Natural Language Question and Answer System on Semantic Health Website

Dini Handayani[1], Mira Kartiwi[2]

Department of Information Systems, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia

dini.handayani@gmail.com, mira@iium.edu.my

***Abstract*— Web search by using generic search engines is the most common activity on the Web. Despite a wide range of information one could gathered from the Internet, there has not been many websites reported providing facility for the user to query the information using a natural language. It is the aim of this research to develop a prototype that implements query technique based on natural language processing to answer such query, extraction and summarization technique to automatically generate paragraph level for a better time saving process. In early 2007, Cervical Cancer Organization reported that a female-related disease, particularly cervical cancer, is the 5th most common cancer in women worldwide. According to that report, the aim of this research is to ensure that the web-based tool developed in this study facilitate the knowledge sharing activities among medical practitioners and public users.**

Keywords*— *Semantic Technology, Ontology, Natural Language processing, Question and Answer, Cervical Cancer

I. Introduction

The advent of Internet has changed the way people live and search for information. Web search by using generic search engines is the most common activity on the Web [1]. A recent research among European citizens [2] indicate that 71% of internet users had used the internet for health purpose, and the most active health users among those were women. Indeed in US, a recent study by Atkinson et al. (2009) [3] indicated that those seeking health information were more likely to be women, aged between 35-49 years. Their study also highlighted three online behaviors among Internet users; these are searching for health information for oneself, participating in a support group for those with similar health or medical conditions, and purchasing medicine or vitamins.

The aforementioned findings support earlier studies by Kummervold et al. [4] that identified Internet as potential media to help increase health and health care awareness among female Internet users.

A study done by Andreassen et al. [2] and G. L. Kreps and L. Neuhauser [5] outlined that the use of Internet in health matters are to find the information, and come to decide whether to see a doctor and to prepare for and follow up on doctor appointment.

Study by Andreassen et al. [2] have found that use of internet does affect patients' use of other health services, but it would appear to supplement rather than to replace medical practitioners'. G. L. Kreps and L. Neuhauser [5] conclude their study that Internet is an important tool with the potential to improve information dissemination and perhaps to improve quality of health care delivery and outcomes and reduce health care errors. One of the aims of this study is to develop a prototype that allows individuals to query information from the website using natural language processing. It is important to emphasize that the prototype being developed in this research is not intended to be used as a replacement for medical practitioners' services.

In early 2007, Cervical Cancer Organization [6] reported that female related diseases, particularly cervical cancer, are the 5th most common cancer in women worldwide. With approximately 471,000 new cases diagnosed each year. In less developed countries, cervical cancer has been renowned as the second most common cause of women mortality after breast cancer for women between 20 and 39 years old. 80% of the cases occur in low income or middle-income countries. Infection by the human papillomavirus is considered as the central risk factor for cervical cancer [7].

Such distressing phenomenon mainly had been attributed to, among others, lack of awareness of both preventive actions and symptoms of cervical cancer [6].

According to that report, the aim of this research is to ensure that the web-based tool developed in this study facilitate the knowledge sharing activities among medical practitioners and public users. This tool also allows medical students in acquiring knowledge of typical enquiries on female related diseases made by patients. It is hoped that this tool will give awareness for the whole community to protect themselves and as a prevention measure against cervical cancer.

Next section is a review on related literature, followed by research methodology, illustration of use and application of semantic technology and natural language on female related diseases, section five will describe about the implication, followed by section six which is limitation, section seven is future research, and lastly section eight is conclusion.

II. Related Work

A. A Brief review about Semantic Web

Information searching using keywords on a search engine often returns too many results that do not contain the information that we need. The vagueness of the results needs to be resolved by semantically annotating resources on the Web, only results with high relevancy to the topic will be retrieved [8]. The current search engine is designed for human consumption; machines are unaware of the actual context and content meaning of different web [9].

The main challenge for the current search engine is to evolve towards a Semantic Web, where information may have explicit semantics, allowing humans and machines to make better use of information, and better integrate available data [10]. Tim Berners-Lee coined the solution through Semantic Web as an extension of the current World Wide Web that does not only provide information at the syntactic level to human users, but also at a machine-understandable [11].

