Language Complexity in On-line Health Information Retrieval

Marco Alfano[1,5][0000-0001-7200-9547], Biagio Lenzitti[2][0000-0003-2664-7788], Davide Taibi[3][0000-0002-0785-6771], Markus Helfert[4][0000-0001-6546-6408]

[1] Lero, Dublin City University, Dublin, Ireland
[2]Dipartimento di Matematica e Informatica, Università di Palermo, Palermo, Italy
[3]Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Palermo, Italy
[4] Lero, Maynooth University, Maynooth, Co. Kildare, Ireland
[5]Anghelos Centro Studi sulla Comunicazione, Palermo, Italy
marco.alfano@lero.ie, biagio.lenzitti@unipa.it, davide.taibi@itd.cnr.it, markus.helfert@lero.ie

Abstract. The number of people searching for on-line health information has been steadily growing over the years so it is crucial to understand their specific requirements in order to help them finding easily and quickly the specific information they are looking for. Although generic search engines are typically used by health information seekers as the starting point for searching information, they have been shown to be limited and unsatisfactory because they make generic searches, often overloading the user with the provided amount of results. Moreover, they are not able to provide specific information to different types of users. At the same time, specific search engines mostly work on medical literature and provide extracts from medical journals that are mainly useful for medical researchers and experts but not for non-experts.
A question then arises: Is it possible to facilitate the search of on-line health/medical information based on specific user requirements? In this paper, after analysing the main characteristics and requirements of on-line health seeking, we provide a first answer to this question by exploiting the Web structured data for the health domain and presenting a system that allows different types of users, i.e., non-medical experts and medical experts, to retrieve Web pages with language complexity levels suitable to their expertise. Furthermore, we apply our methodology to the results of a generic search engine, such as Google, in order to re-rank them and provide different users with the proper health/medical Web pages in terms of language complexity.

Keywords: E-Health, Health Information Seeking, User Requirements, Language Complexity, Structured Data on the Web.

1 Introduction

The number of people searching for on-line health information has been steadily growing over the years [1], [2] so it is crucial to understand their specific requirements in order to help them finding easily and quickly the specific information they

are looking for. Although search engines are typically used by health information seekers as the starting point for their searches [2], [3], they have been shown to be limited and unsatisfactory for finding online health information easily and quickly [4], [5]. In particular, generic search engines (e.g., Google™ or Bing™) exploit the whole Web but make generic searches, often overloading the user with the offered amount of information. Moreover, they are not able to provide specific information to different types of users. At the same time, specific search engines, such as PubMed[1] or the Cochrane Library[2], mostly work on medical literature and provide extracts from medical journals that are mainly useful for medical researchers and experts but not for non-experts. Moreover, they do not consider all the information contained in the Web that is often addressed to non-medical experts.

A question then arises: Is it possible to facilitate the search of on-line health/medical information based on specific user requirements? In this paper, we provide a first answer to this question by exploiting the structured data on the Web for the health domain and presenting a system that allows different types of users, i.e., non-medical experts and medical experts, to retrieve Web pages with language complexity levels suitable to their expertise. Furthermore, we apply our methodology to a generic search engine, such as Google, in order to re-rank its results and to provide different users with the proper health/medical Web pages in terms of language complexity. To this end, we first present a short survey of the main characteristics and requirements related to health information seeking on the Internet. We then analyze the structured data on the Web with particular reference to the health/medical field (by using *health-lifesci.schema.org*) and classify health Web pages based on different audience types such as patients, clinicians and medical researchers. Next, we present the results of some experiments on the language complexity of medical Web pages with structured data and propose a mapping between the language complexity requirements and the *health-lifesci.schema.org* audience types. We then present the architectural and implementation details of FACILE, a meta search engine that provides Web pages ranked in accordance to the audience type. Finally, we show the results of applying FACILE search and ranking capabilities to both the schema.org structured data and the Google results.

Some of the principles presented in this paper are based on the ones discussed in a previous work [6]. The present work, however, extends the previous study by including a literature survey on the health seekers requirements. Moreover, a larger dataset is used by merging the health-lifesci.schema.org structured data of 2017 with the ones of 2018. Furthermore, the description of the FACILE architecture and implementation (with a new ranking formula that takes into account a higher number of parameters) is added together with the application of the FACILE searching and ranking mechanism to both the *schema.org* structured data and the Google results.

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://www.cochranelibrary.com/

2 Characteristics and requirements of On-line Health Information Seeking

We now briefly analyze the main characteristics related to health information seeking on the Internet, based on the following dimensions:

- Who (e.g., number of people searching for health information on the Internet);
- Where (e.g., search engines, social networks);
- When (e.g., time frequency);
- What (e.g., symptoms, pathologies, remedies, drugs);
- How (e.g., user requirements of on-line health information seekers).

The 'Cyberchondriacs' Harris Poll [1] shows that the percentage of all US adults who search for health or medical information online has increased from 27% to 76% from 1998 to 2010. Moreover, the 'Health Online 2013' Pew report [2] says that 72% of adult users in the U.S. were looking for health information online in the previous year. When asked to think about the last time they went online for health or medical information, 39% of online health seekers say they looked for information related to their own situation. Another 39% say they looked for information related to someone else's health or medical situation. An additional 15% of these internet users say they were looking both on their own and someone else's behalf. For what concerns Europe, [7] shows a growth from 14% to 39% in the 2005-2007 period. Moreover, in 2010, national bodies reported that 52,5% of adults in Spain were looking for health content on the Internet [8] and 39% in the UK [9].

According to [2], 77% of online health seekers say they began their last session at a search engine such as Google, Bing, or Yahoo. Another 13% say they began at a specialized site in health information, like WebMD. Just 2% say they started their research at a more general site like Wikipedia and an additional 1% say they started at a social network site like Facebook. According to the survey reported in [10], a general search engine is the most frequently used tool to look for online health information. Other popular sources include Websites providing health information (38%) and Wikipedia or medical search tools such as HONselect and Medline Plus (37%). Forums and blogs are always or often used by 23% of the respondents and 5% use Facebook or other social networks. The same paper affirms that Internet is the second source of information after physicians whereas [11] states that Internet is the most commonly consulted resource for health information followed by conversation with health care providers and use of a medical dictionary.

The 'Cyberchondriacs' Harris Poll [1] shows that the percentage of US adults who often or sometimes search for health or medical information online has increased from 42% to 73% from 1998 to 2010. Moreover, 81% of health information seekers say that they have looked for health information online in the last month and 17% say they have gone online to look for health information ten or more times in the last month. On average, health information seekers do this about 6 times a month. According to the survey presented in [10], 24% of the respondents say they look for health information on the Internet at least once a day and 25% do it few times a week. Moreover, 8% do it once a week, 16% do it few times a month and 16% do it once a month.

