WEB 2018

The Sixth International Conference on Building and Exploring Web Based Environments

ISBN: 978-1-61208-636-1

May 20 - 24, 2018

Nice, France

WEB 2018 Editors

Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany
Kevin Daimi, University of Detroit Mercy, USA

WEB 2018

Foreword

The Sixth International Conference on Building and Exploring Web Based Environments (WEB 2018), held between May 20 - 24, 2018 - Nice, France, continued the inaugural conference on web-related theoretical and practical aspects, focusing on identifying challenges for building web-based useful services and applications, and for effectively extracting and integrating knowledge from the Web, enterprise data, and social media.

The Web has changed the way we share knowledge, the way we design distributed services and applications, the way we access large volumes of data, and the way we position ourselves with our peers.

Successful exploitation of web-based concepts by web communities lies on the integration of traditional data management techniques and semantic information into web-based frameworks and systems.

We take here the opportunity to warmly thank all the members of the WEB 2018 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to WEB 2018. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the WEB 2018 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that WEB 2018 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of Web-based environments.

We are convinced that the participants found the event useful and communications very open. We also hope that Nice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

WEB 2018 Chairs:

WEB Steering Committee
Daniela Marghitu, Auburn University, USA
Erich Schweighofer, University of Vienna - Centre for Computers and Law, Austria
Toyohide Watanabe, Nagoya Industrial Science Research Institute, Japan
Demetrios Sampson, Curtin University, Australia
Mariusz Trzaska, Polish-Japanese Academy of Information Technology, Poland
Imon Banerjee, Stanford University School of Medicine, USA
Alexiei Dingli, University of Malta, Malta
Hossein Sarrafzadeh, Unitec Institute of Technology, New Zealand
Michel Jourlin, Jean Monnet University, Saint-Etienne, France

WEB Industry/Research Advisory Committee
Taketoshi Ushiama, Kyushu University, Japan
Krzysztof Walczak, Poznan University of Economics, Poland

WEB 2018

Committee

WEB Steering Committee
Daniela Marghitu, Auburn University, USA
Erich Schweighofer, University of Vienna - Centre for Computers and Law, Austria
Toyohide Watanabe, Nagoya Industrial Science Research Institute, Japan
Demetrios Sampson, Curtin University, Australia
Mariusz Trzaska, Polish-Japanese Academy of Information Technology, Poland
Imon Banerjee, Stanford University School of Medicine, USA
Alexiei Dingli, University of Malta, Malta
Hossein Sarrafzadeh, Unitec Institute of Technology, New Zealand
Michel Jourlin, Jean Monnet University, Saint-Etienne, France

WEB Industry/Research Advisory Committee
Taketoshi Ushiama, Kyushu University, Japan
Krzysztof Walczak, Poznan University of Economics, Poland

WEB 2018 Technical Program Committee

Rocio Abascal Mena, Universidad Autónoma Metropolitana - Cuajimalpa, Mexico
Alessia Amelio, University of Calabria, Italy
Leandro Antonelli, UNLP, Argentina
Irina Astrova, Tallinn University of Technology, Estonia
Sofia Athenikos, Bank of America Merrill Lynch, USA
Dirk Bade, University of Hamburg, Germany
Imon Banerjee, Stanford University School of Medicine, USA
Carlos Bobed Lisbona, University of Zaragoza, Spain
Tharrenos Bratitsis, University of Western Macedonia, Greece
Marcin Butlewski, Poznan University of Technology, Poland
Tania Calle-Jimenez, Escuela Politécnica Nacional, Ecuador
Rodrigo Capobianco Guido, Paulo State University (Unesp), Brazil
Naděžda Chalupová, Mendel University in Brno, Czech Republic
Dickson Chiu, The University of Hong Kong, Hong Kong
Bouras Christos, University of Patras, Greece
Alexiei Dingli, University of Malta, Malta
Mauro Dragoni, Fondazione Bruno Kessler, Trento, Italy
Mahmoud El-Haj, SCC | Lancaster University, UK
Vadim Ermolayev, Zaporozhye National University, Ukraine
Cécile Favre, University of Lyon | ERIC Lab - Lyon 2, France
Giacomo Fiumara, Università degli Studi di Messina, Italy
Jakub Flotyński, Poznań University of Economics and Business, Poland
Raffaella Folgieri, Università degli Studi di Milano, Italy
Marco Furini, University of Modena and Reggio Emilia, Italy
Christos K. Georgiadis, University of Macedonia, Greece