B. Ontology at a Glance

Ontology supports collaboration in multi discipline of knowledge. The aim of this research is to develop the web-based tool to facilitate the knowledge sharing activities among medical practitioners and novice users who have different level of knowledge. In doing so, researchers cooperate with medical doctors to get more information on female related diseases in the point of view of a doctor as medical practitioner, and a doctor who can give predicted basic question that might arise from a novice user. The ontology has been constructed in our previous work [8]. This study also supported the previous work done by Lv et al. [12] that the complex product development needs multi-disciplinary knowledge; only a few designers with single domain knowledge cannot carry it out.

C. Natural Language Question and Answer System

Medical practitioners have many questions when caring for patients, and frequently need to seek answers for their questions. Speed is an important issue while searching information on the internet, the user will be satisfied while able to access the system in a sort of time. The current information retrieval systems such as PubMed, typically return a list of documents in response to a user's query. On the other hand, question answer techniques are based on automatically analyzing thousands of electronic documents to generate short text answers in response to clinical question that are posed by physicians or other users. The number of returned document from information retrieval system is large and it is time consuming to read it one by one [13].

The aim of this study is to enable question and answer using natural language processing techniques to query all type of medical information, extraction and summarization technique to automatically generate paragraph level for a better time saving process. In this study, part-of-speech technique was implemented. The question from the user was sent into the part-of-speech module, to generate only noun text, as the keyword to get the answer.

Additionally, Ohshima et al. [1] found that women tend to perform more frequent search in question answer sites, this statement support our research due to the target users of our prototype. The target users of our research are women who are seeking personal healthcare information and treatment options, or women who are trying to improve the quality of their lives and overall health.

III. Research Methodology

The methodology used for this study is action research (AR) methodology. "AR is an approach to research that aims both at taking action and creating knowledge or theory about that action" [14]. AR is a combination between theory and practice in one research study. The emphasis of AR is more on what a researcher do rather than on what the researcher say they do [15].

Although there are many types of action research model, the most appropriate model used to develop prototype for implementing semantic technology and natural language processing on female related disease is an action research model defined by Susman and Evered (as cited in [16]). The first step for each cycling is Diagnosing, followed by; Action planning, Action taking, Evaluating, and lastly Specifying learning. The action research cycle can continue, whether the prototype is successful or not. It continues until the problem is solved.

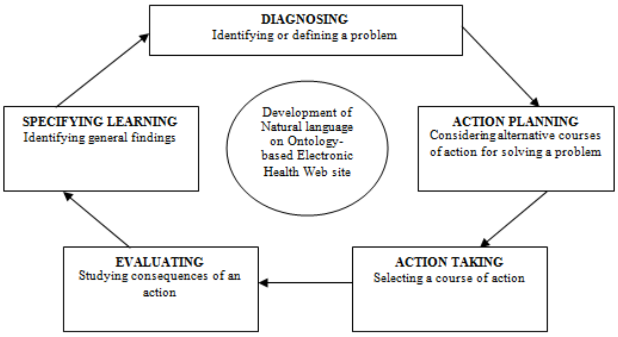

DIAGNOSING
Identifying or defining a problem
ACTION PLANNING
Considering alternative courses of action for solving a problem
ACTION TAKING
Selecting a course of action
EVALUATING
Studying consequences of an action
SPECIFYING LEARNING
Identifying general findings
Development of Natural language on Ontology-based Electronic Health Web site

Figure 1 Action Research cycle

Source: Adapted from [16]

IV. Illustration of Use and Application of Semantic Technology and Natural Language on Female

To analyze the effectiveness of the proposed prototype, interview and observation are conducted to the groups of participants. Participants in this research can be divided into two groups, participants with medical background and participants with non-medical background. Participants with medical background have been selected because of their knowledge in medical field. With that ability, the participant is expected to give comment or suggestion related to the contents of the information provided in the prototype.