The ‘Health Online 2013’ Pew report [2] shows that the most searched health topics are: Specific disease or medical problem (55%), Certain medical treatment or procedure (43%), How to lose weight or how to control your weight (27%), and Health insurance, including private insurance, Medicare or Medicaid (25%). According to the survey reported in [10], the search activity of users is mostly focused on general health information (68%), long-term chronic diseases (59%), healthy lifestyle and nutrition (50%), short-term (up to 2 weeks) acute disease (39%), kids health (22%) and elderly health and care (19%).

A short literature review to evaluate the main user requirements of health information seekers has been carried out in another work [12]. The survey has been revised and extended and the results are reported in Table 1.

Table 1. User requirements of health information seekers.

Paper	Language Complexity	Info Quality	Info Classification/ Customization	Other
N. Pletneva, A. Vargas, C. Boyer. 2011. Requirements for the general public health search [10]	●	●	●	
S. Banna, H. Hasan, P. Dawson. 2016. Understanding the diversity of user requirements for interactive online health services [13]	●	●		
T. Roberts. 2017. Searching the Internet for Health Information: Techniques for Patients to Effectively Search Both Public and Professional Websites [14]		●		
W. Pian, C.S.G. Khoo, J. Chi. 2017. Automatic classification of users’ health information need context: Logistic regression analysis of mouse-click and eye-tracker data [15]			●	
P. C.-I. Pang, K. Verspoor, J. Pearce, S. Chang. 2015. Better Health Explorer: Designing for Health Information Seekers [16]	●		●	●
A. Keselman, R. Logan, C. Smith. 2008. Developing informatics tools and strategies for consumer-centered health communication [17]	●	●	●	
Ardito, S. C. 2013. Seeking Consumer Health Information on the Internet [18]	●	●		

Although limited, the literature review presented above shows that the main requirements of health information seekers are the following:

- Language complexity
- Information quality (mainly intended as information trustworthiness)
- Information classification/customization.

Summarizing, we have found that there is a high number of people seeking for health information on the Internet that has been constantly increasing over the years (who). Search engines are the most used means to access medical information (where) and they are used more and more often (when) to seek information on a broad range of medical subjects (what). Moreover, the main requirements of health information seekers are language complexity, information quality and information classification and customization (how).

As stated in the Introduction, this paper mainly focuses on presenting the principles and design/development details of a system that allows to provide different types of users (e.g., medical experts and non-experts) with health/medical Web pages with different language complexity levels so to allow them to immediately find Web medical contents that present a language suitable to their expertise. In another work [12], we explore the other two user requirements, information quality and information classification/customization, and provide a mapping model among those user requirements and the schema.org elements.

As seen in Table 1, the papers dealing with the language complexity user requirement are [10], [13], [16], [17] and [18]. In particular, [10] presents a survey on user requirements which shows that users want to know if the information they search for is explained in the same way their doctor would but they do not present a solution for providing this type of information as we do in this work. Similarly, [13] shows that users feel that the language used must be easy to understand but there is no practical indication on how to achieve it. The system presented in [16] contains a slider that allows to specify the reading level but the system only works with a small amount of information (few pages created by hand) whereas our system automatically works in real time with the health/medical resources provided by *schema.org* (tens of thousands of Web pages) and, in non-real time, with the whole Internet (through Google). [17] suggests that increased understanding can be accomplished by facilitating precise information retrieval with optimized, domain-specific search engines without providing any specific example. They also suggest automatic text translation to simpler text in order to enhance text readability. In other works [19], [20], we have also tackled the problem of translating medical/technical terms in lay terms so to facilitate their comprehension by non-medical experts. In the work presented here, however, our system directly finds the easy-to-understand Web pages available on the Web. Finally, [18] lists some consumer medical information reputable sites and suggests that patients should be taught to search PubMed, that is a collection of scientific medical articles mainly devoted to medical researchers. Our system, as already said, exploits the whole Internet and automatically provides either more complex or simpler web content depending on the user requirements.

3 Structured Data in Health science domain on the Web

In the last few years the use of *schema.org* vocabularies, to include semantic information in Web pages, has rapidly increased. The *schema.org*[3] initiative has been promoted in 2014 by major players in the search engine market with the aim to create, maintain, and reuse vocabularies for structured data on the Internet. In particular, *schema.org* defines types (e.g., *Product*, *Organization*, *People*) and related properties (e.g. *name*, *title*, *description*) that are interleaved within the HTML code and used to visualize that information in specific parts of a Web page. At present, the vocabularies defined by *schmea.org* are used in over ten million Web sites and search engines leverage the structured data to provide users with more appropriate results. Along with the core schema, that is used to describe a huge number of different types of entities from learning resources [21][22] or products and organizations, *schema.org* also defines extensions with the focus on specific sectors such as automotive, Internet of Thing (IoT) and health.

In our study, we are interested in exploiting structured data to match the requirements identified in Section 2 with particular respect to the requirements related to the complexity of the language used by the Web pages containing health related information. To this aim, we refer to the *health-lifesci* extension[4] of *schema.org* that contains 93 types, 175 properties and 125 enumeration values related to the health/medical field. They can be used, among others, to extract data related to the requirements of information quality, information classification and language complexity. In particular, for the language complexity, the *MedicalAudience*[5] type plays a key role to identify searching mechanisms that provide targeted information. This type describes the target audiences for medical Web pages and it includes *Patient*[6], *Clinician*[7] and *MedicalResearcher*[8] as more specific types. As reported in *schema.org*, a patient is any person recipient of health care services. Clinicians are medical clinicians, including practicing physicians and other medical professionals involved in clinical practice. Medical researchers are professionals who make research on the medical field.

In order to explore the use of the *schema.org* vocabulary to support health information seeking on the Web, we have evaluated the adoption of the types and properties defined in this vocabulary through the analysis of the *schema.org* information made available by the Web Data Commons initiative. The Web Data Commons (WDC) [23] contains all Microformat, Microdata and RDFa data extracted from the open repository of Web crawl data named Common Crawl (CC). At the time of writing, the latest release of the WDC dataset is dated November 2018 and it is based on 2.5 billion crawled pages with about 37% of them including structured data. We extended the work presented in [6] by merging the dataset extracted by WDC in 2017 with the one of 2018. The dataset dumps of the two years are made available by WDC

[3] https://schema.org/
[4] https://health-lifesci.schema.org/
[5] http://schema.org/MedicalAudience
[6] http://schema.org/Patient
[7] http://schema.org/Clinician
[8] http://schema.org/MedicalResearcher

as compressed files (8,433 files for 2017 and 7,263 for 2018). Each file is around 100 MB large and contains information in the form of RDF quadruples. A quadruple is a sequences of RDF terms in the form {s, p, o, u}, where s, p and o represent a triple consisting of subject, predicate, object and u represents the URI of the document from which the triple has been extracted.

