Преглед НЦД 22 (2013), 47–54

Aleksandar Mihajlović, Vladisav Jelisavčić,
Bojan Marinković, Zoran Ognjanović
Mathematical Institute of the Serbian Academy of Sciences and Arts, Serbia
Veljko Milutinović
School of Electrical Engineering, University of Belgrade, Serbia

THE SERBIA-FORUM CULTURAL HERITAGE DIGITIZATION PROJECT WITH EMPHASIS ON SEMANTIC INDEXING

Abstract: Within this paper a new electronic encyclopedic (e-encyclopedia) web application named Serbia-Forum is introduced. Serbia-Forum is compared with existing e-encyclopedias such as Wikipedia, Europeana and Austria-Forum. The limitations of the mentioned e-encyclopedias are addressed and the goals of Serbia-Forum established. Serbia-Forum is governed by two sets of guidelines referred to as "axioms". It distinguishes itself from existing electronic encyclopedias in that it focuses on the collection and digitization of national cultural heritage element and concepts of Serbia. An insight is given into how some of the axioms of Serbia-Forum can be implemented by other (encyclopedic) applications. Introduction to upcoming research in novel semantic indexing schemes which will be implemented by Serbia-Forum is provided.

Keywords: e-encyclopedia, cultural heritage, cultural artifacts, cultural elements, Serbia-Forum, sub-content

1. Introduction

The increased use of computers at homes and in the office environments naturally creates a new space for introduction of new paradigms in the domain of presentation and processing of digitized documents. In an increasingly digital and environmentally aware world, the benefits of digitization are apparent. Fitting large and numerous documents onto a PDA or a laptop are simply more economical, allowing ease of access, reducing the need for physical storage space and paper. Aside from the financially, organizationally, and environmentally economic benefits, document preservation and digitization of scholastic content gave rise to new modes of rapid knowledge acquisition, abstract learning and augmented reality experiences. Computer based interactive multimedia has a potential of completely replacing legacy learning modes based on the chalk board, hardcopy notes and books, as well as photographs and analog audio/video play-back/recording devices.

The continuous growth of the Internet's availability gave birth to a number of new and revolutionary data sharing ideas. Among these ideas is the "knowledge across the wire," whose goal initially was to reduce frequent trips to libraries, book stores, and bodega newspaper stands giving people quick and easy access to the information they need, when they need it. Due to costly and limited data storage, capacities on servers, modest bandwidth, and limited access speeds in the early stages of the Internet development, the digital transposition of conventional hard copy volumes of knowledge, such as dictionaries and encyclopedias, was reserved for standalone installations at large costs. Nonetheless, as blue-chip technology improved, so the revolution of free online electronic volumes began. Volumes of particular interest are electronic encyclopedias.

2. Problem Statement

The popularization of electronic encyclopedias can be attributed to the development of so called Wiki technology. Wiki is a two-way website technology which allows users to modify web site content via a web browser using a simplified markup language or a rich-text editor. Wikis are powered by wiki software which is sometimes developed collaboratively via evolving wiki. By means of flexibly controllable content modification rights (edit rights), wikis may serve many collaborative purposes, such as community websites dynamic maintenance, collaboration on proprietary intranets, knowledge management, note taking, etc. The most successful use of wiki technology is one engineered by Wikimedia, the founders of Wikipedia which is the fifth most popular website of 2012. It is a free online encyclopedia that has opened a new pathway for freely sharing scholastic information on the web [1]. Funded exclusively by donations (major benefactors being Dell Computers and Virgin Unite), the annual operating cost for Wikipedia in 2009 was 9.4 million dollars [2]. Wikipedia is a revolutionary application, with an astounding 330 million monthly visitors [3]. It allows dedicated internet users to upload, read, write and modify their own encyclopedic articles under some level of editorial supervision [2]. Even though Wikipedia seems to be an extremely useful source of encyclopedic and digitized content, Wikipedia and similar web applications available online have their limitations. These limitations can be relaxed through other similar and improved initiatives. Within this paper the corresponding limitations of current free e-encyclopedias are presented and a new and improved initiative is introduced known as the Serbia-Forum cultural heritage digitization project or shortly Serbia-Forum.

