eHealth Beyond the Horizon – Get IT There
S.K. Andersen et al. (Eds.)
IOS Press, 2008

Cataloguing and displaying Web feeds from French language health sites: a Web 2.0 add-on to a health gateway

G. KERDELHUÉ[a], B. THIRION, B[a]. DAHAMNA[a], SJ. DARMONI[a]
[a] *CISMeF, Rouen University Hospital, 76031 Rouen, France*

Abstract. Among the numerous new functionalities of the Internet, commonly called Web 2.0, Web syndication illustrates the trend for better and faster information sharing. Web feeds (a.k.a RSS feeds), which were used mostly on weblogs at first, are now also widely used in academic, scientific and institutional websites such as PubMed. As very few French language feeds were listed or catalogued in the Health field by the year of 2007, it was decided to implement them in the quality-controlled health gateway CISMeF ([French] acronym for Catalogue and Index of French Language Health Resources on the Internet). Furthermore, making full use of the nature of Web syndication, a Web feed aggregator was put online in to provide a dynamic news gateway called "*CISMeF actualités*" (http://www.chu-rouen.fr/actualites/). This article describes the process to retrieve and implement the Web feeds in the catalogue and how its terminology was adjusted to describe this new content. It also describes how the aggregator was put online and the features of this news gateway. *CISMeF actualités* was built accordingly to the editorial policy of CISMeF. Only a part of the Web feeds of the catalogue were included to display the most authoritative sources. Web feeds were also grouped by medical specialties and by countries using the prior indexing of websites with MeSH terms and the so-called metaterms. *CISMeF actualités* now displays 131 Web feeds across 40 different medical specialities, coming from 5 different countries. It is one example, among many, that static hypertext links can now easily and beneficially be completed, or replaced, by dynamic display of Web content using syndication feeds.

Keywords. Information Dissemination, Internet, Quality-controlled health gateway, Syndication feed

Introduction

The core of the World Wide Web is an extensive repository of documents linked by hypertext. It is now often referred to as "Web 1.0". In this context, several quality-controlled health gateways have been developed [1]. These gateways were defined by Koch [2] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery. Their 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. CISMeF ([French] acronym for Catalog and Index of French Language Health Resources on the

Internet) [1] was designed to catalog and index the most important and quality-controlled sources of institutional health information (N=36,103). Its URL is http://www.chu-rouen.fr/cismef.

Web 2.0 is a metaphor describing new Internet applications and services which emphasize greater user participation in developing and managing content [3]. As the use of Internet becomes more and more "social", the need for bigger and faster information exchange grows. It manifests by the growth of Web syndication.

Web syndication is a form of information dissemination in which the (partial or complete) content of a website is made available for end-users or other sites to use. It is inherent to a large number of Web services, blogs, wikis, social networks, etc. The two main families of web syndication formats (XML-based) are RSS (Really Simple Syndication) and Atom [4]. These so-called Web feeds provide a way to rapidly disseminate awareness of new information. They permit continuous instant alerting to the latest ideas in medicine [5]. The adoption of the RSS format by PubMed to display search results is a remarkable example [6].

Convinced by the inherent qualities of Web feeds and of their interest for the medical domain, the CISMeF team decided to catalogue health Web feeds available in French in January 2007. At that time, they were totally scattered across the Web. Some general lists and portals were existing but they mostly include Web feeds from journalistic or blogs sources. French language Web feeds from scientific, academic or institutional sources were neither listed nor catalogued in the health field.

The goal of this work is to describe how Web feeds were included in the existing catalogue and how a dynamic news gateway called *CISMeF actualités* [7] was created.

1. Material and Methods

1.1. CISMeF terminology

CISMeF uses two standard tools to organize information: the MeSH thesaurus [8] from the US National Library of Medicine and several metadata element sets, in particular the Dublin Core metadata format [9]. In order to customize the MeSH to the field of health Internet resources, several enhancements [1] to this thesaurus were developed with the introduction of two new concepts, respectively metaterms (MT) and resource types (RT).