Fig. 1 presents an example of RDF quads, for the *Patient* subtype, extracted from WDC. It clearly shows the subject, predicate, object and URI of the quadruples. In compliance with the Open Science model, we have made the RDF quads subsets, for the *Patient*, *Clinician* and *MedicalResearcher* specific types, available at the address http://h-easy.lero.ie/opendata/, in order to allow other researchers to use and lead further research on these data.

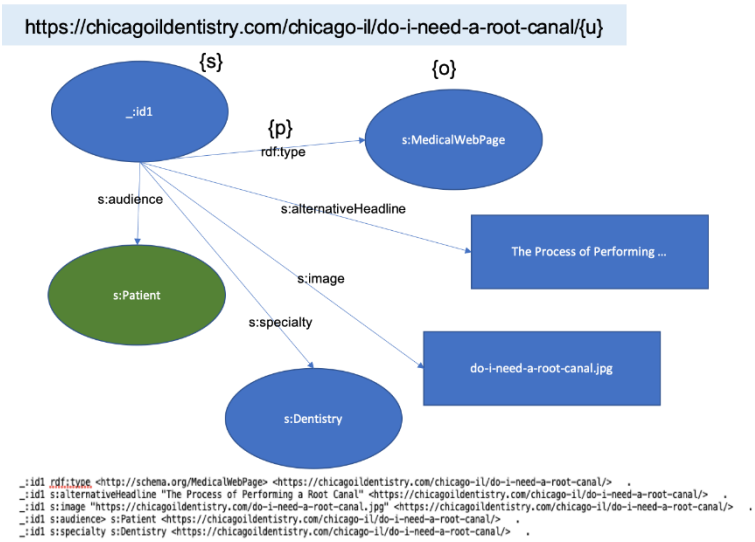
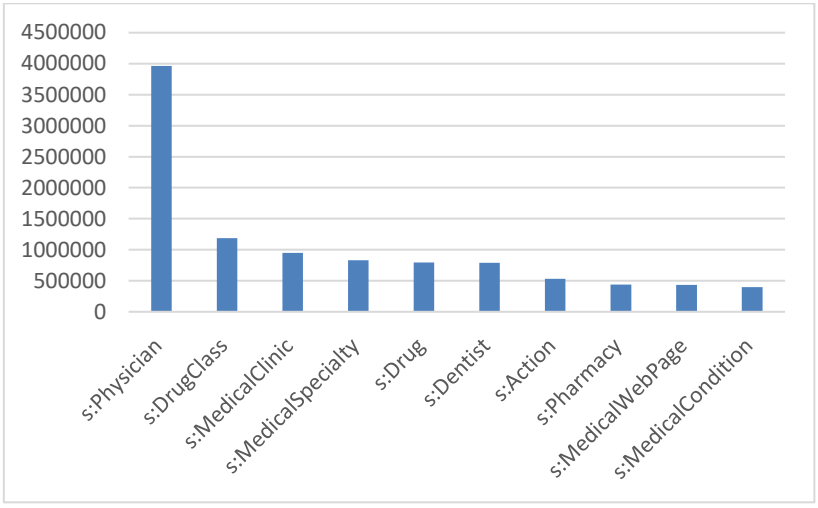


Fig. 1. Example of RDF quads for the Patient subtype.

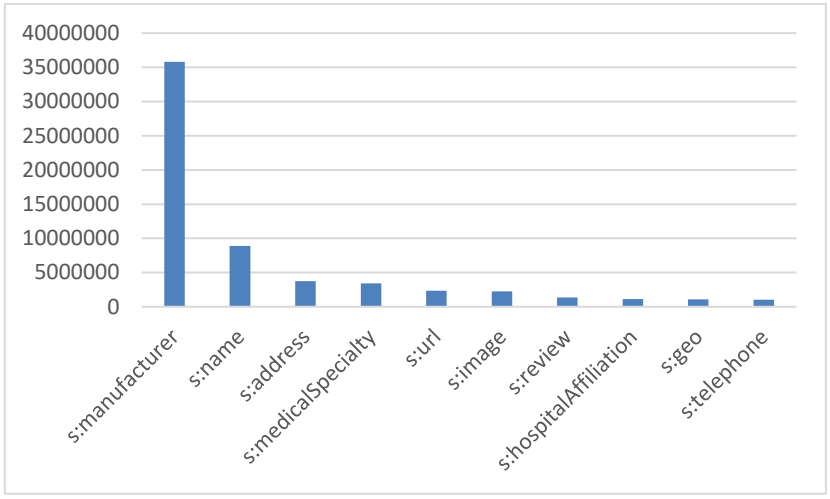
From the dataset dumps by WDC, we have filtered the quadruples that contain types and properties related to the health domain. The resulting dataset that we have used in our study consists of 103 billion RDF quadruples.

Fig. 2 (a) and (b) respectively show the top ten types and properties of the dataset we use for this study. Notice that, although, we have extracted types, properties and enumeration values of health-lifesci.schema.org, some types, such as *Action*, are generic and belong to the *schema.org* core vocabulary, but they assume a specific meaning in the context of *health-lifesci*. For example, the *Action* type is linked to the potential actions of a specific group of drugs. The same applies to properties such as *manufacturer* (presenting the highest frequency) which is generic and belongs to the *schema.org* core vocabulary but, in the context of *health-lifesci*, it refers to the organiza-

tion producing a specific *Drug*. Finally, notice that Physician is not used as a synonym of doctor but indicates the doctor office[9].



(a)



(b)

Fig. 2. Top ten types (a) and properties (b) of health-lifesci.schema.org.

[9] https://schema.org/Physician

We have also analyzed the distribution of the so called Pay Level Domains (PLDs) in the dataset including 2017 and 2018 dumps. The complete results of this analysis are available at the address http://h-easy.lero.ie/opendata/ while Table 2 shows the top ten results. In this list, we also indicate whether each PLD is related to the health/medical domain.

Table 2. PLDs with # of quads and health/medical indication.

#quads	PLD	Health/Medical
10544968	lybrate.com	yes
7082432	patents.google.com	no
3346339	vidal.fr	yes
2556287	vitals.com	yes
1567948	estdoc.jp	yes
1368641	restonhospital.com	yes
1309007	md.com	yes
1157954	carroya.com	no
1065347	spreadshirt.com	no
957936	doctoranytime.gr	yes

With regards to the *medicalAudience* property, we have computed the number of quads for each audience types and the results are reported in Table 3.