Erich Schweighofer, University of Vienna - Centre for Computers and Law, Austria
Blerina Spahiu, University of Milano - Bicocca, Italy
George Tambouratzis, Institute for Language and Speech Processing - Athena R.C., Greece
Mariusz Trzaska, Polish-Japanese Academy of Information Technology, Poland
Domenico Ursino, University "Mediterranea" of Reggio Calabria, Italy
Taketoshi Ushiama, Kyushu University, Japan
Sofie Van Hoecke, IDLab - Data Science Lab | Ghent University-imec, Belgium
Costas Vassilakis, University of the Peloponnese, Greece
Maurizio Vincini, Universita' di Modena e Reggio Emilia, Italy
Krzysztof Walczak, Poznan University of Economics, Poland
Toyohide Watanabe, Nagoya Industrial Science Research Institute, Japan
Jasy Liew Suet Yan, Universiti Sains Malaysia, Malaysia
Jian Yu, Auckland University of Technology, New Zealand
Fouad Zablith, American University of Beirut, Lebanon
Yongfeng Zhang, University of Massachusetts Amherst, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Radical Transparency on the Web

Terrance Goan
Stottler Henke Associates, Inc.
Seattle, USA
email: goan@stottlerhenke.com

***Abstract*—Web users face massive and ongoing challenges in ascertaining the legitimacy and originality of information they discover. Intellectual property is commonly misappropriated; false news is propagated virally; and the original source of information is typically very difficult to determine. Recent technological advances have opened the door to supporting real-time transparency for all web content. In this paper, we outline a way forward, and open a discussion of the potential impact of radical transparency on the Web.**

Keywords- plagiarism detection; web search; intellectual property; source discovery; fact checking.

I. INTRODUCTION

As much as the Web is a resource for valuable and legitimate information and services, it has also become increasingly riddled with copyright violations, urban legends, rumors, fraud, and misinformation. Further, the vast scale and scope of the Web makes it ungovernable by any centralized authority. The only means of combatting this challenge is by empowering Web users by revealing evidence of credibility and sourcing. There are, of course, individuals and organizations that seek to fact-check hoaxes and scams, but the processes of source-discovery and fact-checking are laborious, and the products of these investigations typically reach only a small percentage of those who could benefit.

What is needed now is a complement to traditional Web indices—one that makes it easier for users to follow information “bread crumbs” back to original source materials so that they can assess for themselves validity and originality. The technical solution may appear to the end-user as a plagiarism detection system that can highlight all passages on a Web page that share a common origin with one or more other Web pages. Such a solution would not only reveal potential illegal reuse and fraud, but also could serve as a foundation for a new form of Web navigation that allows users to navigate amongst closely related materials that have not been explicitly hyperlinked (e.g., news stories that share significant quotations). This radical form of transparency would fundamentally change the way content is created and consumed on the Web.

In Section 2 we present related research, and in Section 3 we describe a key technology that could pave the way forward. In Section 4 we then describe some initial steps that we have taken. In Sections 5 and 6 we then discuss the potential impacts of radical transparency on the Web and our conclusions.

II. RELATED WORK

The proposed approach to Web transparency is most closely related to research conducted in the areas of plagiarism detection and author identification. Over the past decade, numerous useful tools have been developed to detect plagiarism [3]. State-of-the-art plagiarism detection systems rely on matching similarities between documents, and thus the effectiveness of these tools is limited by the scope and characteristics of the text collections they index.

While details vary, contributions to the annual Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) competition [4] follow a two-stage strategy that was originally proposed by Stein et al. [6]. First, the contents of a suspicious document are analyzed to generate queries that may return potential source documents. Each of these candidates is then compared with the suspicious document to identify matching passages.

