner}, P{All | Registered | Group}, U{Owner}, R{Owner}
- Status: C{Owner}, A{Owner}, P{All | Registered | Group}, U{Owner}, R{Owner}
- Comment: C{All | Registered | Group}, A{Owner}, P{All | Registered | Group}, U{Owner}, R{Owner}

The above configurations demonstrate the ability of the IB to represent via appropriate privacy scope different types of health websites. Table 1 summarizes the vocabularies used for the IBs.

TABLE I VOCABULARIES OF DIFFERENT TYPES OF HEALTH WEBSITES USING IBS

Health Website	Vocabularies
Article-based	Menu, Article, Comment
Blog	Category, Post, Comment
Wiki	Category, Wiki Post
Forum	Forum, Topic, Thread
Community	Wall, Post, Status, Comment

VI. CONCLUSION AND FUTURE WORK

The Internet is an ideal tool for promoting public health goals of prolonging life, health, and improving the quality of life. This holds even truer because health is a hot topic on the Internet, and many websites are available for users to search for health advice. However, these websites can lead to harm if ethical issues are not taken into consideration. Existing CMSs and CMFs are not necessarily the best for making health websites with incorporated ethics. Moreover, disclaimers in terms of service may satisfy legalistic requirements, but ethically more ought to be done to ensure safety and trustworthiness of their content. We propose an HCMS that is aware of ethical aspects and allows automated incorporation of the necessary features within the content lifecycle. In addition, we also suggest unique data mining and recommender features that are not available in current health websites to the best of our knowledge. It should be noted that the work presented here is a work in progress. These initial conceptual specifications of the system are meant to facilitate the next steps of implementation and evaluation. The survey results of existing health websites, CMSs, and CMFs are valuable in providing a snapshot of the state-of-the-health-web as well as quantifying the need to continue with the proposed system. Moreover, this study presented a unique approach to analysing health content via lifecycles. The next phase of this research will involve pilot studies. Fulfilling legal obligations merely requires a disclaimer and transferring the burden of verification to the end-user, but fulfilling ethical obligations for non-maleficence requires a cultural shift and more action on our part. It remains to be seen how this HCMS will fare and what level of adoption will be possible. However, we hope this will encourage deeper questions about ethical obligations to patients' well-being in the internet age.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant.Privacy-Preserving Data Mining. ACM SIGMODRecord, 29:439–450, 2000.

[2] James G. Anderson and Kenneth W. Goodman. Ethics and Information Technology:A Case-Based Approach to a HealthCareSystem in Transition.Springer New York, 2002.

[3] Angela Arner and Mary Wolcott-Breci.Greystone.Net, Web Site Health Content Management. Journal of the Medical Library Association, 94(2):235–237, 2006.

[4] Susannah Fox. Health Information Online. Pew Internet and American Life Project. Retrieved August 10, 2012 from http://www.pewinternet.org/PPF/r/156/report_display.asp, 2005.

[5] Milt Freudenheim. Health Care Is High Among Web Searches. The New York Times. Retrieved August 21,2012 from http://prescriptions.blogs.nytimes.com/2011/02/01/health-care-is-high-among-web-searches/, 2011.

[6] Frank Gilbane. What is Content Management. White paper, Retrieved August 10, 2012 from http://gilbane.com/gilbane_report.pl/6/What_is_Content_Management.html, 2000.The Gilbane Reports, 8(8).

[7] Greystone.Net.Greystone.Net Website. Retrieved August 10, 2012 fromhttp://www.greystone.net.

DOI:10.5963/PHF0202005

[8] YiHong, TimothyB. Patrick,and Rick Gillis.Protection of Patient's Privacy and Data Security in E-Health Services. In Proceedings of the 2008 International Conference on Bio Medical Engineering and Informatics, pages 643–647, 2008.

[9] James B. Weaver III, Nancy J. Thompson, Stephanie Sargent Weaver, and Gary L. Hopkins. Healthcare Non-adherence Decisions and Internet Health Information. Computers in Human Behavior, 25:1373–1380, 2009.

[10] EIBSLtd. Content Management for NHS & Health Care Websites, Extranets, Intranets, Portals-EasySite CMS. Retrieved August 10, 2012 from http://www.eibs.co.uk/sectors-and-clients/ nhs-content-management-for-health-care/.

[11] Genetics Ltd. Content Management Solutions for NHS & Health. Retrieved August 10, 2012 from http://www.contentmanagement.co.uk/discover/solutions/content-management-nhs-health.aspx.

[12] Microsoft. The Microsoft Connected Health Framework –Architecture and Design Blueprint. Retrieved August 10, 2012 from http://www.microsoft.com/industry/healthcare/technology/Healthframework.mspx.

[13] Compare Stuff Network. The CMS Matrix. Retrieved August 10, 2012 from http://www.cmsmatrix.org.

[14] U.S. National Library of Medicine. Medical Subject Headings (MeSH).RetrievedAugust 10, 2012from http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[15] Stanley Oliveira, Stanley R. M. Oliveira, and Osmar R. Zaïane. Toward Standardization in Privacy-Preserving Data Mining. In Proceedings of the 3rd Work shop on Data Mining Standards, in conjunction with KDD 2004, pages 7–17, 2004.