Table 3. Number of RDF Quads extracted for each specific type.

Schema.org types	RDF Quads
Patient	62,251
Clinician	17,416
MedicalResearcher	3,770

These three types, related to *MedicalAudience*, facilitate the identification of pages targeted to patients, clinicians and medical researchers. Table 4 shows an extract of five quads from each subset (the audience appears in the third column).

Notice that, at this stage, we have found Web pages that have been targeted to the different user types by their author, but we do not exactly know the reason behind the choice of considering a page more suitable for a specific audience type. In fact, the motivation could be related to the language complexity level (e.g., more or less tech-

nical) or to the treated subject (e.g., pathology symptoms and remedies, for patients, or technical aspects, for medical researchers), or to something else. In the next section, we present a mapping between the language complexity levels and the different audience types so to provide users with Web pages related to their specific requirements.

Table 4. An extract of five RDF quads extracted from Patient (a), Clinican (b) and MedicalResearcher (c) subsets.

Subject	Predicate	Object	Uri
_:genid2d65f95a781e614808bccfde1f41b001c32db0	<http://schema.org/audience>	http://schema.org/Patient	<https://dentistinsurrey.ca/cosmetic-dental-procedures-to-enhance-your-smile/>
<https://medlineplus.gov/spanish/ency/article/001054.htm>	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Patient	<https://medlineplus.gov/spanish/ency/article/00105.html>
<https://medlineplus.gov/ency/article/001525.htm>	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Patient	<https://medlineplus.gov/ency/article/001525.htm>
<https://medlineplus.gov/ency/patientinstructions/000391.htm>	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Patien	<https://medlineplus.gov/ency/patientinstructions/000391.htm>
_:node266bc63ad0aaf66daea4a87983675233	<http://schema.org/MedicalWebPage/audience>	https://health-lifesci.schema.org/Patient	<http://mis-varices-info.es/es/conexiones>

(a)

Subject	Predicate	Object	Uri
_:genid2dde430d3d6a664e8796b9654a5fa312882db88	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Clinician	<https://fpnotebook.com/cv/Exam/PlsPrdxs.htm>
_:node3651c910a570c21033d04278bfa589a8	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Clinician	<https://fpnotebook.com/cv/Exam/JPnt.htm>
_:nodebd34e3af7dbf1d2c29d520cd3372c32e	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Clinician	<https://fpnotebook.com>
_:node76312b2a953eb616b45ab7fe34f88c	<http://schema.org/MedicalScholarlyArticle/audience>	http://schema.org/Clinician	<http://www.creteilophtalmo.fr/en/2012/neovascularisation-choroidienne-compliquant-une-dmla-atrophique/>
_:node12c5ae94a53b3b39196fac4bc1aaaa9	<http://schema.org/MedicalWebPage/audience>	http://schema.org/Clinician	<http://www.choosingwisely.org.au/recommendations/gesa>

(b)

Subject	Predicate	Object	Uri
_:node4edd6b853592234609e785dd74bfa28	<http://schema.org/MedicalWebPage/audience>	http://schema.org/http://schema.org/MedicalResearcher	<https://www.malacards.org/>
_:nodea78f8069a42267ba126819c0543d237	<http://schema.org/MedicalWebPage/audience>	http://schema.org/http://schema.org/MedicalResearcher	<https://www.malacards.org/card/chronic_leukemia>
_:nodeb246e0cf395edb3ff3564dbf73d916	<http://schema.org/MedicalWebPage/audience>	http://schema.org/http://schema.org/MedicalResearcher	<https://www.malacards.org/search/results/atorvastatin>
_:genid2d8ba0b032efee4268945f68fa2bd1f2442db0	<http://schema.org/audience>	http://schema.org/http://schema.org/MedicalResearcher	<https://www.nanostring.com/products/gene-expression-panels/gene-expression-panels-overview>
_:node57b22f2149e6112a71febf24e34f9d67	<http://schema.org/MedicalWebPage/audience>	http://schema.org/http://schema.org/MedicalResearcher	https://www.malacards.org/card/inflammatory_breast_carcinoma

(c)

4 Mapping Language Complexity User Requirements to Audience Types

As seen above, users have different requirements when searching for health information on the Web. In particular, one of the most important requirement for non-expert health information seekers is that the language used in the Web pages must be easy to understand. On the opposite, medical experts require that the info they are looking for presents a proper technical and rigorous terminology. We then consider two classes of users:

- Non experts (e.g., patients or citizens);
- Experts (e.g., physicians or medical researchers).

We have taken the three subsets presented in the previous section, related to *Patient*, *Clinician*, and *MedicalResearcher* audience types, and, for each quadruple, we have analysed the related Web page in order to estimate its language complexity. To this end, we have evaluated the 'term familiarity index', as described in [6], [24], [25] of the English and non-empty Web pages (around 50% of the total). In particular, for each Web page, we have computed the term familiarity of each word by using the number of results provided by the Google search engine and we have then computed the page familiarity index by averaging all the term familiarity indexes. This information has been stored in a database to avoid work duplication.

In particular, for each Web page, we have computed and stored the number of unique words, the related page familiarity, the total number of words and the related page familiarity. The results of the performed experiments, for the three audience types, are available at the address http://www.math.unipa.it/simplehealth/simple2/ResSchema.php and the first six results of each audience type are shown in Fig. 3.

ID	URL	# Distinct Words	Page Familiarity DW (billions)	# Total Words	Page Familiarity TW (billions)
1	https://tatefamilydentistry.com/...	256	5.95	502	8.51
2	https://midtownoms.com/corrective-jaw-surgery/...	208	5.98	446	8.30
5	https://midtownoms.com/implant-bone-grafting/...	239	4.90	525	8.15
8	https://midtownoms.com/contact-us/...	76	7.80	139	7.18
9	https://www.restylaneusa.com/specialist...	361	4.72	1037	8.60
11	https://midtownoms.com/referring-doctors/...	92	7.56	168	7.68

(a)

ID	URL	# Distinct Words	Page Familiarity DW (billions)	# Total Words	Page Familiarity TW (billions)
3	https://www.onlinedentalmarketing.com/targeted-dental-market...	230	7.03	471	10.35
4	https://www.onlinedentalmarketing.com/targeted-dental-market...	346	5.57	790	9.96
10	https://www.onlinedentalmarketing.com/privacy-policy/...	552	5.36	1422	10.07
16	https://www.onlinedentalmarketing.com/blog/...	299	6.74	664	9.31
18	https://www.onlinedentalmarketing.com/meet-us/...	183	6.59	312	8.28
20	https://www.onlinedentalmarketing.com/targeted-dental-market...	334	5.69	692	9.23

(b)

ID	URL	# Distinct Words	Page Familiarity DW (billions)	# Total Words	Page Familiarity TW (billions)
6	http://hcvhub.deusto.es/	122	6.71	194	8.57
7	http://www.malacards.org/card/geniculate_herpes_zoster...	677	2.39	2047	3.13
15	http://www.malacards.org/card/yaws...	2815	0.87	6994	1.82
37	http://www.malacards.org/card/klippel_feil_syndrome_3_autoso...	314	3.08	1095	2.83
39	http://www.malacards.org/card/chorioretinitis...	2374	0.85	5592	1.67
54	http://www.malacards.org/card/spindle_cell_hemangioma...	555	2.51	1597	3.29

(c)

Fig. 3. First six test results for Patient (a), Clinican (b), and MedicalResearcher (c) audience types.

