Semantic and Traditional Website Design Systems

Nwagwu, Honour Chika, Abhadiomhen, Stanley & Okereke, George Emeka
Department of Computer Science
University of Nigeria
Nsukka, Enugu State, Nigeria
E-mail: honour.nwagwu@unn.edu.ng

ABSTRACT

The design, implementation and technology of traditional websites are different from that of the semantic websites. But unlike the traditional websites which maintain a massive heterogeneous amount of information and semi-automated search systems, Semantic websites are composed of structured data and enable automatic recognition of its data through its tools and techniques. This paper explores the advantages of implementing semantic websites over the traditional websites. It compares data processing in both traditional and semantic websites and recommends the way forward to establishing an internet system which enables users to easily retrieve needed information.

Keyword: Traditional Websites, Semantic Websites, Semantic Technologies, Information, Web data, Search

1. INTRODUCTION

Due to the huge amount of heterogeneous information available on traditionally designed websites, web users do not effectively retrieve their needed information. The heterogeneity of information in traditional web ranges from different data formats such as Hyper Text Mark-up Language (HTML), and Hyper Text Transfer Protocol (HTTP) to structured and unstructured data. The processing of these different data types are difficult to automate. However, the success story of the semantic Web has projected the existence of particular machine readable data formats which can automatically be processed by semantic applications.

Nevertheless, there is a challenge that while the semantically designed websites provide meaningful data that enhances the processing of its information, the traditionally designed websites produce data which are difficult to process and analyse. Indeed, there are issues that limit the extent to which information is processed by web tools and techniques in traditionally designed websites. Some of these issues have been explored by Ahmed, Hussain, Hameed and Ali (2012) and it is observed that although traditional search engines provide support for the automatic retrieval of information, however, due to the absence of semantics in keywords used in queries assigned to its repositories, it becomes challenging to retrieve relevant information from the search engines. Consequently, the authors of this work assess that the traditionally designed websites provide semi-automatic search techniques to its website users. For example, it is observed that different results are obtained through the use of different search engines such as Google.com , info.com , and Bing.com , when the string “types of Bus” is entered in the search engines.

The displayed result of the search engines are arranged differently and ranges from data describing types of bus as "expansion bus types e.g ISA - Industry Standard Architecture. EISA - Extended Industry Standard Architecture. MCA - Micro Channel Architecture" to data describing types of bus as "types of road vehicles such as omnibus, multibus, motorbus, and autobus". These search engines do not take into consideration, the meaning of the key words as intended by the web user in the query. Also, the information on the traditionally designed Websites lacks semantics. Consequently, the displayed results by the search engines may not be relevant to the web user when retrieved from several traditionally designed websites. This type of response from traditionally designed websites forces the user to begin to read through all the retrieved results, which more often than not is boring especially when the retrieved result is large. Sometimes, the search will not produce any results mainly because there are no exact match between the searched keyword and available keywords in the website. For example, the search for "honourable" may not recognise "Honourable" in the website

From the described challenges, this paper highlights the importance of semantic website design by reviewing the existing literature of web data processing. It reviews how information is retrieved in the traditional and semantic websites. Section two provides an overview of the traditionally and semantically designed websites. Section three reviews related works on how information can be automatically processed from the data in the differently designed websites. Section four specifies a proposal for automated retrieval and processing of data from traditionally designed websites. Finally, section five presents the conclusion and future work.

2. OVERVIEW OF THE TRADITIONALLY AND SEMANTICALLY DESIGNED WEBSITES

The traditional web is assumed to be the biggest global database (Madhu, Govardhan and Rajinikanth, 2011). It was introduced by Tim Berners-Lee in late 1989 (Pranay and Bijoy, 2015) and in 1990, Tim Berners-Lee and Robert Cailliau proposed the idea of interlinking web documents whose contents are primarily formatted text, video and images. Some underlying technologies of the traditional web are Hyper Text Mark-up Language (HTML)-for document authoring, Hyper Text Transfer Protocol (HTTP)-for document request, Universal Resource Locator (URL)-for document reference and Cascading Style Sheet (CSS)-for document Styling.