[16] Julio Ponce, Alberto Hernndez, Alberto Ochoa, Felipe Padilla, Alejandro Padilla, Franciscolvarez, and Eunice Poncede Len.Data Mining and Knowledge Discovery in Real Life Applications, ed. Julio Ponceand Adem Karahoc.I-Tech, Vienna, Austria, 2009.

[17] J. M. Roberts and K. L. Copeland. Clinical Websites are Currently Dangerous to Health. International Journal of Medical Informatics, 62(2-3):181–187, 2001.

[18] Hamman W. Samuel, Osmar R. Zaïane, and Dick Sobsey. Towards a Definition of Health Informatics Ethics.In Proceedings of 1st ACM International Health Informatics Symposium, IHI '10, pages 257–264, 2010.

[19] Tony C. Shan and Winnie W. Hua. Taxonomy of Java Web Application Frameworks. In IEEE International Conference on e-Business Engineering, pages 378–385, 2006.

[20] Bhavani Thuraisingham. Web Data Mining and Applications in Business Intelligence and Counter-Terrorism. CRC Press LLC, Florida, USA, 2003.

[21] Bhavani Thuraisingham, Latifur Khan, Chris Clifton, John Maurer, and Marion Ceruti. Dependable Real-Time Data Mining. IEEE International Symposium on Object-Oriented Real-TimeDistributedComputing,0:158–165, 2005.

[22] H. Joseph Wen and Joseph Tan. The Evolving Face of Tele Medicine& E- Health: Opening Doors and Closing Gaps in E-Health Services Opportunities & Challenges. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.

[23] Wikipedia. List of Content Management Frame works. Retrieved August 10, 2012 fromhttp://en.wikipedia.org/wiki/List_of_content_ management frame works, 2008.

[24] Jin Xu, Yingping Huang, and Gregory Madey.A Research Support System Framework for Web Data Mining. In Workshop on Applications, Products and Services of Web-based Support Systems at the Joint International Conference on Web Intelligence and Intelligent Agent Technology, pages 37–41,2003.

[25] Susan Mc Keever. Understanding Web Content Management Systems: Evolution, Lifecycle and Market. Industrial ManagementandDataSystems,103(9): 686–692, 1970.

[26] Ann Rockley. Managing Enterprise Content: A Unified Content Strategy. New Riders Press, 2002.

[27] Gerry McGovern and Rob Norton. Content Critical. FT Press, 2001.

[28] State Government, Victoria, Australia. Web Content Lifecycle. Retrieved August 21,2012 from http://www.egov.vic. gov.au/victorian-government-resources/reports-victoria/web-content-lifecycle-and-content-management-roles/1-web-content-lifecycle.html, 2010.

[29] Alan Thomson and Daniel Schmoldt. Ethics in Computer Software Design and Development. Journal of Computers and Electronics in Agriculture 2001; 31:85-102.

[30] Trz sicki K. Medical Informatics Ethics (Subject and Major Issues). Studies in Logic, Grammar and Rhetoric 2005; 8(22):55-72.

[31] Elizabeth Layman. Health Informatics: Ethical Issues. Health Care Manager 2003; 22(1):2-15.

[32] Kenneth Goodman and Randolph Miller. Ethics and Health Informatics: Users, Standards, and Outcomes. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer New York, 2006; pages 379-402.

[33] Ken Masters. Health Informatics Ethics. Health Informatics: Practical Guide for Healthcare and Information Technology. Lulu.com, 2012; pages 195-215.

[34] Peter Winkelstein. Ethical and Social Challenges of Electronic Health Information. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer US, 2005; pages 139-159.

[35] International Standards Organization (ISO). TC 215 - Health Informatics. Retrieved December 22, 2012 from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=54960.

[36] U.S. Department Of Health and Human Services, Food and Drug Administration (FDA), Center for Devices and Radiological Health, Center for Biologics Evaluation and Research. General Principles of Software Validation; Final Guidance for Industry and FDA Staff. FDA 2002. Retrieved January 15, 2013 from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm085371.pdf.

[37] U.S. Department Of Health and Human Services, Food and Drug Administration (FDA), Center for Devices and Radiological Health, Center for Biologics Evaluation and Research. Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices. FDA 2005. Retrieved January 15, 2013 from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm089593.pdf.

DOI:10.5963/PHF0202005

[38] U U.S. Department Of Health and Human Services. Understanding Health Information Privacy. Retrieved January 15, 2013 from http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html.

[39] Tara Lagu, Elinore Kaufman, David Asch, and Katrina Armstrong,. Content of Weblogs written by Health Professionals. Journal of General Internal Medicine, 2008:23(10), 1642-1646.

[40] Mayo Clinic. Retrieved January 15, 2013 from http://www.mayoclinic.com/health/DiseasesIndex/.

[41] Health Link Alberta. Retrieved January 15, 2013 from http://www.albertahealthservices.ca/223.asp.

DOI:10.5963/PHF0202005