Next, we have computed some statistics related the term familiarity indexes of the Web pages for the different target audiences and we have obtained the results reported in Fig. 4. It shows, for each specific type, the box plot of the average of the term familiarity indexes computed for all words (page familiarity). A box plot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). Overall, the median and the first-third quartile interval of *Patient* is much higher of those of *Clinician* and *MedicalResearcher* that partially overlap. The outliers above the maximum mainly refer to pages that contain informative/commercial data for the different types of users and then use a simple language. The outliers below the "minimum" mainly refer to pages, such as those of the www.malacards.org domain, which indicate all three classes, as target audiences, but have a low term familiarity index clearly indicating that they should be targeted only to medical experts for what concerns the language complexity.

The experimental results show that the Web pages targeted to *Patient*, present, on average, a much higher term familiarity index and thus a simpler terminology whereas the Web pages targeted to *Clinician* and *MedicalResearcher* present, on average, a lower term familiarity index and thus a more complex terminology, even though *Clinician* pages are a little closer to *Patient* pages. As a consequence, *Patient* pages, falling in the intervals shown in Fig. 4, can be used for the Non-expert class and *Clinician/MedicalResearcher* pages, falling in the intervals shown in Fig. 4, can be used for the Expert classes producing then the following mapping:

- Non-experts -> *Patient*
- Experts -> *Clinician* and *MedicalResearcher*

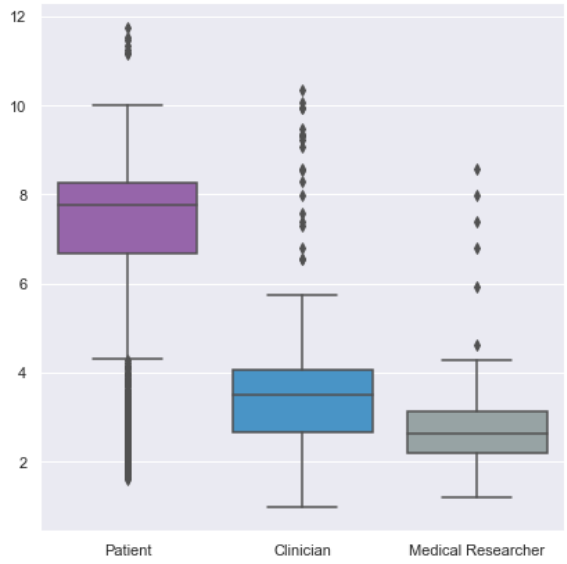


Fig. 4. Box plot of the average of term familiarity indexes for all words (computed in [6]).

This allows us to provide different types of users with health Web pages targeted to their specific language complexity requirements. Notice that the presence of structured data inside a Web page can also be seen, somehow, as a basic guarantee of information quality even though an evaluation of the quality level of a Web page content requires a specific analysis that is outside the scope of this work.

5 FACILE architecture and implementation

Once created the mapping model, as shown in the previous Section, we have built a meta search engine, FACILE, that provides the different audience types with the proper Web contents in terms of language complexity. The meta search engine can be accessed at the address http://www.math.unipa.it/simplehealth/facile and Fig. 5 reports the input interface of the engine. Notice that it provides the user with two search possibilities:

- A **Search on semantic Web (schema.org)** that allows a real-time search by using the health-lifesci.schema.org URLs analysed in the previous sections and allows to specify the audience type, i.e., non-expert (Patient) or expert (Clinician or Medical Researcher);
- A **Search on Google** that uses the Google search engine in order to explore the whole Internet and find the Web pages related to the searched keyword(s) and recomputes the page ranking on the basis of the term familiarity of each Web page. Since this computation takes some time, the search, in this case, is not in real time in the sense that it is not providing the user with an answer in a time comparable to that of a generic search engine. Notice that the interface allows to specify the number of Google results (maximum fifty, higher than the twenty-thirty results usually analysed by a user [26]).

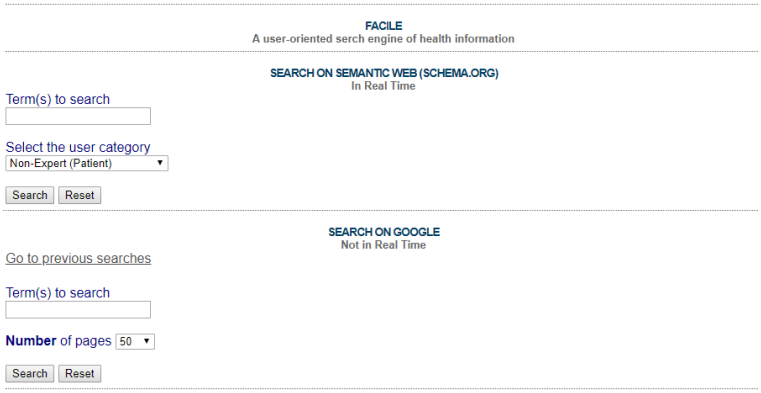


Fig. 5. Input interface of FACILE search engine.

Fig. 6 presents the Facile architecture. From top to bottom, we have the following:

- The **Client** allows to search for the medical keyword(s)**.**

- The **Search Engine** behaves slightly differently depending of the two types of search:
  - In the case of **Search on semantic Web (schema.org),** it looks for the lifesci.schema.org URLs related to the keyword(s) into the **FACILE DB** and selects the ones related to the chosen medical audience, i.e., *Patient*, *Clinician* or *MedicalResearcher*. Moreover, it provides a list of URLs sorted in terms of keyword(s) occurrences and term familiarity (see Section 5.1);
  - In the case of **Search on Google**, it first uses Google to find a number of URLs (max 50) related to the keyword. It then uses the **Web page retriever** and **Feature extractor** and loads the results into the **FACILE DB** (this operation requires some time). Finally, it provides a list of URLs sorted in terms of term familiarity (see Section 5.2).
- The **FACILE DB**, contains the the information related to the URLs. In the case of the **Search on semantic Web (schema.org)**, each URL is associated to the page words and number of occurrences, the associated medical audience and the page familiarity. In the case of the **Search on Google**, each URL is only associated to the page familiarity.
- The **Web page retriever** retrieves Web pages from the Web and the **Feature extractor** extracts/computes page features such number of words, term familiarity, etc.
- The **Health-life.sci.schema.org Quads** contains the quadruples related to *Patient*, *Clinician*, and *MedicalResearcher* health-lifesci.schema.org elements.