Although, the traditional web made some significant improvement with an evolution from Web 1.0 to Web 2.0 (Sareh,Mohammad and HadiKhosravi, 2012) however, the traditionally designed websites lack the semantic structure required in order to automatically retrieve and process data. This challenge arises because of the inherent limitations of the traditional web which includes:

i. The technology does not allow for precise identification of the content of its web documents. For example, it can be observed that a search for "types of bus" in a search engine may not return pages that contain only types of computer bus; it might also return minibuses and other types of road vehicles. This is because the search is restricted to keywords.
ii. The technology lacks proper mechanism to express the relationship between web documents. Unlike in semantic web documents whose Web Ontology Language can be defined by relationships such as "owl:sameAs", and "owl:differentFrom, the traditional web does not have such relationships in its documents. As a result, information from different traditionally designed websites cannot be easily combined to answer a query.
iii. The words in a traditionally designed website are usually ambiguous for existing algorithms to analyse hence; contents can only be read and interpreted by humans.

In this context, Sir Tim Berners-Lee envisioned Web 3.0-Semantic web (W3C, 2001; Allenmang and Hendler, 2011) which provides some infrastructures, for example, Resource Description Framework (RDF), Simple Protocol and RDF Query Language (SPARQL), and Web Ontology Language (OWL) that express information in formats that can be understood and processed by both machine agents and humans (Troullinou, Kondylakis, Daskalaki and Plexousakis, 2017).

2. 1 Semantic Web Technologies

The technology needed to design semantic website are available but they are not fully adopted by website developers for web designs. The rationale behind not fully adopting semantic web design approaches by website developers can be related to cost and lack of awareness. This section presents an overview of some notable semantic web technologies.

2.1.1 Semantic Web Ontology

Troullinou, G., Kondylakis, H., Daskalaki, E., Plexousakis, D. (2017) explain that one major component of the semantic web is the ontology, because, it plays an important role in fulfilling the interoperability of web data. Ontology is the conceptualization of a domain into machine-readable and human understandable format consisting of entities, instances, classes, relationships and axioms (Suqin and Zixing 2010). In the same vein, Cob and Abdullah (2008) state that ontologies are neither knowledge nor information but they are information about information employed to represent the relation holding between various terms within the information. Ontology can be used not only to annotate web-pages but; also to annotate other data source, for example, the collection of XML documents and relational databases. Ontology uses a semantic web language known as Web Ontology Language (OWL) to represent rich and complex knowledge about things, groups of things and relations between things. OWL is a computational logic-based language that allows for any knowledge stated to be analysed by computer programs either to verify the consistency of that knowledge or in order to make implicit knowledge clear. Its documents known as ontologies can be distributed on the World Wide Web and may be referenced from other OWL ontologies (Pascal, Markus, Krotzsch, Bijan, Sebastian and Rudolph (2012).

2.1.2 Resource Description Framework (RDF)

RDF is a language employed to represent meta-data about resources in the World Wide Web (WWW). It can also be used to represent information about things on the web. RDF provides a framework for information to be processed and shared between applications without loss of semantics. It uses Uniform Resource Identifier (URI) in order to identifying things and can be embedded into web pages via RDFa (Ben and Birbeck, 2012). Apart from RDFa, there are other data formats such as Open Graph Protocol (Pierfrancesco, Paolo and Alessandro, 2014) and Schema.org (Khalili and Auer, 2013) that are meaningful and used by leading organisations such as Facebook, Google, and Yahoo.

2.1.3 Simple Protocol and RDF Query Language (SPARQL)

SPARQL is a query language for RDF recommended by the World Wide Web consortium (W3C) in 2008 (Steve, Andy and Seaborne, 2013) and it is used to express a query over diverse data sources. It allows for data to be retrieved not only from one SPARQL endpoint but from multiple endpoints (Muhammad, Yasar, Ali, Ivan and Axel-Cyrille, 2014). A SPARQL endpoint is a URL to which queries can be sent to a knowledge store via SPARQL language and it returns answers to the queries in one or more machine-process-able formats (Fujino and Fukuta, 2012).

2.2 Comparative analysis of the traditionally and semantically designed websites

Table 1 presents a summary of the differences between the traditionally and semantically designed websites.

Table 1: A comparative analysis of traditional and semantic websites

SN	Attributes	Traditional websites	Semantic websites
1	**Data-Types**	Assume that individuals will convert values to appropriate data types	Explicitly indicate the data type of each value.
2.	**Data-Format**	Designed using HTML, CSS, HTTP and XHTML e.t.c.	Designed using RDF, RDFS, OWL, and RDFa embedded in triples within a human-readable XHTML document
3.	**Information processing**	The content can be easily read and interpreted by human beings.	The content can be understood and processed by both humans and machines.
4	**Information Access**	It is designed like a stovepipe system where information can only flow. Thus, information cannot be easily shared by other websites or organization that requires it.	Knowledge sharing and discovery is the key concept in the design i.e., information can be shared between websites and organizations that require it.