What is important to note is that while these methods may sufficiently scale to the challenge of detecting plagiarism in college essays, they are wholly inadequate for application at Web-scale [5][7]. The primary challenges in applying existing methods are the costs of finding and retrieving potential matches and in doing pairwise comparisons. However, new approaches to text indexing and comparison are available that could be implemented at Web-scale. In the next section we outline one of these possible approaches along with evidence of its sufficiency.

III. A PATH TO WEB TRANSPARENCY

In 2011, Mansour et al. published a paper describing a parallelizable string indexing scheme that can index the human genome in just 19 minutes on an ordinary desktop computer [2]. Since then, we have witnessed continued progress in the rapid indexing, search, and analysis of huge text sequences (e.g., string similarity calculations) (e.g., [8]). While this research originated with a focus on bioinformatics, the potential applications for these new algorithms are far-reaching and include the potential for full-text comparisons on the Web.

The conceptual foundation for Mansour’s work is the suffix tree [1]. A suffix tree is a data structure that indexes all possible suffixes of a string (e.g., a Web page). Both the construction, and querying, of a suffix tree can be done in linear time (in the length of the input). Further, the suffix tree allows us to search a very large corpus (e.g., the Web) for *all* occurrences of a String S in $O(|S|)$. This ability to conduct full text comparisons against an arbitrarily large corpus of text in time that grows linearly with the size of the query string opens the door for the Web transparency we propose. Finally, Mansour's approach allows for parallel disk-based processing on commodity hardware.

IV. Initial Steps

We recently constructed a limited form of this concept. We implemented Mansour's algorithm and applied it to the challenge of detecting the intentional or unintentional release of sensitive intellectual property from an organization. Essentially, we built an index of proprietary text and then conducted searches against that index with any suspect material. We found that using four commodity desktop machines, our initial implementation could index 1TB of data in four days. Even with this non-optimized implementation, we believe indexing petabytes of Web data in this manner is now feasible.

Searching for matches is more challenging than indexing since identifying *all* overlaps between a query document and a stored corpus typically would require many restarts. The process involves finding a match starting at position X in the query document, then reinitiating the search at position X+1. There are, however, numerous optimizations that are possible, including ignoring short matches and extremely common matches which are likely of little import. In our initial tests, searching 1,000 documents on a single machine took under two seconds.

This approach of course has limitations. Plagiarism detection, for instance, can be complicated by careful obfuscation that would limit the number of longer matches. However, the scale of the index would act to counter this force as numerous variants of the original would also be indexed. Further, many of the useful applications of these indexes would not involve intentional obfuscation.

V. Impacts of Web Transparency

It is worth considering the impact of easily accessible transparency on the Web. While, on balance, the effect of transparency would be positive, there remains a possibility of unfortunate side-effects. Consider the following:

- Intellectual property protection. Given a Web-scale index, unauthorized content reuse would be easily detected. This would likely have the biggest impact on the research community and in journalism, where the material is made public and plagiarism can have dire professional ramifications. Interestingly, the transparency we envision may spark improvements in automated plagiarism approaches where words are strategically replaced with synonyms to avoid detection.
- Implicit reference chaining. Many articles, Web pages, and academic papers legitimately include quotations and share bibliographic references. By revealing these text overlaps, Web browsers could enable users to navigate between related documents without the need for explicit hyperlinks.
- Fact checking and urban legend detection. The proposed approach to Web transparency could enable automated source identification through a combination of text overlaps and timestamps. This would allow information consumers to quickly ascertain for themselves the likely veracity of published reports.
- Change detection. Many legal documents are created from templates or through the reuse of material found on the Web. Examples include usage agreements, licenses, non-disclosure agreements, and the fine print associated with financial instruments, such as credit cards. The proposed web indexing scheme would present the opportunity to highlight differences between a legal document and others that were found online. This in turn would allow users to detect modifications of terms that warrant their attention.