3. Existing Solutions and Their Limitations

Wikipedia is constrained by three particular limitations that Serbia-Forum seeks to relax: quality, structural, and legal limitations. These limitations attest to the preliminary character of the encyclopedic project itself. It focuses on decentralized, free and growing content based on dedicated user collaboration that is completely reusable by all, absolutely neutral and verifiable by all. However, considering the fact that anyone can author articles and remain anonymous, the quality of information and the credibility of the authors are questioned.

3.1 Wikipedia. Structurally, the multilingual approach of Wikipedia and similar e-encyclopedias only contributes to the universality of the initiative. However, there are several structural hurdles that can be relaxed in Serbia-Forum. First, Wikipedia has an assortment of articles. Not many books are digitized and made available on Wikipedia. Digitized books can be found on Google Books. Archival documents and other sensitive content is also insufficiently presented, (Due to costly and limited data storage capacities on servers, modest bandwidth, and limited access speeds in the early stages of Internet development, the digital transposition of conventional hard copy volumes of knowledge such as dictionaries and encyclopedias where reserved for standalone installations). Secondly, specifically for Serbia, even if there exists a significant number of articles and documents published in the Serbian language, a significantly smaller number of these semantically represent concepts related to Serbian cultural heritage. Serbia and Wikipedia are connected by only one string, the Serbian language. In the midst of insufficient representation, a need arises for the systematic collection and presentation of concepts, significant historical figures and documents related to events in Serbian history and culture. Thirdly, version tracking in Wikipedia is quite complicated to perform. Finally, semantic organization of the content is also not one of the strong points of Wikipedia. Image searching and relevant keyword searches are crude. Article queries are performed according to query syntax, thus delivering lower quality hits.

Legally, publications in Wikipedia are protected through only one license, the creative commons or (CC) license. Publishing anything outside of the bounds of the CC license agreement on Wikipedia is rather complicated, costly and would defy the meaning of a free encyclopedia. However, protection exclusively under the CC license also affects the quality of the content and reduces content diversity and integrity. Content outside of Wikipedia is protected by other authorship licenses and thus cannot be made available through Wikipedia. This is rather unfortunate, because having a diverse set of corresponding or semantically related content under a set of different licenses would enrich the experience of Wikipedia users. For example: a user can explore archival documents and 3D scans of places relevant to the subject, or the user can read about the same topic in a number of different articles written by different authors at different time periods.

3.2 Europeana. Similar to Wikipedia, another e-encyclopedia type application is Europeana. Europeana isn't an encyclopedia in the classical sense. It is rather a collection of information concerning European cultural heritage. Users of Europeana have access to millions of books, pictures, museum pieces, movies and archive data [4]. All the contents are also under the supervision of the Europeana foundation and over 2000 institutions all over Europe are contributing. Every institution individually is responsible for the selection and presentation of its contents, and contribution is exclusively reserved for these institutions [5]. By restricting contributions to select and legitimate institutions, the quality of the content is expected to be of high caliber. Unlike Wikipedia which is funded solely by donations, Europeana and the projects contributing data to Europeana have been funded by the European Commission under eContentplus, the Information and Communications Technologies Policy Support Program (ICT PSP) and similar programs [6]. Under ICT PSP, the European Commission alone invested 6.200.000 Euros into the development of Europeana; the sum covers only 50%-100% of the costs of the total project [7]. Therefore, Europeana is also reliant on Member States' ministries of culture and education for an element of its funding [6]. Concisely, accumulation of diverse content related to "European culture" and the enforcement of protected ownership and authorship rights are the main points of the Europeana project. Content credibility is the advantage of Europeana which distinguishes it from more popular Wikipedia. However, Europeana is not an encyclopedia. It is only a portal to relevant multimedia content. Another disadvantage of Europeana is that it focuses exclusively on countries within the European Union; therefore several countries with rich European cultural heritage such, as Serbia for example, are unfortunately excluded from this initiative.