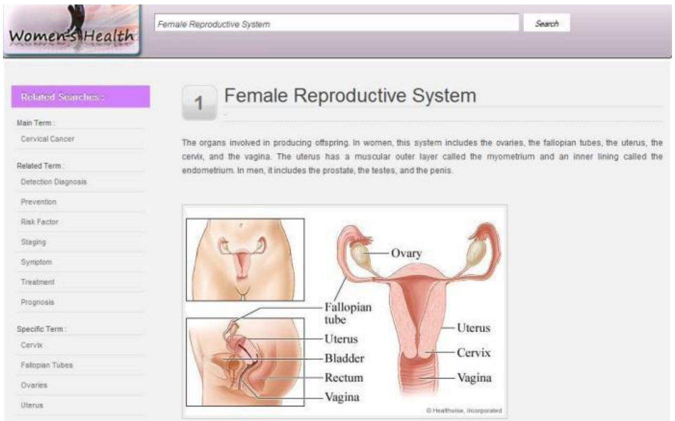
The expected feedbacks from both groups are in order to assess the effectiveness of the website including the effectiveness of feature, functionalities, and content. Moreover the expected feedbacks are in the easiness of learnability and understandability of the website, especially for the novice user. The principle of collaborative participants from different background is to get feedback and idea from each person as potential input to enhance prototype development.

A. *Prototype as support tool for participant with Medical Background*

Medical practitioner groups; consist of general medical doctors, nurses and medical students. Participants with medical background group, consist of medical practitioner with IT background (a) and medical practitioner with non-IT background (b).

a) *Participant with IT Background*

During the observation session, the participant entered a variety of types of questions such as: to get advice on a treatment (e.g. using "how" keyword), to get definition on certain disease (e.g. using "what" keyword), and other question to test the implementation of semantic technology and NLP (Natural Language Processing). The respondent also entered a question that would give an image as part of its answer, as depicted in Figure 2, as such; the prototype developed in this study is different from the existing application, such as askMEDLINE, which is only upon emphasis on the result on text rather than image [17].



Female Reproductive System

The organs involved in producing offspring. In women, this system includes the ovaries, the fallopian tubes, the uterus, the cervix, and the vagina. The uterus has a muscular outer layer called the myometrium and an inner lining called the endometrium. In men, it includes the prostate, the testes, and the penis.

Figure 2 Observation of Respondent in A group

b) *Participant with Non IT Background*

The respondent queried the prototype using medical terminologies, to check whether this application is suitable for medical practitioner. Interestingly, while in the prototype the term "period" refers to menstruation, in medical jargon it can be "staging", and in this observation what the respondent meant is "staging" as depicted in Figure 3.

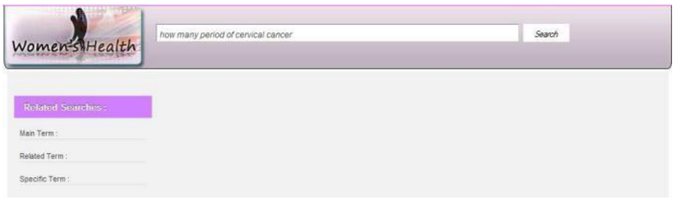


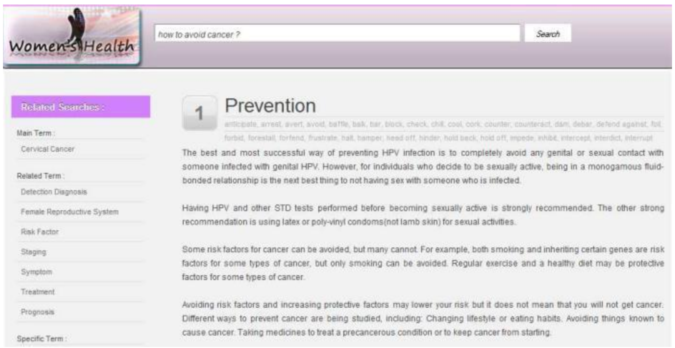
Figure 3 Observation of Respondent in B group

B. *Prototype as support tool for participant with Non-Medical Background*

Participants with non-medical doctor consists of a group of women, with different background of education, IT background (c) and non-IT background (d) later we call this group as novice user.