VI. Conclusions

In this paper, we have proposed the utilization of highly scalable string indexing and search methods to provide an extreme form of transparency on the Web. The implications of revealing shared text during everyday use of the Web are worthy of consideration, particularly given the lack of any governing authority that can act to thwart the growing threat posed by plagiarism, fake news, and government-sponsored propaganda initiatives on the Web. As a next step we anticipate a subject-specific Web-index that will allow us to further explore issues of scalability and utility.

References

[1] D. Gusfield, Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge university press, 1997.

[2] E. Mansour, A. Allam, S. Skiadopoulos, and P. Kalnis, "ERA: efficient serial and parallel suffix tree construction for very long strings," Proc. of the VLDB Endowment, vol. 5, no. 1, pp. 49-60, 2011.

[3] V. Martins, D. Fonte, P. Henriques, and D. da Cruz, "Plagiarism detection: A tool survey and comparison," OASIcs-OpenAccess Series in Informatics, vol. 38, pp. 43-58, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[4] Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) Workshops http://pan.webis.de/, [retrieved March, 2018]

[5] M. Sanchez-Perez, G. Sidorov, and A. Gelbukh, "A winning approach to text alignment for text reuse detection at PAN 2014," In CLEF (Working Notes), Sep. 2014, pp. 1004-1011.

[6] B. Stein, S. zu Eissen, and M. Potthast, "Strategies for retrieving plagiarized documents," Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Jul. 2007, pp. 825-826.

[7] N. Unger, S. Thandra, and I. Goldberg, "Elxa: scalable privacy-preserving plagiarism detection," Proc. of the 2016 ACM Workshop on Privacy in the Electronic Society, Oct. 2016, pp. 153-164.

[8] M. Yu, J. Wang, G. Li, Y. Zhang, D. Deng, and J. Feng, "A unified framework for string similarity search with edit-distance constraint," The VLDB Journal, vol 26, no. 2, Apr. 2017, pp. 249-274.

Cultural Based Adaptive Web Design For Wellington Institute of Technology

Lakshmi Sivadas
School of IT
Wellington Institute of Technology
Wellington, New Zealand
Email:Lakshmi.Sivadas1988@gmail.com

Abdolreza Hajmoosaei
School of IT
Wellington Institute of Technology
Wellington, New Zealand
Email:Reza.moosa@weltec.ac.nz

Abstract— **Currently academic institutions' websites are used for evaluating the standard of institutions. The institution's website is an important marketing tool because it is an advertising forum to students. Website is a kind of promotional material which exchanges academic images to students which will in turn provides revenue to institution. Poorly designed website is risky because it will create a negative impact about academic institution. The role of academic website is very important for student's decision to select an academic institution because every student will examine website before enrolling. There are many features that attract users to a website; one among those factors is cultural attributes. Web designers should adapt a cross cultural web design by considering the culture of the targeted audience. A well-designed website with improved user interface which would incorporate various cultural factors will definitely increase the revenue of institution. The aim of this research is to evaluate the web interface of Wellington Institute of Technology institute (Weltec) based on the various cultural dimensions of Asian students and to provide some suggestions to design an interface layout for Weltec web interface, which can satisfy Asian students' cultural attributes**.

Keywords - Web Interface Design; Academic Institution Website; Cultural Attributes; Website Satisfaction.

I. INTRODUCTION

Nowaday's lots of researches have been conducted in the field of anthropology on the cultural differences and similarities. The primary reason for increasing the number of research in this area is because the web has become a medium for promotion and marketing. Therefore, it is essential that a web interface design reflects the cultural preferences of target audience [1]. According to Hofstede, the culture is the collective program of mind that differentiates the people from a group from another group where the mind stands for thinking, feeling and acting towards beliefs, attitudes and skills [2]. Culture can also be defined as patterns of thinking that influence people how to communicate with user interface [3].