3.3 Austria Forum. The initial template of Serbia-Forum is another similar e-encyclopedia initiative, "Austria Forum". The Austria Forum project brings almost all of the advantages of a Wikipedia like e-encyclopedia with the diverse multimedia content advantages of Europeana. It is also even more region specific in terms of its content then Europeana, bringing the content to the level of a country and thus to the level of a specific culture. Austrian culture in this case. The Austria Forum has about 20.000 units of well indexed content with an array of different content types. Its content includes books, articles written by credible authors (Wikipedia like), photographs and videos [8]. By developing a semantically configured search/indexing scheme, the Austria Forum is a great tool for shedding light on some of the jewels of Austria's national and cultural heritage, geared towards the German speaking public. A relatively advanced e-book reading web application is available for fast retrieval of scanned books and photo albums [9]. However, unlike the Serbia-Forum initiative, the Austria Forum also has its limitations. Firstly, author credibility is questioned; qualified article authorship which is however compensated by required author's biographies, allows the user/reader of the article to justify himself/herself whether the author's credibility is sufficient. Serbia-Forum adds another level of assurance of authorship credibility. This approach has only one dimension of editorial redaction and false information censorship. Secondly, archive documents, 3D scans of spaces and efficient semantic searching and correlative analysis between digitized documents, articles and other content is to be better addressed in the Serbia-Forum as well.

4. The Serbia-Forum Project

Cultural heritage is the set of tangible and intangible artifacts that support the cultural identity of members of the culture. The set of tangible and intangible artifacts of cultural heritage, referred to as cultural heritage elements, define a peoples. The beliefs, customs, traditions, foods, mythology/bards, literature, history, science, technology, language, music, arts, architecture, sports, militarism, philosophy, politics of leadership and much more can be found under the set of cultural heritage elements. Digitization of cultural heritage elements has many advantages such as improving public access to cultural heritage elements both tangible and intangible and facilitating teaching and research [10]. Central to this advantage is content preservation.

4.1 Motivation. According to Em. Univ. Prof. Dr. Hermann Maurer, "we have to preserve the past and the present of countries and regions, and to continue to do so, so that people today know what was important in the region e.g. 20, 30, 100 years ago, but also that in 2050 people will know about what the region was like in the past" [9]. For Serbia, the Serbia-Forum digitization initiative is welcomed with open arms. Primarily, because of Serbia's historically and culturally turbulent location, where many times through history Serbian cultural heritage elements have been stolen and destroyed. For instance, the NAZI bombing of Belgrade as part of Hitler's operation "Punishment", targeting the Kingdom of Yugoslavia on April, 6th 1941, which resulted in the destruction of the National Library of Serbia (NBS) that led to the complete loss of over 500,000 volumes of Serbian medieval literature. Thousands of maps, scientific documents, and other cultural artifacts as well, were completely destroyed. Many works of art, wisdom and knowledge of the Serbian medieval forefathers had been lost forever [11]. It is one of the Serbia-Forum's goals to help prevent such disastrous losses. Secondly, the costs of the initiative itself are drastically lower than those of similar initiatives in Europe and the world.

Finally, as mentioned before it is important to improve access to cultural heritage information which facilitates teaching and research. Archivists, librarians and museum professionals are among the many groups that are increasingly involved in creating digital resources to improve access and understanding to their collections [10]. Such information should be available to everyone. The Serbia-Forum project is a unique project because it attempts to localize content regionally to the cultural level of the region, to unify a community of credible authors to continually write articles for the forum and to provide high-quality trustworthy cultural heritage content. Sources of content are carefully chosen and are more often than not, government protected institutions such as formidable universities, national and city archives, national and city libraries, museums monasteries and much more. The details of the Serbia-Forum are governed by a strict set of guidelines referred to as "axioms".

4.2 Axioms of the Serbia Forum. The Serbia Forum project has two characters: encyclopedic and preservative. The Serbia-Forum's purpose is to present credible encyclopedic articles and digitally captured elements (cultural heritage elements) of Serbian cultural and national heritage. The two characters mentioned above are governed by two sets of axioms that distinguish Serbia-Forum from all similar initiatives in use today. These axioms are referred to as the primary and secondary axioms. The primary set of axioms is local in nature and bound to the Serbia-Forum project as follows:

1. Content is selected and controlled by government funded and owned cultural, and academic institutions. Such as the National library of Serbia or the National Archives of Serbia.
2. There exists a licensing scheme where each document is copy protected by an adequate license. A legal document stating how each digitized piece of content is to be treated online with respect to its owner and/or author.
3. Quality instead of quantity where the extremes of popular content meet rare content. Excluding the midsection of the content range.