1) *Participant with IT Background*

The respondents typed in different kind of questions to test the implementation of semantic technology and NLP. As previously discussed in literature review, GoWeb Search engine cannot answer question types like manner of action, degree or interval ('how many', 'how much' or 'how long') and procedure ('how to') (Tomuro and Lytinen, 2001) as cited in [18]. The figure 4 shows that the prototype can answer the ('how') question. This kind of question is a bit similar with definition question type, which is given a term as a back end to process the question.



Prevention

anticipate, arrest, avert, avoid, baffle, balk, bar, block, check, chill, cool, cork, counter, counteract, dam, debar, defend against, foil, forbid, forestall, forfend, frustrate, halt, hamper, head off, hinder, hold back, hold off, impede, inhibit, intercept, interdict, interrupt

The best and most successful way of preventing HPV infection is to completely avoid any genital or sexual contact with someone infected with genital HPV. However, for individuals who decide to be sexually active, being in a monogamous fluid-bonded relationship is the next best thing to not having sex with someone who is infected.

Having HPV and other STD tests performed before becoming sexually active is strongly recommended. The other strong recommendation is using latex or poly-vinyl condoms(not lamb skin) for sexual activities.

Some risk factors for cancer can be avoided, but many cannot. For example, both smoking and inheriting certain genes are risk factors for some types of cancer, but only smoking can be avoided. Regular exercise and a healthy diet may be protective factors for some types of cancer.

Avoiding risk factors and increasing protective factors may lower your risk but it does not mean that you will not get cancer. Different ways to prevent cancer are being studied, including: Changing lifestyle or eating habits. Avoiding things known to cause cancer. Taking medicines to treat a precancerous condition or to keep cancer from starting.

Figure 4 Observation of Respondent in C group

Tomuro and Lytinen (2001) as cited in [18] in their previous study, found that GoWeb search engine cannot answer the interval question ('how long'), as depicted in Figure 5, the prototype cannot answer the same type of question. When the answer relate to absolute time, the NLP can process an answer to the question, however for relative time, it requires a deeper semantic interpretation (Niu et al.,2003) as cited in [19].

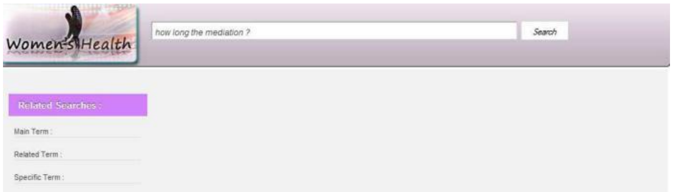


Figure 5 Observation of Respondent in C group

Figure 6 shows that the prototype cannot answer the ('how') question. The term from the question is ambiguous; therefore the process cannot give the result.

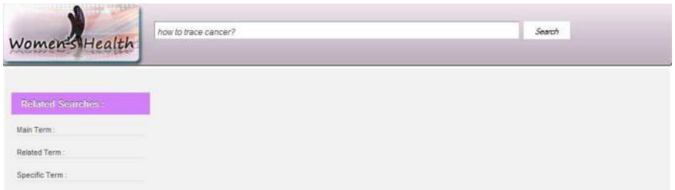
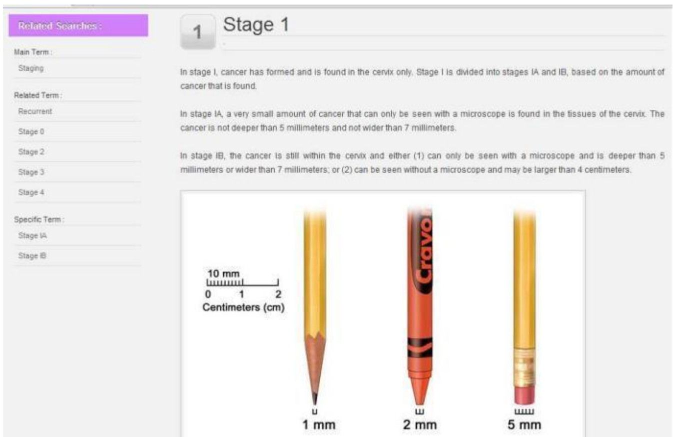