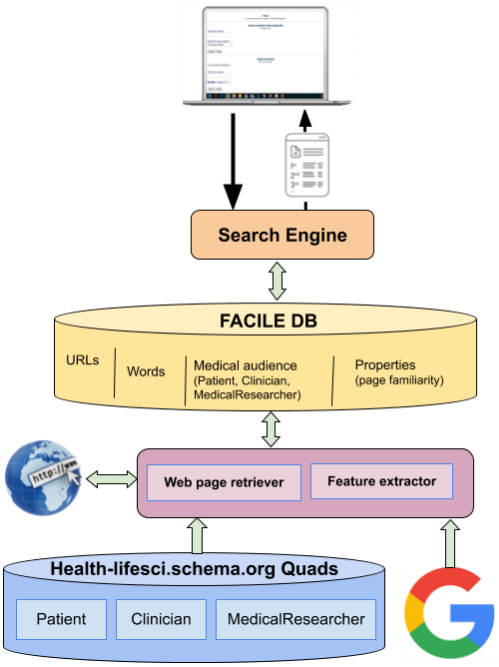


Fig. 6. FACILE Architecture

5.1 Use of FACILE with Health-lifesci.chema.org structured data

This option, as seen above, gives the user the possibility to input one or more keywords and to indicate the audience, i.e., Non-expert or Expert. The system looks for the lifesci.schema.org URLs related to the keyword(s) into the **FACILE DB** and selects the ones related to the chosen audience. It then provides a list of URLs sorted by using the following ranking formulas:

- Non-Expert (Patient)
  $R = \alpha * (Term\_Occurrences/Max\_Occurrences) + (1 - \alpha) * (Page\_Familiarity\_Index) / Max\_Familiarity\_Index)$ (1)

- Expert (Clinician and MedicalResearcher)
  $R = \alpha * (Term\_Occurrences/Max\_Occurrences) - (1 - \alpha) * (Page\_Familiarity\_Index) / Max\_Familiarity\_Index)$ (2)

  Where:
  - *Term_Occurrences* is the number of occurrences of the keyword(s) in the page;
  - *Max_Occurrences* is the maximum number of occurrences of the keyword(s) in all found Web pages;
  - *Page_Famililarity_Index* is the page familiarity, i.e., the mean of the term familiarity indexes of the Web page;
  - *Max_Famililarity_Index* is the maximum page familiarity of all found Web pages.
  - α allows us to differently weighs the number of occurrences and page familiarity.

Notice the non-expert formula is a sum because we want meaningful pages (with high number of occurrences of the searched item) but with the simplest language, whereas the expert formula is a difference because we want meaningful pages (with high number of occurrences of the searched item) but with the most complex/technical language.

We have made some preliminary experiments with the weight and found out that a value of $\alpha = 0.3$ provided us with the best results in terms of correspondence between the intended audience and the provided Web pages. For example, Fig. 7 reports the top ten results of FACILE for the 'diabetes' keyword for the Non-Expert (*Patient)* and Expert *(Clinician and MedicalReseracher)*. For each URL, the number of occurrences of the keyword (diabetes in this case), the page familiarity and the R result of the ranking formula are shown.

By examining Fig. 7 we can easily see that the top links of *Patient* present a high term familiarity index and belong to medlineplus.gov which is notoriously a Web portal for non-experts. The top links of *Clinician* present a medium-low term familiarity index and belong to the fpnotebook.com Web portal - which acts as a medical dictionary - and presents a technical language even though understandable by users with some medical skills or to malacards.org Web portal that is a human disease database and presents a very technical and complex language. The top links of *MedicalResearcher* present a low term familiarity index and belong to malacards.org Web

portal that, as said, presents a very technical and complex language. Notice that some malacards.org pages contain all the three audience types and may appear in more than one ranking (as in the case of the *Clinician* and *MedicalResearcher* web pages) because often present a high number of occurrences of the searched item. Of course, the ranking mechanism presented here is just a first proposal and needs to be refined and enriched to transform FACILE in a proper user-oriented search engine. To this end, each result page contains a link to a "detailed page" that presents, among others, the possibility for the user to choose different values of α and thus to experiments with the different ranking possibilities.

1100 PAGES FOR TERM DIABETES OF TYPE PATIENT

Back to Facile Home Page

Details page

#	URL	Occurrences	Familiarity	R
1	https://medlineplus.gov/ency/article/000305.htm	88	8.54	1.17
2	https://medlineplus.gov/ency/article/000313.htm	86	8.49	1.16
3	https://medlineplus.gov/ency/article/001214.htm	73	8.14	1.08
4	https://medlineplus.gov/ency/patientinstructions/000328.htm	40	9.25	1.07
5	https://medlineplus.gov/ency/patientinstructions/000086.htm	38	9.07	1.04
6	https://medlineplus.gov/ency/patientinstructions/000083.htm	28	8.87	0.99
7	https://medlineplus.gov/ency/patientinstructions/000322.htm	21	9.12	0.99
8	https://medlineplus.gov/ency/patientinstructions/000014.htm	4	9.74	0.99
9	https://medlineplus.gov/ency/patientinstructions/000079.htm	32	8.69	0.98
10	https://medlineplus.gov/ency/article/003640.htm	17	9.10	0.97

(a)

1148 PAGES FOR TERM DIABETES OF TYPE CLINICIAN

Back to Facile Home Page

Details page

#	URL	Occurrences	Familiarity	R
1	http://www.fpnotebook.com/Endo/DM/DbtsMlts.htm	181	3.10	0.29
2	http://www.malacards.org/card/diabetes_mellitus_permanent_neonatal	164	2.56	0.29
3	http://www.malacards.org/card/neonatal_diabetes_mellitus	128	2.09	0.22
4	http://www.malacards.org/card/monogenic_diabetes	143	2.65	0.21
5	http://www.malacards.org/card/diabetes_insipidus_neurohypophyseal	139	2.52	0.21
6	http://www.fpnotebook.com/Endo/DM/DbtcKtcds.htm	118	2.98	0.09
7	http://www.malacards.org/home/malalist/D	41	1.20	0.02
8	http://www.fpnotebook.com/Endo/DM/HyprsmlrHyprglycmcSt.htm	96	3.12	0.01
9	http://www.fpnotebook.com/Renal/Endo/CntrlDbtsInspds.htm	89	3.29	-0.03
10	http://www.malacards.org/card/microvascular_complications_of_diabetes_1	61	2.48	-0.05