3. AUTOMATED WEB DATA RETRIEVAL SYSTEMS

The amount of data available on the web is overwhelming, yet most web users do not have the capabilities to easily retrieve their information needs from these data. This is because web data are mostly in varied formats which pose a challenge to automated retrieval of the data. Certainly, web users can afford to read through the entire text in a webpage or several webpages to identify their information needs, but such a technique will not enable efficient retrieval of information because it will take so much time to identify the needed information. There are however, other approaches for automating the retrieval of data that can meet the information needs of a web user. Such approaches include the use of web search tools and Scraping programs.

The web user can retrieve needed information through querying the web search engine. Govathoti and Babu (2018) note that search engines should be enriched with semantic web capabilities that analyse webpage content and provide relevant results corresponding to the user query. The web user can search within a web page using the search tool embedded in the web page. Such search engines receives a user query in form of keyword(s) and uses its search algorithm to search the webpage(s) in order to locate any associated information. Consequently, a ranked list of documents that match the user searched keyword(s) is/are returned. This is also the approach adopted by leading search engines such as Google and Yahoo.
But in traditionally designed websites, the most relevant documents do not necessarily appear at the top of the returned list of documents as a result, presenting inaccurate information to the web user where wrong meanings are associated to the searched words. For example, a web user who searches for “types of bus” and is interested in different types of vehicles may find the search engine presenting on the top of the ranked list of output documents, the types of bus in a computer system. This is unlike in semantic websites where the associated data are linked with meaningful vocabulary such as Web Ontological Languages (OWL) and Resource Description Framework Scheme (RDFS).

The second approach for automating the retrieval of data that can meet the information needs of a web user is the use of web scraping programs. Web scraping is the programming for an automated and targeted extraction of data from Web content. It can be used in mining scientific publications and bibliographic data as explained in Meschenmoser, Meuschke, Hotz and Gipp (2016). Web scraping is described in Haddaway (2015) as the use of a program to extract data from HTML files on the internet. Fernández (2011) outlines some services that make use of semantic data extracted from unannotated web resources to include opinion miners, recommenders, and mashups that index and filter pieces of news. Some notable web scraping tools include WebSundew , Easy Web Extractor , Handy Web Extractor and FMiner.

Web scraping programs are easily used in webpages containing semantic markup (Semantic webpages). In traditionally designed website, RDF data can be added to represent content in the HTML documents. It can also be extracted from the website using specialised tools. For example, Golbeck, Grove, Parsia, Kalyanpur, and Hendler (2002) describe an RDF web scraper that helps users to specify how to extract RDF mark-up from webpages. They explain that users of their tool can analyse the HTML in a webpage and create a wrapper that describes how the tag structure relates to the contents.

The RDF web scraper parses the page based on the wrapper and generates a table of data. Finally, the web scraper user as explained in Golbeck, Grove, Parsia, Kalyanpur, and Hendler (2002), can indicate the ontological specifications for each column of the table, and generates the corresponding RDF. Also Huynh, Mazzocchi and Karger (2007) implemented Piggy Bank. The Piggy bank can invoke screenscrapers to re-structure information within webpages of websites which do not publish RDF into RDF and store same in a repository. It allows Web users to extract individual information items from within webpages and save them in RDF data format thereby evolving the extraction of web data from only HTML files to varieties of data types. Unlike in traditionally designed websites, RDF embedded HTML documents enable an automated knowledge extraction and retrieval of relevant data from web documents.

Obviously, semantically designed websites have great advantages over traditionally designed websites. It is easy to retrieve relevant data that are associated with the context of the query within a semantic or across semantic webpages than a traditionally designed website. Some examples of websites that adopt some semantic technologies include DBPedia , British Broadcasting Cooperation (BBC) website and Wikipedia as described in (Ismayilov, Kontokostas, Auer, Lehmann, and Hellmann, 2018; Liu, Mikroyannidi and Lee, 2015; Lehmann et al., 2015). Even so, the cost of designing and implementing a semantic website can be huge. Also, there is little awareness among users of the web on the amazing advantages of semantically designed websites. Some of the advantages of implementing a semantic website include the following:

- Semantic data removes the challenges of homonyms
- Applications can be built to automatically process data from semantic websites for specific purposes.
- Semantic websites provide structured data for search engines, thereby saving the web user the pains of having to read through unrelated document for needed data by providing relevant search results

4. PROPOSAL: AN AUTOMATED DATA RETRIEVAL AND PROCESSING SYSTEM FOR TRADITIONALLY DESIGNED WEBSITES

The authors of this work propose a methodology for an automated retrieval and processing of data from traditionally designed websites. A compiler is attached as an add-on to a browser and web users can query the compiler through the browser for information about website(s). Web information on any traditionally designed website which is accessed by the browser is automatically compiled using RDF format and stored in a unique RDF repository in a semantic database. Such information can be queried using keywords by the web user to access reputable and sound information from the web server. Note the searched keywords are converted into a "Select search in SPARQL query" and descriptive information is returned to the web user through the web server. Figure 1 provides a pictorial description of this proposed methodology.

5. CONCLUSION AND FUTURE WORK

Web users are mostly insensitive about the efficacy of the information retrieval system of a website. Undoubtedly, a well-designed website is useful to the web user because he can easily retrieve his needed information. This is unlike a website whose data are not structured in a format that is searchable through available online technologies. This work has explored the evolving technologies in website design. The traditionally designed websites is compared to the semantically designed websites and a methodology is proposes on how the traditionally designed websites can be made semantic. The authors hope to implement their proposed methodology as explained in section 4.0. They are also exploring options on how to dynamically convert traditionally designed web sites to a semantic one.

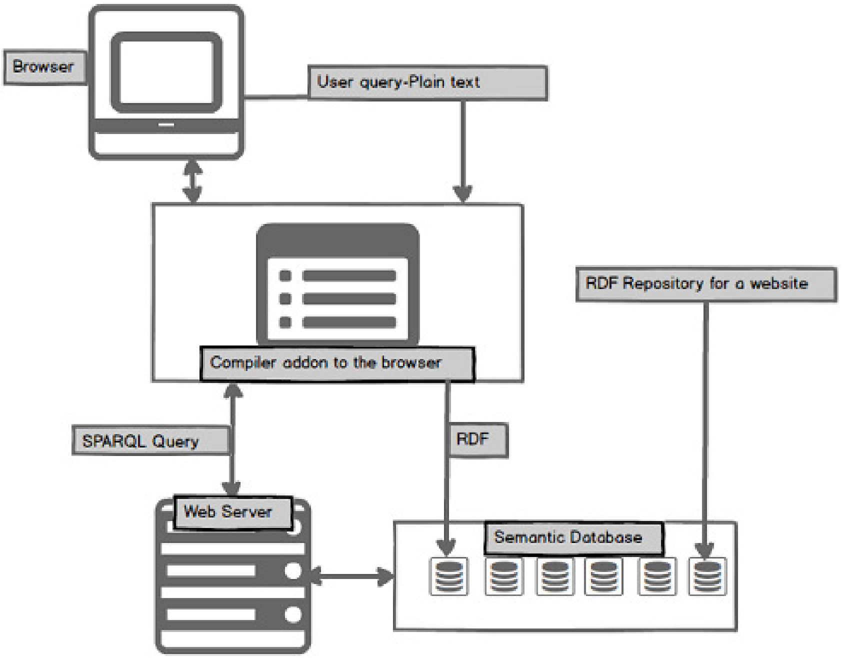

Figure 4: An automated retrieval and processing system for traditionally designed websites