Figure 6 Observation of Respondent in C group

2) *Participant with Non IT Background*

Through the observations it was found that most of the activities conducted by the respondent with non-medical and non IT background were typing a question based on a keyword, read the result with most of them followed a link of the information in related searches. The questions that often arise from this group are related to symptoms and staging.

In observation session of participant in-group D, their behaviour reflects the statement on previous study done by Andrenucci [19] that some question requires an answer beyond the question. After obtaining the answer of the question, participant in group D continued to ask more questions by clicking related term on related searches column. As depicted in Figure 7 – 10.



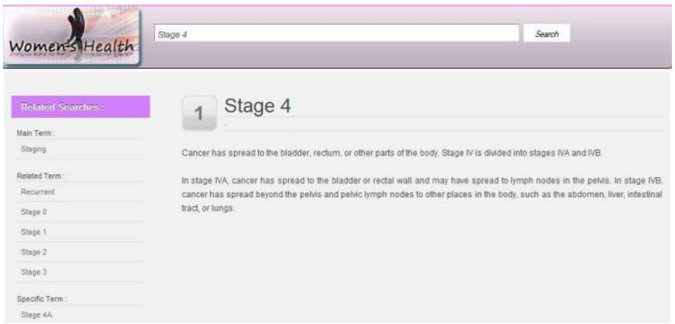
In stage I, cancer has formed and is found in the cervix only. Stage I is divided into stages IA and IB, based on the amount of cancer that is found.

In stage IA, a very small amount of cancer that can only be seen with a microscope is found in the tissues of the cervix. The cancer is not deeper than 5 millimeters and not wider than 7 millimeters.

In stage IB, the cancer is still within the cervix and either (1) can only be seen with a microscope and is deeper than 5 millimeters or wider than 7 millimeters; or (2) can be seen without a microscope and may be larger than 4 centimeters.

Figure 7 Observation of Respondent in D group

As it can be seen in Figure 8 participant in group D continued by clicking on "Stage 4" term in related searches column

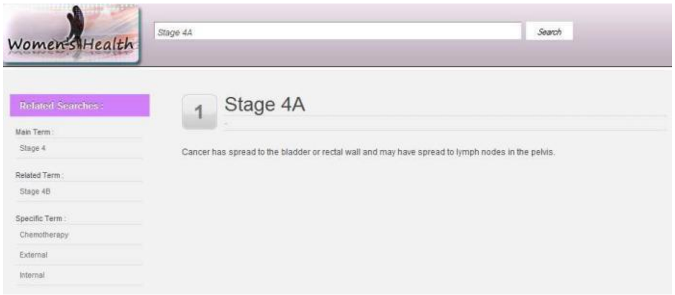


Cancer has spread to the bladder, rectum, or other parts of the body. Stage IV is divided into stages IVA and IVB.

In stage IVA, cancer has spread to the bladder or rectal wall and may have spread to lymph nodes in the pelvis. In stage IVB, cancer has spread beyond the pelvis and pelvic lymph nodes to other places in the body, such as the abdomen, liver, intestinal tract, or lungs.

Figure 8 Observation of Respondent in D group

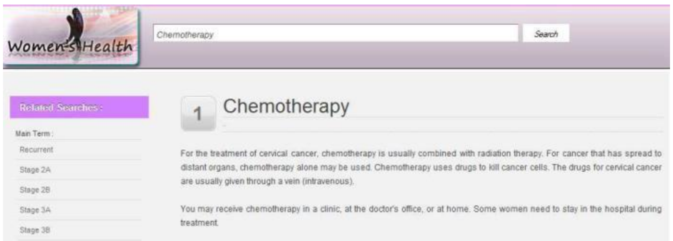
Again participants in group D continued with click on "Stage 4A" in the related searches after reading the answer from the previous question (Figure 9).