(b)

515 PAGES FOR TERM DIABETES OF TYPE MEDICALRESEARCHER

Back to Facile Home Page

Details page

#	URL	Occurrences	Familiarity	R
1	http://www.malacards.org/card/diabetes_mellitus_permanent_neonatal	164	2.56	0.34
2	http://www.malacards.org/card/neonatal_diabetes_mellitus	128	2.09	0.26
3	http://www.malacards.org/card/monogenic_diabetes	143	2.65	0.26
4	http://www.malacards.org/card/diabetes_insipidus_neurohypophyseal	139	2.52	0.26
5	http://www.malacards.org/home/malalist/D	41	1.20	0.03
6	http://www.malacards.org/card/microvascular_complications_of_diabetes_1	61	2.48	-0.02
7	http://www.malacards.org/card/diabetic_polyneuropathy	43	2.18	-0.06
8	http://www.malacards.org/search/results/sepsis?retired=1	6	0.91	-0.07
9	http://www.malacards.org/search/results/ADP	7	0.96	-0.07
10	http://www.malacards.org/search/results/nephritis?retired=1	8	1.03	-0.07

(c)

Fig. 7. Diabetes outputs for *Patient* (a), *Clinician* (b), and *MedicalResearcher* (c).

5.2 Use of FACILE with Google

The use of structured data related to the intended audience, in combination with the term familiarity of a Web page, provides a method for ranking Web pages in terms of the complexity level of the text. Generalising this approach, the term familiarity analysis can be used for ranking Web pages even when they do not contain any specific structured data about their audience. The **Search on Google** section of the FACILE meta search engine follows this approach by re-ranking the results, obtained through the generic Google search engine, in terms of page familiarity.

An example of this approach is shown in Fig. 8. The results for the "Antibiotics" search keyword in Google, are ranked according to the page familiarity, as provided by FACILE. The table reports each URL with the Google rank and the page familiarity. URLs are ranked by taking into account the page familiarity value of the corresponding Web page, from the highest to the lowest.

The results present a background colour that gives an indication of the intended audience. In particular, the green colour is used to highlight URLs that have a value of page familiarity above 6 that, as seen in Fig. 4, somehow indicates pages suitable to Non-Expert audience. The yellow color is used to indicate URLs that have a value of page familiarity between 5 and 6 and is related to an interval that lies between the Expert or Non-Expert "zone". The red color is used to indicate URLs that have a value of page familiarity below 5, indicating Web pages more suitable, in principle, to an Expert audience.

In our example, the top result is a web page that explains, in lay terms, what antibiotics are and how they work. The other top results of the list refer to nhs.uk and medicalnewstoday.com domains and also represent Web pages with information for non-expert users. On the opposite, the Web pages appearing at the bottom of the list

are related to concepts such as *Tetracycline* and the *Timeline of antibiotics* that use a language more suitable for experts. It is interesting to note how ranking the results according to the term familiarity notably changes the order of the resulting URLs.

15 PAGES FOR TERM ANTIBIOTICS

#	URL	Googe Ranking	Familiarity
1	https://microbiologysociety.org/members-outreach-resources/outreach-resources/antibiotics-unearthed/antibiotics-and-antibiotic-resistance/what-are-antibiotics-and-how-do-they-work.html	14	8.55
2	https://www.medicalnewstoday.com/articles/10278.php	1	8.36
3	https://www.nhs.uk/conditions/antibiotics/	12	8.30
4	https://www.drugs.com/article/antibiotics-and-viruses.html	10	8.06
5	https://www.emedicinehealth.com/antibiotics/article_em.htm	15	7.85
6	https://medlineplus.gov/antibiotics.html	11	7.85
7	https://www.nhsinform.scot/tests-and-treatments/medicines-and-medical-aids/types-of-medicine/antibiotics	13	7.77
8	https://www.drugs.com/article/antibiotics-for-uti.html	8	7.35
9	https://www.drugs.com/drug-class/glycopeptide-antibiotics.html	9	6.87
10	https://www.drugs.com/article/antibiotics.html	7	6.74
11	https://en.wikipedia.org/wiki/Antimicrobial	6	6.19
12	https://en.wikipedia.org/wiki/Tetracycline_antibiotics	5	5.61
13	https://en.wikipedia.org/wiki/Antibiotic	2	5.32
14	https://en.wikipedia.org/wiki/Timeline_of_antibiotics	4	4.51
15	https://en.wikipedia.org/wiki/List_of_antibiotics	3	4.11

Fig. 8. Re-ranking Google search for the keyword “Antibiotics”.

6 Conclusions

The World Wide Web has more and more become the privileged source for an increasingly number of people looking for health information. The typologies of available information are able to satisfy the needs of different types of users, with different levels of expertise. The wide range of information, from practical suggestions to scholarly papers, matches the requirements of both experts and not experts when it comes to using the Web for health information seeking. However, generic and specialized search engines are not able to immediately and easily provide information to different audience types while, at the same time, exploiting all the health/medical information contained in the Web.

In this work, we have identified the main requirements related to health information seekers on the Web and have proposed an approach to classify Web pages in the health domain that satisfies the language complexity requirement. The proposed approach is based on structured data on the Web. In particular, the *schema.org* vocabulary and, more specifically, the types and properties of its *health-lifesci* extension have been used to classify health Web pages according to the different audience types.

The use of structured data in combination with the evaluation of the term familiarity index has led to a mapping between the language complexity user requirement and the different audience types. Preliminary experiments have been conducted to validate this approach and creating a mapping model. The results of those experiments have guided the design of a meta search engine that allows different users to find Web pages related to their language complexity requirements.

The performed texts and experiments have provided us with satisfying results but a more comprehensive set of tests needs to be undertaken for a evaluating more effectively the correlation between language complexity levels and the different audience types, thus, better identifying the thresholds for what concerns the term familiarity index of a Web page that led to classify the Web page as suitable for experts or non-experts. Moreover, the ranking mechanism of the meta search engine presented here should be refined in order to weight the term familiarity index in combination with the number of the keyword(s) occurrences and other parameters related to further user requirements. The time for re-ranking the Google results also needs to be optimized so to to provide users with results in real or near-real time.