A metaterm denotes a medical specialty (e.g. *cardiology*), a biological science (e.g. *bacteriology*), or a health topic (e.g. *diagnosis*), which has semantic links with one or more MeSH terms, subheadings and RTs (N=123). The idea of creating metaterms came up to optimize information retrieval in CISMeF and to cope with the relatively restrictive nature of these medical specialties as MeSH keywords. The metaterms are also used reciprocally to categorize a document from its specific indexing [10].

CISMeF resource types (RT) are an extension of the publication types of MEDLINE. They are used to describe the nature of the resource, while MeSH terms describe its topic (N=275).

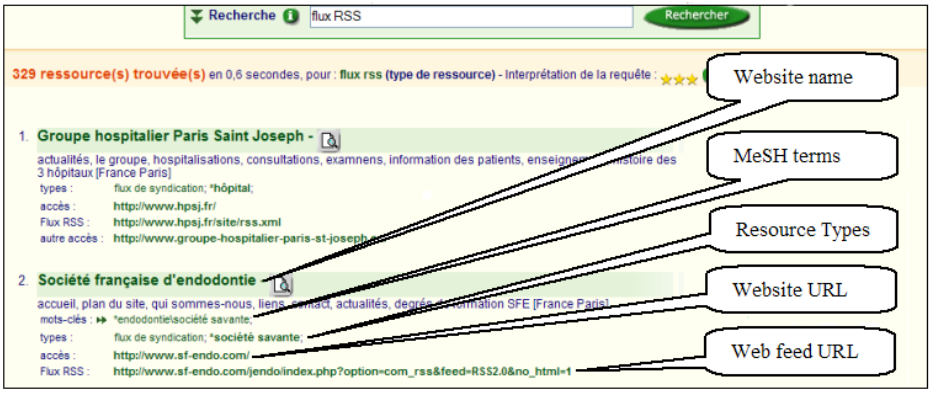

Figure 1: RSS Feeds as they appear in the catalogue after the query "RSS feeds".

1.2. Implementing feeds in the catalogue

The first step was to retrieve the Web feeds. As nothing previously existed, it was done progressively and still it is an ongoing, not automated, process composed by several elements: survey of new or already indexed websites (progressively checked during day to day activities), use of search engines which permits to search Web feeds specifically.

The second step was to add two new elements: the web feed URL and a new RT (see Figure 1). As syndication feeds are always attached to a single website, it was decided not to create completely new entries in the catalogue but to add the feed description to the already existing entries for websites, thus attaching the prior indexing with MeSH terms to the feed itself. Two URLs are mentioned for each website providing web syndication: one for the website homepage, one for the web feed itself. The new RT created was called "syndication feed", ("flux de syndication" in French). This resource type can be used in simple or advanced search to rapidly target the Web feeds among all the resources of the catalogue.

1.3. One step beyond the directory: displaying web feeds

Though the number of existing Web feeds is constantly growing, they remain unknown to a large part of the CISMeF audience. It was decided to display the feeds in order to make this content accessible for anyone without any prior technical knowledge.

Several free services offer to display a selection of Web feeds such as Google reader, Bloglines or Netvibes. However it was preferred to use an open-source software in order to have greater control over *CISMeF actualités* and to guarantee the independence and persistence of the service.

The Software chosen was Gegarius [11]. Its most deciding features were: to be completely web-based and running on a web server, to support multiple feeds formats, to provide a search engine, to be translated in French language, to be committed to Web standards (it renders XHTML/CSS), and finally to be a free software released under the *GNU General Public License.* The technical requirements were quite low: an Apache web server, PHP, and a MySQL Database. The software was then adapted to the CISMeF graphic charter by simply modifying its style sheet.