Nowaday's academic institutions' websites are used for evaluating the standard of institution [4]. The most important marketing tool for an institution is to have an effective website because website is an advertising forum to students. The website is a kind of promotional material which exchange academic idea to students, which in turn provides revenue to institution. A poorly designed website is risky because it will create a negative impact about academic institution. When students visit academic institution website they must find it easy to navigate and access information otherwise they will leave the website and institute will lose potential candidates. According to Davis and Lindridge as cited by [5] , cultural factors must be considered in web design which can increase aesthetic quality and success of website. A well-designed website with improved user interface incorporated with various cultural factors will definitely increase the revenue [6]. Therefore, while designing a web interface, relevant visitors' culture attributes should also be considered.

Wellington Institute of Technology (Weltec) is an internationally recognized Centre of education, providing education for more than 11,000 students every year and offering more than 150 programs. In the competitive world, website should be designed efficiently to satisfy and to attract more students to Weltec. The Weltec web interface design should adapt a cross cultural web design by considering the culture of the targeted audience. The aim of this research is to evaluate web interface of Weltec based on the cultural dimensions of Asian students and to provide some suggestions to design an interface layout for Weltec web interface which can satisfy cultural dimensions of Asian students.

In Section II, we describe the influence of cultures on web design. Section III describes the Hoftstede's cultural dimensions. Related works are explained in Section IV. Research approach and findings are described in Section V and finally, Section VI covers the conclusion of research.

II. THE INFLUENCES OF CULTURES ON WEB DESIGN

In 2001, Sun [7] studied about the effect of culture on web site and stated that users use cultural priorities to evaluate the quality of a website. Studies conducted by Smith, Dunkly, Minoch [8] stated that use of design components according to user's culture will definitely increases the user satisfaction, usability and friendliness of a website. Marcus and Gould stated in [9] that it is possible to analyze the culture effects in terms of web design components like images, icons and navigation. Galdo, Fernandes, Russo and Boor in [10] emphasized that designer should include the cultural factors such as icons, symbols and colors in web page design. The culture is very important aspect while designing a web interface because it can help users to interact much better with the interface. Therefore, it is necessary to incorporate cultural characteristics in interface design [11]. Good understanding of cultural preference is necessary before designing web interface. Culture is a major factor that has to be considered to create a global interface design because users will feel more comfortable while interacting with the interface that is designed according to their culture [12]. Global interface should include a diversity of culture to provide support for users [13]. Users will be from different background so that their expectations towards the interface will be also different so it is very important to satisfy cultural features to increase usability of web interface [14].

III. HOFSTEDE'S CULTURAL DIMENSIONS

One of the goals of this study is to identify the cultural dimensions that must be considered in website design. For this purpose, we have done literature review on several cultural models, such as Victor's Model [15], Hall's Model [16], Trompenaar's Model [17] and Hofstede's Model [18] . In this study, we selected Hofstede's cultural model because Hofstede's theory describes clearly the attributes of culture. Hofstede's theory is very popular in field of cultural research and has been cited more than 3500 times and included in more than 9000 articles [1]. Hofstede is the ninth most cited european in 2011 according to social science citation index [19]. Ford and Kotze argue in their article [1] "Designing usable interfaces with cultural dimensions" that the web interface that follows the Hofstede's cultural dimensions will provide more user friendlyness in web layout.

Greet Hofstede worked as a psychologyst for IBM and conducted a study during the period from 1978 to 1983. He collected data from more than 100.000 IBM employees from 53 countries through interview and survey. Stastical analysis of large set of data was done and each country was given a rating from 0 to 100 [9] . In 1990, Hofstede published 'Software of mind' with more details of culture in an organization with his five cultural dimensions Power Distance, Individualism Vs Collectivism , Masculinity Vs Femininity , Uncertainty avoidance and Long Term orientation Vs Short term orientation. Power distance refers to inequalities among the people in society and how power is distributed [20]. Individualism refers to individuals are expected to take care of themselves and collectivism expect family member or relatives can look after a person [21]. Masculinity refers to difference in the emotional roles between genders [22] . Uncertainty Avoidance deals with feel of anxiety for a member of society in a particular situation [23]. Long term orientation stands for encouragement towards the future growth whereas short term orientation stands for present and past [24].