Finally, other user requirements, such as the quality of information and the information classification/customization, have to be taken into account and other types and properties of the schema.org vocabulary have to be included in the proposed method in order to provide users with on-line resources that satisfy the different user requirements and allow them to easily acquire, comprehend and learn health/medical information by exploiting the Web [26], [27], [28].

Acknowledgements

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754489 and by Science Foundation Ireland grant 13/RC/2094 with a co-fund of the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie).

References

1. Taylor, H. 2010. HI-Harris-Poll-Cyberchondrics. Harris Interactive. https://theharrispoll.com/the-latest-harris-poll-measuring-how-many-people-use-the-internet-to-look-for-information-about-health-topics-finds-that-the-numbers-continue-to-increase-the-harris-poll-first-used-the-word-cyberch/.
2. Fox, S.; Duggan, M. 2013. Health Online 2013. Pew, Research Center's Internet & American Life Project, http://www.pewinternet.org/2013/01/15/health-online-2013/.
3. Spink, A. et al. A study of medical and health queries to Web search engines. 2004. Health Information and Libraries Journal, 21, 44–51.
4. Keselman, A.; Browne, A.C.; Kaufman, D.R. 2008. Consumer health information seeking as hypothesis testing. J. Am. Med. Inform. Assoc, 15, 4 (2008), 484– 495
5. Luo, G.; Tang, C.; Yang, H.; Wei, X.2008. MedSearch: A specialized search engine for medical information retrieval. In Proc. CIKM '08.

6. Alfano, M.; Lenzitti, B.; Taibi, D. and Helfert, M. 2019. Facilitating Access to Health Web Pages with Different Language Complexity Levels.In Proceedings of the 5th International Conference on Information and Communication Technolo-gies for Ageing Well and e-Health - Volume 1: ICT4AWE, ISBN 978-989-758-368-1, pages 113-123. DOI: 10.5220/0007740301130123.
7. Kummervold E., Chronaki C.E., Lausen B., Prokosch H.U., 2008. eHealth Trends in Europe 2005-2007: A Population-Based Survey. J Med Internet Res., Vol. 10.
8. Instituto Nacional de Estadística. 2010. Encuesta sobre Equipamiento y Uso de Tecnologías de la Información y Comunicación en los hogares.
9. UK national statistics, 2010. Statistical bulletin: Internet Access 2010. Office for National Statistics. 27 Aug 2010.
10. Pletneva, N., Vargas, A. & Boyer, C., 2011. Requirements for the general public health search. Khresmoi Public Deliverable D8.1.1.
11. Keselman, A. & Slaughter, L., 2007. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. Proc. AMIA Annual Symposium Proceedings, pp.399–403.
12. Alfano, M., Lenzitti, B., Taibi, D., Helfert, M.: Provision of Tailored Health Information for Patient Empowerment: An Initial Study. In: CompSysTech'19: 20-th International Conference on Computer Systems and Technologies, June 21–22, 2019, University of Ruse, Bulgaria. ACM, New York, NY, USA.
13. Banna, S., Hasan, H. & Dawson, P., 2016. Understanding the diversity of user requirements for interactive online health services. International Journal of Healthcare Technology and Management, 15(3).
14. Roberts T. 2017. Searching the Internet for Health Information: Techniques for Patients to Effectively Search Both Public and Professional Websites. SLE Workshop at Hospital for Special Surgery Tips For Evaluating the Quality of Health, 1–12.
15. W. Pian, C.S.G. Khoo, J. Chi. 2017. Automatic classification of users' health information need context: Logistic regression analysis of mouse-click and eye-tracker data. Journal of Medical Internet Research, 19(12). https://doi.org/10.2196/jmir.8354.
16. Pang, P. C.-I.; Verspoor, K.; Pearce, J.; Chang, S. 2015. Better Health Explorer: Designing for Health Information Seekers. In OzCHI '15 Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (pp. 588–597). https://doi.org/10.1145/2838739.2838772.
17. Keselman, A., Logan, R., Smith, C. 2008. Developing informatics tools and strategies for consumer-centered health communication. Journal of the American Medical Informatics Association: JAMIA, 15(4), 473–483. https://doi.org/10.1197/jamia.M2744.
18. Ardito, S. C. 2013. Seeking Consumer Health Information on the Internet, 37(4), 1–5. Retrieved from http://www.infotoday.com/OnlineSearcher/Articles/Medical-Digital/Seeking-Consumer-Health-Information-on-the-Internet-90558.shtml.
19. Alfano, M.; Lenzitti, B.; Lo Bosco, G.; Perticone, V. 2015. An Automatic System for Helping Health Consumers to Understand Medical Texts, Proc. of HEALTHINF 2015, Lisbon, pp. 622-627.
20. Alfano, M., Lenzitti, B., Lo Bosco, G., and Taibi, D., 2018. Development and Practical Use of a Medical Vocabulary-Thesaurus-Dictionary for Patient Empowerment. Proc. of ACM International Conference on Computer Systems and Technologies (CompSysTech'18), Ruse.
21. Dietze S., Taibi D., Yu R., Barker P., d'Aquin M., 2017. Analysing and Improving Embedded Markup of Learning Resources on the Web. Proc. of the 26th International Conference on World Wide Web Companion (WWW '17 Companion). International World Wide

Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 283-292. DOI: https://doi.org/10.1145/3041021.3054160.

22. Taibi, D., Fetahu, B., Dietze, S. 2013. Towards integration of Web data into a coherent educational data graph. WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web, 419-424.
23. Meusel, R., Petrovski, P., and Bizer, C. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. Proc. of the 13th International Semantic Web Conference (ISWC14), Springer-Verlag New York, USA, 277-292.
24. Kloehn, N. et al., 2018. Improving consumer understanding of medical text: Development and validation of a new subsimplify algorithm to automatically generate term explanations in English and Spanish. Journal of Medical Internet Research, 20(8).
25. Leroy, G. et al., 2012. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. AMIA Annual Symposium Proceedings. pp.522–31.
26. Alfano, M.; Lenzitti, B.; Taibi, D. and Helfert, M. 2019. ULearn: Personalized Medical Learning on the Web for Patient Empowerment. Lecture Notes in Computer Science: Advances in Web-Based Learning – ICWL 2019. Springer.
27. Alfano, M., Lenzitti, B., and Lo Bosco, G., 2015. U-MedSearch: A Meta Search Engine of Medical Content for Different Users and Learning Needs. Proc. of International Conference on e-Learning (e-Learning'15), Berlin.
28. Alfano, M., Lenzitti, B., and Lo Bosco, G., 2014. A Web search methodology for health consumers, Proc. of ACM International Conference on Computer Systems and Technologies (CompSysTech'14), Ruse, pp. 150-157.