Cancer has spread to the bladder or rectal wall and may have spread to lymph nodes in the pelvis.

Figure 9 Observation of Respondent in D group

Participant in group D continued by clicking "Chemotherapy" in the related searches after reading the answer from previous question as depicted in Figure 10.



For the treatment of cervical cancer, chemotherapy is usually combined with radiation therapy. For cancer that has spread to distant organs, chemotherapy alone may be used. Chemotherapy uses drugs to kill cancer cells. The drugs for cervical cancer are usually given through a vein (intravenous).

You may receive chemotherapy in a clinic, at the doctor's office, or at home. Some women need to stay in the hospital during treatment.

Figure 10 Observation of Respondent in D group

V. IMPLICATIONS

This section explains the implication of the research on female related diseases focusing on cervical cancer. There are several implications that can be drawn from the results of this research. The theoretical as well as practical implications are highlighted in the following section.

A. *Theoretical Implications*

The study makes important contributions in developing a prototype that implement semantic technology and NLP on female related diseases. Initially, the study offers theoretical explanation that requires semantic technology on the current website.

Moreover, the study also gives the explanation that the lack of awareness on both preventive measures and symptoms are the reason a lot of women were infected by cervical cancer.

B. *Practical Implications*

The study applies the semantic technology and NLP technique on the prototype. Positive result of the finding obtained during interviews and observations also provide strong implication of the proposed technique for the developed prototype on female related disease.

Furthermore, the research has shown that the user is more satisfied while using our prototype that applied semantic technology. In fact, among the benefits highlighted by the respondents were time savings and more effective results.

Moreover, the implementation of NLP technique on the website is not the focus for some respondents. The respondents queried the question using keyword even though the researcher has given a brief overview on NLP.

VI. LIMITATIONS

There are several limitations on this study that should be noted. Firstly, the study was conducted on cervical cancer; it is limited to the study on female related diseases.

Secondly, for the question with absolute time, the prototype cannot answer the question. In medical domain, the question about time is more on relative time, which requires a semantic analysis for answering such question.

Thirdly, the researcher utilized an existing ontology from NCI (National Cancer Institute) [20] which is nciOncology follow the needs of prototype being developed in this study. The existing ontology cannot fulfill the needs of both medical practitioner and novice user for some terms, some of the terms were only provided in plain language while others in medical jargon.

Lastly, the respondents were limited to both women with medical background and women with non-medical background. The findings of this study may not generalize and draw overall conclusion to male perception, as for the female respondent the researcher assumed that they have basic knowledge about female related disease. Due to time constraint, only female were chosen as respondent for this study.

VII. Future Research

The study is limited to the implementation of semantic technology and NLP in cervical cancer. Thus, future study should focus more on female related diseases. The study applied the qualitative approach using interviews and observations for data collection; therefore another approach for future study could begin by applying a different data collection technique.

Furthermore, some of the question that arises on medical domain in the context of time needs answers that relate to interval or relative time. The current prototype cannot provide such answer. Future research is needed to address semantic analysis for answering relative time questions.

Moreover, the study used the existing ontology which is nciOncology from National Cancer Institute (NCI) [20]. The lack of information for some medical jargon and plain language highlight the need for continuous improvement and modification of the existing ontology.

Future research should also aims to fulfill the needs of both medical practitioners and novice users regarding the use of term in the ontology.

Additionally, future research should be driven with more diverse demographic characteristics of the respondent research. Demographic characteristics might give different finding on the study.

VIII. Conclusion

Today, an increasing number of people have become health conscious and turned to the Internet in searching for information. Such behaviours arise due to the astronomical cost of getting information on female related disease from medical practitioners. Also, few websites use semantic technology and NLP to present information on female related diseases. We proposed a solution by developing a prototype that allows both individuals and medical practitioners to refer and enquire using natural language. This prototype also enables the web users to share knowledge on female related diseases. Thus far this study gains understanding on the implementation of semantic technology and NLP in electronic health research.