All of the previously mentioned popular e-encyclopedia like web applications are not governed by these axioms. The secondary set of axioms is global in nature and not exclusively bound to the Serbia-Forum:

1. Semantic search and correlation of all content (particularly articles and digitized content and digitized sub-content i.e. pictures in a book)
2. Version tracking of every document/written article. Thus, facilitating "multidimensional" research and learning (explained in the follow-up paper). Tracking how a particular topic was perceived through the course of time.
3. Information about the author of each document is supplied in the form of a small biography. Related to the content selection axiom of the primary set of axioms, this axiom serves as a second credibility evaluation of authorship and information (authors are sifted through and selected from the total pool of authors by a highly credible and unbiased editorial board).

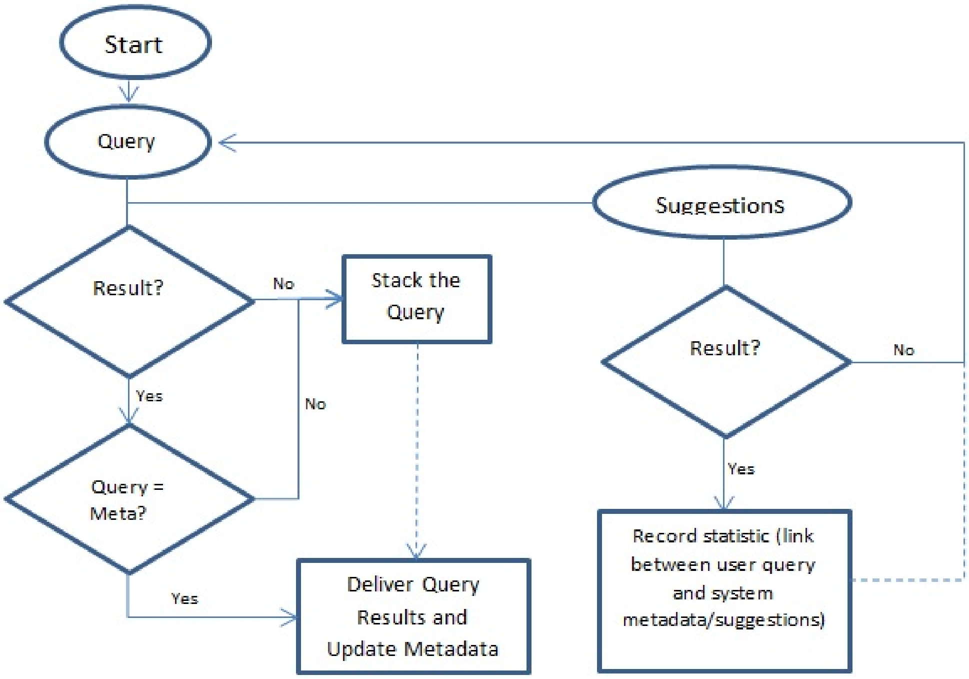
Some existing projects of a similar nature may have already implemented some or all of the axioms within the set of secondary axioms. However, the purpose of the Serbia-Forum project is not solely the implementation of these secondary axioms, but also their improvement and documentation thereof. Such documentation based on Serbia-Forum research, experimentation, and experience will result in several follow-up papers.

4.3 Details of Importance. The current main research idea of concern within the Serbia-Forum project is data structuring, organization and semantic searching of content and sub-content. To illustrate sub-content, let us consider an example of a book which represents content, all of the book contents, including all illustrations, images and text would be considered as sub-contents. There is much room for improvement in the area of efficient semantic searching of content and content correlation. The generic search algorithm might deal with the tuple (input type, output type). In conventional syntactical search engines these two values are generally the same (input type = output type). Google however might obfuscate this principle i.e. image searching. However, when one searches for images in Google, the search isn't performed semantically, rather, syntactically. Google returns the names of the image files whose filenames and metadata (which are in textual form) are syntactical related to the user's input query, along with a small snapshot of the image labeled under that resulting filenames. But in semantics any combination of the two query types is acceptable. In syntactical engines, searches are performed exclusively using the input type. One cannot yet efficiently search for images with an image as a query type or for that matter with a combination of images and text. The meaning of the input, hidden behind its type, can broaden or narrow down a set of query hit results. Since in semantic search schemes the input type is irrelevant, the output type is irrelevant as well. This gives the user a plethora of sources of information to choose from/sift through when performing a search, therefore enriching his or her learning experience.