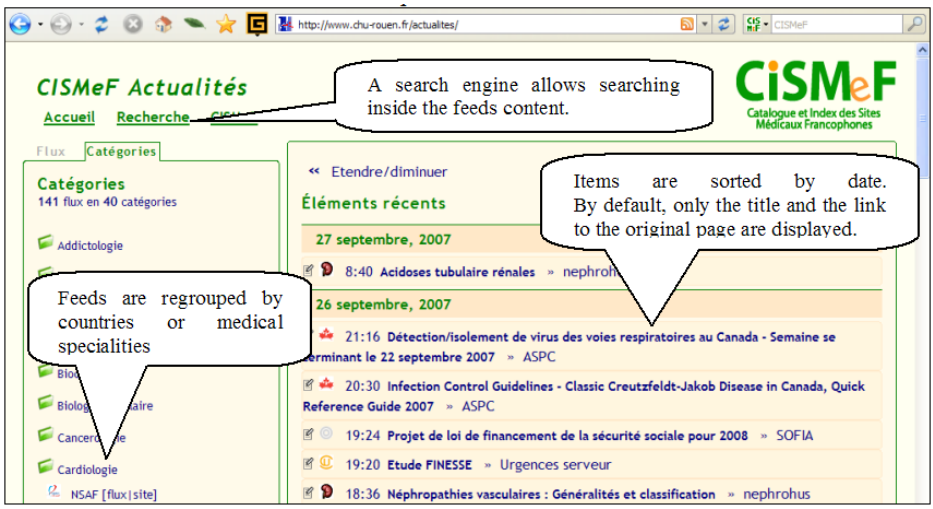

Figure 2. The *CISMeF Actualités* homepage.

1.4. Which feeds to display? Extending the CISMeF editorial policy to Web feeds

Not all the feeds catalogued in CISMeF are displayed in *CISMeF actualités* due to editorial choices but also to technical problems.

The CISMeF editorial policy is to reference only documents from the most authoritative sources within the large number of websites referenced. Thus, only the feeds provided by institutions, hospitals, medical schools and health professionals associations were displayed in *CISMeF Actualités*. On the other hand, feeds provided by the patients associations websites, the professional unions, though being indexed in the catalogue, were discarded.

We did not choose to display feeds from medical journals as there are still very few of them in French language and as their content is mostly not freely available.

The technical problems concern mostly Web feeds that do not comply with the RSS specification [12].

1.5. Categorize the Web feeds

Categorization is performed using semantic links between MeSH terms, MeSH qualifiers, CISMeF resource types and CISMeF metaterms (See Figure 2). For example, the web feed for the website CETL (URL: http://www.cetl.net/) indexed with the MeSH term “lysosomal storage diseases” was categorized in Genetics.

2. Results

2.1. Features of “CISMeF actualités”

The most important features of *CISMeF actualités* are:

- Almost real-time updates (while the CISMeF catalogue itself is updated weekly). Feeds are regularly checked for new items. As Gregarius itself does not have

scheduled tasks built in, a planned task using crontab executes every two hours the php script in charge of updating the feeds.
- Search engine. The content of each feed is accessible through full text search.
- Feeds are regrouped by countries and medical specialities (see Figure 2).
- Feeds aggregation and RSS output. For the whole CISMeF actualités, for each category and for each search results, Web syndication is available.
- Name and Logo of the publisher clearly indicate the source of each item. For the CISMeF team, the publisher is the main criterion of quality from the HON code of conduct [13].

2.2. Statistics

The number of Web feeds indexed in the catalogue is 329 (11 October 2007). 131 feeds are displayed in *CISMeF Actualités* (11 October 2007). These feeds are categorized in 40 medical specialities. Thus 40 out of 123 (32%) CISMeF metaterms are used. The number of unique visitors showed constant augmentation between July and September 2007 from 4675 to 7657.

2.3. Syndicate the aggregated feeds from CISMeF Actualités

CISMeF actualités provides a Web feed for each of these categories, aggregating the existing feeds. These new feeds can also be displayed in specialized context. We used this feature on a specialized gateway about handicap (http://www.chu-rouen.fr/handicap) where the feed of the category Handicap from *CISMeF Actualités* is the core of the news section.

In the near future, we could implement a similar feature in a specialized search engine (http://doccismef.chu-rouen.fr/servlets/KISMeF) in collaboration with the French National Cancer Institute (http://www.e-cancer.fr/).

3. Discussion

Integrating Web feeds in the CISMeF gateway fulfils several needs. Users convinced by their usefulness can now find feeds in the French language medical domain through the catalogue itself. Users ignorant of Web feeds or lacking the technical knowledge to use them can now easily access their content through *CISMeF Actualités*.