IV. RELATED WORK

Work1- Cultural Similarities and Differences in the design of University web sites [25]*:* This research was done by Ewa Callahan to understand the cultural differences and similarities in the web interface design of universities based on the Hofstede's Cultural dimensions. To conduct the study 900 universities website from 45 different countries like Malaysia, Austria, United States, Japan, Sweden, Greece and Denmark were chosen. Graphical elements and information organization of home page were mainly analyzed by content analysis methods. Correlation between each cultural dimension was calculated. The result of analysis with respect to power distance positively, individuals/collectivism negatively masculinity/femininity positively and uncertainty avoidance positively were correlated. The result of analysis shows web page layout choices are different all around the world but still there were a few similarities like simple menu that were chosen by most of the countries.

Work 2- Arabic Interface analysis based on cultural markers [26]. The primary goal of this research was to analyze the most important cultural markers in the educational websites of Arabic countries. Nine universities including Zayden University, UAE University, Sharjah University, the global university of science, Kuwait institute of Medical specialization from Saudi Arabia, Kuwait, Dubai, Abu Dhabi and Sharjah were selected for analysis. After identifying the cultural markers, the markers were related to Hofstede's cultural dimensions. Hofstede's power distance value for Arabic nation is high. The findings of this study for power distance is also high. The individualism and collectivism is less according to Hofstede's dimension. The result supports Hofstede's claim. Hofstede said Arab countries have a more score in masculinity. This study supports masculinity therefore value of Hofstede for Arabic country can be partially correlated to findings. Hofstede claims that Arab countries have high uncertainty avoidance and there is no score for the long term versus short term orientation. Moderate support for long term orientation was found in this study.

Work 3- Cultural Variability in Web Content [27]*:* A comparative analysis of American and Turkish Websites has been done in this work. This research was conducted in Bebek Istanbul University in Turkey in January 2010. US based company websites and their Turkish counterparts were selected for study. Web content analysis was used to verify Hofstede's cultural dimension in 88 websites. Firstly, after literature review a list of website features were identified which represents different cultural traits. Then focus group study with six students where conducted to find out the website features. The web features identified by students in US and Turkish websites were evaluated. Finally, fourteen web features that represent Hofstede's cultural dimensions were discovered. The websites were analyzed with the absence and presence of the website features. There was a difference in ten features among the fourteen features between Turkish and Us Websites. There was no Power distance in Turkish website. Turkish website displayed more collectivism than US website. Turkish websites show lower Masculine. US websites displayed more Uncertainty avoidance than Turkish websites.

Work 4- Cultural Values and Interpersonally in Spanish and British University Websites [28]*:* This research was conducted in Spain by Francisco Miguel Ivorra. The primary goal of research was to study the impact of Individualism in Peninsular Spanish and British University websites. According to Hofstede findings, Spain is moderately individualistic and UK is highly individualistic. In this study 30 university websites from Spain and UK were selected randomly. The section "Reason to study at the university" was selected to analyze. The Observation and quantitative analysis were used and statistical analysis with Chi square test was done to find result. The findings of the research show Spain has highly tribal culture with the presence of Individualism. Similarly, UK also has tribal culture in a moderate rate.

Work 5- Website Design and localization [29]*:* A Comparison of Malaysia and Britain has been done in this research. Two researchers Tanveer Ahmed and Haralambos Mouratidis conducted a study to explore the cultural values in Malaysian and British websites and how these values are reflected in web interface design based on Hofstede's two cultural dimensions Power distance and Individualism/Collectivism. For this purpose, the author selected six different websites from three areas like Banking, Tourism and Education. Research method for analyzing the cultural elements, was content analysis framework and analysis procedure for analyzing the cultural elements. After analysis the authors found a considerable difference in the cultural values in Malaysian and British websites. The findings of research show Malaysia have high index of power distance whereas British value for power distance is low. The Malaysian's have high collectivism in contrast Britain is highly individualistic.