References

[1] H. Ohshima, A. Jatowt, S. Oyama, S. Nakamura, and K. Tanaka, "Towards Improving Web Search: A Large-Scale Exploratory Study of Selected Aspects of User Search Behavior," presented at the Proceedings of the 10th International Conference on Web Information Systems Engineering, PoznaD, Poland, 2009.

[2] H. K. Andreassen, M. M. Bujnowska-Fedak, C. E. Chronaki, R. C. Dumitru, I. Pudule, S. Santana, H. Voss, and R. Wynn, "European citizens' use of E-health services: A study of seven countries," vol. 7, 2007.

[3] N. L. Atkinson, S. L. Saperstein, and J. Pleis, "Using the internet for health-related activities: findings from a national probability sample," *US National Library of Medicine National Institutes of Health,* 2009.

[4] P. E. Kummervold, C. E. Chronaki, B. Lausen, H. U. Prokosch, J. Rasmussen, S. Santana, A. Staniszewski, and S. C. Wangberg, "eHealth trends in Europe 2005-2007: a population-based survey," *US National Library of Medicine National Institutes of Health* vol. 10, 2008.

[5] G. L. Kreps and L. Neuhauser, "New directions in eHealth communication: Opportunities and challenges," *Patient Education and Counseling,* 2010.

[6] CervicalCancer.org, "Cervical Cancer Statistics," ed, 2007.

[7] J. S. Smith, L. Lindsay, B. Hoots, J. Keys, S. Franceschi, R. Winer, and G. M. Clifford, "Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: A meta-analysis update," *International Journal of Cancer,* vol. 121, pp. 621-632, 2007.

[8] D. Handayani, M. Kartiwi, and S. Hakiem, "An Implementation of Natural Language Processing on Ontology-based Electronic Health Website: A Case of Female-related Diseases Domain," in *ICT4M*, Jakarta, 2010.

[9] M. N. K. Boulos, A. V. Roudsari, and E. R. Carson. (2001, Towards a semantic medical web: healthcybermap's dublin core ontology in protege-2000. *Fifth International Protégé Workshop,Sowerby Centre for Health Informatics at Newcastle , University of Newcastle upon Tyne, Newcastle, England.*

[10] C. Golbreich. (2008, Combining content-based retrieval and description logics reasoning. *multimedia.semanticweb.org*.

[11] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, and B. Parsia, "The National Cancer Institute's Thesaurus and Ontology," *imap.websemanticsjournal.org,* vol. 1, 2011.

[12] L. Kun, L. Z. Qing, and F. L. Yan, "Ontology-Based Knowledge Modeling for Collaborative Product Development " *Key Engineering Materials* vol. 455, pp. 662-666, 2010.

[13] M. Lee, J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. Ely, and H. Yu, "Beyond Information Retrieval Medical Question Answering," in *Medical Question Answering. AMIA Symposium Proceedings*, 2006, pp. 469–473.

[14] P. Coughlan and D. Coghlan, "Action research for operations management," *International Journal of Operations & Production Management,* 2002.

[15] D. E. Avison, F. Lau, M. D. Myers, and P. A. Nielsen, "Action research," *Commun. ACM,* vol. 42, pp. 94-97, 1999.

[16] P. Järvinen, "Action Research is Similar to Design Science," *Computer and Information Science,* vol. 41, pp. 37-54, 2007.

[17] P. Fontelo, F. Liu, and M. Ackerman, "askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed," *BMC Medical Informatics and Decision Making,* vol. 5, 2005.

[18] H. Dietze and M. Schroeder, "GoWeb: a semantic search engine for the life science web," *BMC Bioinformatics* vol. 10, 2009.

[19] A. Andrenucci, "AUTOMATED QUESTION-ANSWERING TECHNIQUES AND THE MEDICAL DOMAIN," in *International Conference on Health Informatics*, 2008.

[20] T. Berners-Lee and M. Fischetti, *Weaving the Web: The Original Design* and Ultimate Destiny of the World Wide Web by Its Inventor, 1st ed. Harper San Francisco, 1999.