5. Future Research & Follow-up Papers

Two novel ideas for the improvement and enrichment of the user's Serbia-Forum experience are an interactive semantic search and index definition method using hidden Markov models HMMs and correlation analysis and dynamic story-lining of content [12].

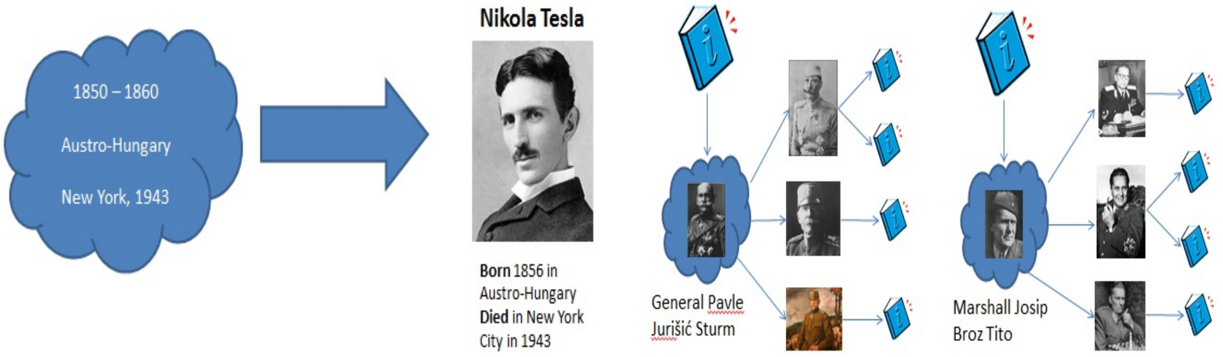
5.1 Interactive Semantic Search and Index Definition. Interactive semantic searching and index definition would be comprised of two modules. The first is a suggestions module that delivers a list of suggested "next query" inputs upon providing the initial query in case that the content results of the initial query is not what the user is looking for. The second module is the index or metadata definition module. Within this module a sequence of unsuccessful but related queries is stored onto a stack and finally appended as metadata to the final and desired content result. Thus enriching the metadata definitions of the queried content and facilitating efficient query hits.


Start
Query
Suggestions
Result?
No
Stack the Query
Result?
No
Yes
No
Query = Meta?
Yes
Yes
Deliver Query Results and Update Metadata
Record statistic (link between user query and system metadata/suggestions)

Picture 1: The crude idea in flow chart form of the interactive semantic search and indexing scheme; two modules are linked directly at the query-suggestions edge. Suggestion results are not the actual query results, but rather suggested next query search parameters delivered by the system based on existing content metadata.

The two modules can be used together in unison, as shown in Picture 1, to enrich the metadata description of content which will aid the user in finding the desired content faster; facilitating query hits. For example, the user enters the first semantic query string upon which a list of possible content results is displayed along with suggested keywords for a possible next query (if the desired content result is not displayed in the list). If the desired content is not in the displayed list then the user is prompted to insert another relevant keyword(s) query, either from the suggested keyword set or any new keyword(s) query. The initial keyword(s) is/are saved in a stack, and the next semantic keyword(s) search is processed. The next set of results are displayed and if the desired content is now present in the list of results then the initial keyword(s) is/are popped off the stack and become a new metadata describing the desired content. The initial idea behind the proposed suggestions module is based on an HMM. The hidden states of the chain are the desired results of the query and the observed states are the suggestions. Further research and results will be found in the follow up paper.

5.2 Correlation Analysis and Dynamic Story-lining. Story-lining is a form of dynamic content presentation under content collaboration scheme. Semantically related content is presented in an organized fashion to form a logical sequence of knowledge points or stories, hence a line of stories. These stories can be temporal in nature i.e. a history book structure covering the events of World War Two, or they can be a union based on nature where a knowledge set requires some prerequisite knowledge set, building up on a sequence of prerequisite and forthcoming knowledge. A story line consists of a sequence of points, where each point is some short and easily consumable information. Each point has the option to branch out into topics the user is potentially interested in [12].