REFERENCE

1. Ahmed, F., Hussain, S., Hameed, S. & Ali, S. (2012). Semantic web E-portal for tourism. Second International Conference on Digital Information and Communication Technology and its Applications (DICTAP).
2. Allenmang, D. & Hendler, J. (2011). Semantic Web for the Working Ontologist. Second ed. Effective Modeling in RDFS and OWL. USA: Morgan Kaufmann.
3. Ben, A. & Birbeck(2012). RDFa Core 1.1. Retrieved from http://www.w3.org/TR/rdfa-core
4. Cob, Z. & Abdullah, R. (2008). Ontology-based Semantic Web services framework for knowledge management system. International Symposium on Information Technology.
5. FernándezVillamor, J. I., Blasco Garcia, J., Iglesias Fernandez, C. A., &GarijoAyestaran, M. (2011).A semantic scraping model for web resources-Applying linked data to web page screen scraping.
6. Fujino, T. &Fukuta, N. (2012). A SPARQL Query Rewriting Approach on Heterogeneous Ontologies with Mapping Reliability. International Conference onAdvanced Applied Informatics
7. Golbeck, J., Grove, M., Parsia, B., Kalyanpur, A., & Hendler, J. (2002, October). New tools for the semantic web. In International Conference on Knowledge Engineering and Knowledge Management (pp. 392-400). Springer, Berlin, Heidelberg.
8. Govathoti, S., & Babu, M. P. (2018). An Implementation of a New Framework for Automatic Generation of Ontology and RDF to Real Time Web and Journal Data. IJCSIS.
9. Haddaway, N. R. (2015). The use of web-scraping software in searching for grey literature.Grey J, 11(3), 186-90.
10. Huynh, D., Mazzocchi, S., &Karger, D. (2007). Piggy bank: Experience the semantic web inside your web browser. Web Semantics: Science, Services and Agents on the World Wide Web, 5(1), 16-27.
11. Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2018). Wikidata through the Eyes of DBpedia.Semantic Web, (Preprint), 1-11.
12. Khalili, A., Auer, S. (2013). Wysiwym authoring of structured content based on schema.org. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) WISE 2013, Part II. LNCS, vol. 8181, pp. 425–438. Springer, Heidelberg.
13. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ...&Bizer, C. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2), 167-195.
14. Liu, D., Mikroyannidi, E., & Lee, R. (2014).Semantic web technologies supporting the BBC knowledge & learning beta online pages. In Proceedings of the Linked Learning Meets LinkedUp Workshop: Learning and Education with the Web of Data (LILE 2014).
15. Madhu, G., Govardhan, T. &Rajinikanth. (2011). Intelligent Semantic Web Search Engines: A Brief Survey. Retrieved from http://airccse.org/journal/ijwest/papers/0111ijwest03.pdf.
16. Meschenmoser, P., Meuschke, N., Hotz, M., & Gipp, B. (2016).Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction.D-Lib Magazine, 22(9/10).Retrieved from http://www.dlib.org/dlib/september16/meschenmoser/09meschenmoser.print.html
17. Muhammad, S., Yasar, K., Ali, H., Ivan, E. & Axel-Cyrille, N. (2014). A Fine-Grained Evaluation of SPARQL Endpoint Federation Systems. Semantic Web Journal, 10.3233/SW-150186.
18. Pascal, H., Markus, Krotzsch.,Bijan, P., Sebastian &Rudolph. (2012). OWL 2 Web Ontology Language Primer (Second Edition).Retrieved fromhttp://www.w3.org/TR/owl-primer/

19. Pierfrancesco, B., Paolo, N., Alessandro, V. (2014). Linked Open Graph: browsing multiple SPARQL entry points to build your own LOD views, International Journal of Visual Language and Computing, 2014. Retrieved from http://dx.doi.org/10.1016/j.jvlc.2014.10.003
20. Pranay, K. &Bijoy,C. (2015). Evolution of World Wide Web: Journey from Web 1.0 to Web 4.0. International Journal of Computer Science and Technology, Vol. 6, pp. 134- 138, 2015.
21. Sareh, A., Mohammad, A., Nemat, B. &HadiKhosravi,F.(2012). Evolution of the World Wide Web: From Web 1.0 to Web 4.0". International Journal of Web& Semantic Technology (IJWesT), Vol. 3, No. 1, pp. 1-10, 2012.
22. Steve, H., Andy &Seaborne (2013).SPARQL 1.1 Query Language Retrieved from http://www.w3.org/TR/sparql11-query
23. Suqin, T. & Zixing, C. (2010). Using the format concept analysis to construct the tourism information ontology. Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).
24. Troullinou, G., Kondylakis, H., Daskalaki, E., Plexousakis, D. (2017) Ontology understanding without tears: the summarization approach. Semantic Web Journal. IOS press.
25. World Wide Web: Proposal for a HyperText Project. Retrieved from http://www.w3.org/Proposal.html
26. W3C. World Wide Web Consortium. Retrieved from http://www.w3.org
27. W3C (2001). Semantic Web Activity. Retrieved from http://www.w3.org/2001/sw

WEB Links (Footnotes)

1. https://www.google.com/
2. http://www.info.com/
3. https://www.bing.com
4. www.websundew.com
5. www.webextract.net
6. www.scraping.pro/handy-web-extractor
7. www.fminer.com
8. http://wiki.dbpedia.org/
9. https://www.bbc.com/educationn
10. https://www.wikipedia.org/