Very few papers in the medical literature refer to web syndication (n=15 in the MEDLINE bibliographic database September 27, 2007 with the query: "web syndication" or "RSS feed"). Most of them consist of explanations aimed at the end-users, very few concern the information providers [14] [15].

Two important websites provide service comparable to *CISMeF Actualités,* though at a larger scale and concerning English language. They were major sources of inspiration.

The National Library for Health (UK) provides a large directory of RSS feeds accessible at this URL http://www.library.nhs.uk/rss/Directory/ [16]. It shares with CISMeF the human selection of the feeds, the ability to browse by categories and a search engine. But they do not display the feeds themselves and the search engine do not allow to search within the feed content.

Medworm (http://www.medworm.com/) [17] is a medical RSS feed provider as well as a search engine built on data collected from RSS feeds. It provides a search engine that search through the content of the feeds and sort the results by date and relevance. The content is also available through different categories.

The most important limit of *CISMeF Actualités* is the lack of Web feeds from major institutions. This is especially true for France. For example, The major Belgian and Canadian Health agencies provide Web feeds while the French don't.

The second limitation is the contents of the feed themselves. As they consist mostly of Press announcements, some major events in the Health field may not appear at all, while small events, like charity events for example, may appear out of range. *CISMeF Actualités* provides raw information, without the selection or the synthesis a journalist would do.

In the future, the growing number of websites deciding to provide Web feeds raises a number of questions. The discovery of Web feeds itself may need to be automated so that feeds from important sources would not be missed. Considering *CISMeF Actualités*, its content could be evaluated so that only the most important and relevant feeds are displayed.

References

[1] Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ: Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004 Dec: 21(4):253-61.

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

[3] O'Reilly T. What is Web 2.0? Design patterns and business models for the next generation of software. http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed Jun 2007).

[4] Web syndication - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Web_syndication (accessed October 4, 2007).

[5] Giustini D How Web 2.0 is changing medicine. BMJ, Vol. 333, No. 7582. (23 December 2006), pp. 1283-1284.

[6] NLM Technical Bulletin, May-June 2005, RSS Feeds Available from PubMed. http://www.nlm.nih.gov/pubs/techbull/mj05/mj05_rss.html (accessed October 4, 2007).

[7] CISMeF Actualités. http://www.chu-rouen.fr/actualites/ (accessed October 4, 2007).

[8] Medical Subject Headings - Home Page. http://www.nlm.nih.gov/mesh/ (accessed October 4, 2007).

[9] Dublin Core Metadata Initiative (DCMI). http://dublincore.org/ (accessed October 4, 2007).

[10] Darmoni SJ, Névéol A, Renard JM, Gehanno JF, Soualmia LF, Dahamna B, Thirion B. A MEDLINE categorization algorithm. BMC Med Inform Decis Mak. 2006 Feb 7;6(1):7

[11] Gregarius - A Free, Web-based Feed Aggregator. http://gregarius.net/ (accessed October 4, 2007).

[12] RSS 2.0 Specification (RSS 2.0 at Harvard Law). http://cyber.law.harvard.edu/rss/rss.html (accessed October 4, 2007).

[13] HONcode: Principles - Quality and trustworthy health information. http://www.hon.ch/HONcode/Conduct.html (accessed October 4, 2007).

[14] Barsky E Introducing Web 2.0: RSS trends for health librarians JCHLA / JABSC, Vol. 27, No. 1. (Winter 2006), pp. 7-8.

[15] J Robinson, S de Lusignan, P Kostkova, B Madge, A Marsh, C Biniaris The Primary Care Electronic Library: RSS feeds using SNOMED-CT indexing for dynamic content delivery. Inform Prim Care, Vol. 14, No. 4. (2006), pp. 247-252.

[16] National Library for Health - RSS Directory. http://www.library.nhs.uk/rss/Directory/ (accessed October 4, 2007).

[17] MedWorm: Medicine RSS. http://www.medworm.com/ (accessed October 4, 2007).