Picture 2: The image on the left is an example of semantic search of an image; (text, image). The image on the right is an example of semantic search of and image accompanied by, a further semantic search of correlated images, and texts in other units of a content/sub-content.

Image recognition and relevance analysis are an important part of semantic correlation, which is an important tool in the story-lining concept. It would help in binding whole texts, segments of text and keywords (sub-content type text), and other images and content/sub-content types of some query of a type image. This widens the correlation spectrum by increasing the number of file types considered in story-lining. Picture 2, is an example of a short semantic sub-content correlation, of content type book "A", sub-content type image "A-1" with other sub-content type image sub-content in other books, say image 23 in book B, "B-23". In this case, the image of famous Serbian World War One General Pavle Jurišić Sturm in one book is semantically correlated to other images of him in other books.

6. Conclusion

Wikipedia is not enough [9]. Even though Wikipedia seems to be an extremely useful source of encyclopedic and digitized content, Wikipedia and similar web applications available online have their limitations. Unlike such existing e-encyclopedias, the Serbia-Forum aims to localize content regionally to the culture of the region, to unify a community of credible authors to continually write articles for the forum and to provide high-quality trustworthy cultural heritage content. Serbia-Forum is a currently progressing digitization project whose goal is to incorporate contributed knowledge of cultural heritage and qualitatively digitized cultural heritage content in an encyclopedic manner as an enhanced Wikipedia-like website. Its aim is to unify a regionally and culturally distinguished community, such as Serbia, and to facilitate digital access to otherwise inaccessible digital resources of the Serbian culture. Such an initiative will benefit the European community by serving as a model for European cultural preservation in Europe’s fast melting multinational and multiracial setting. By building a better future together, projects such as Serbia Forum will enhance the bonds between the peoples of Europe, facilitating intercultural harmony through a profound understanding, sincere appreciation, cordial respect and maximized cooperation.

References

[1] Wikimedia META-WIKI, *Wikipedia.org is more popular than…*, http://meta.wikimedia.org/wiki/Wikipedia.org_is_more_popular_than... (18.06.2012)

[2] Qualman E., *How Wikipedia Makes its Money and Exists*, Socialnomics, http://www.socialnomics.net/2009/12/13/how-wikipedia-makes-money-exists/, 13.12.2009, (18.06.2012)

[3] Wikipedia article traffic statistics, *Main_Page*, http://stats.grok.se/en/201012/Main_Page, (18.06.2012)

[4] Europeana.eu, *About Europeana: What is Europeana?*, http://pro.europeana.eu/web/guest/europeana-faq, (18.06.2012)

[5] Academia Europaea, Europeana, http://pro.europeana.eu/about?utm_source=portalmenu&utm_medium= portal&utm_campaign=Portal% Bmenu, (18.06.2012)

[6] Europeana.eu, *Financial: How is Europeana financed?*, http://pro.europeana.eu/web/guest/europeana-faq, (18.06.2012)

[7] European Commission Information Society, *Funded projects - eContentplus: Call 2008*, http://ec.europa.eu/information_society/activities/econtentplus/projects/econtentplus/index_en.htm, (18.06.2012)

[8] Austria-Forum, *Kurze Einführung in das Austria-Forum*, http://www.austria-lexikon.at/af/Hilfe/Kurze%20Einf%C3%BChrung%20in%20das%20Austria-Forum, (18.06.2012)

[9] Maurer H., *Why Wikipedia is Not Enough*, MISANU Belgrade Presentation, 12.02.2012, (18.06.2012)

[10] National Archives and record administration, http://www.archives.gov/digitization/strategy.html, (18.06.2012)

[11] Savic C., Belgrade 41: *Hitler’s Attack*, http://www.serbianna.com/columns/savich/081.shtml, (18.06.2012)

[12] Milutinovic V., et. al., *Proposal: Attention Capturing Story-based Individualized Information Integration*, FP7 ACS III Project Proposal, (18.06.2012)

ambiz2005@gmail.com