V. RESEARCH APPROACH

Based on Hofstede's cultural model, we designed a questionnaire and distributed among 80 Asian students at Weltec that come from India, China, Philippine, Sri Lanka, Nepal and Middle East countries. Subsequently, based on the survey results, Weltec website was evaluated to examine whether the website interfaces meet all the cultural criteria of Asian students.

A. Data Collection

In this study, a quantitative survey was conducted. Among the 19 questions of questionnaire, 5 questions were related to power distance, 3 questions were related to uncertainty avoidance, 4 questions were related to masculinity, another 4 were related to collectivism and last 3 questions were about short-term orientation. For each question, respondents may select five rating fields strongly agree, agree, neither agree nor disagree, disagree and strongly disagree respectively shown by number 5 to 1.

B. Data Analysis

The opinion of respondents was firstly separated based on nationality and entered in separate spreadsheets. Statistical data analysis was performed using Microsoft Excel. The response for each question was entered as numeric value; strongly disagree=1 to strongly agree=5. For each cultural dimension average was calculated.
The table below shows the result of data analysis (PD= Power Distance, MAS = Masculinity, COL=Collectivism, ST = Short- term orientation, UA = Uncertainty avoidance).

TABLE I. RESULT OF CULTURAL DIMENSIONS VALUE

	PD	MAS	COL	ST	UA
INDIA	2.52	2.35	3.94	3.92	4.62
CHINA	2.61	2.58	3.67	3.55	4.33
PHILIPPINES	2.16	2	3.77	3.33	4.90
SRI LANKA	2.60	2.70	3.72	3.43	4.40
NEPAL	1.9	1.72	3.47	3.43	4.36
MIDDLE EAST	2	2.42	3.90	3.10	4.53

The survey results for all six asian countries (India,China,Philiphines,Sri Lanka,Nepal and Middle East) were similar for five cultural dimension as shown in Figure1. The cultural dimension values for asian countries in Uncertainity, Collectivism, Short term orientatin are high, and low for power distance and masculinity.

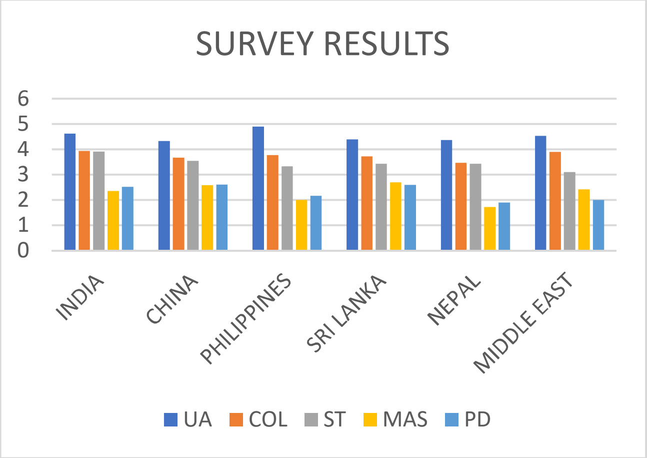


Figure 1. Cultural dimensions value for 6 asian countries

C. Research Findings

The major aim of our research is to investigate whether cultural dimensions of Asian students are reflected in Weltec website. In this research, we examined design characteristics of Weltec website using a set of guidelines developed by Marcus for mapping of Hofstede’s cultural dimensions into web design components.

Power Distance: Survey data shows Asian countries have low power distance. Therefore, the Weltec website should have the five design features listed in the table below.

TABLE II. CULTURAL WEB COMPONENTS FOR LOW POWER DISTANCE

	Components of cultural factors	Identified in Weltec Website
1	Less Structured access to Information	Yes (Pass)
2	Less Focus on Authority	Yes (Pass)
3	Shallow Hierarchies	Yes (Pass)
4	Photos of Students	Yes (Pass)
5	Images of public space and everyday activities	No (Fail)

Among the five web characteristics, only four web characteristics were satisfying, the fifth feature images of public space and everyday activities is absent in Weltec website (please refer to TABLE 2). As example, there is no public image in homepage of Weltec (shown in Figure 2).

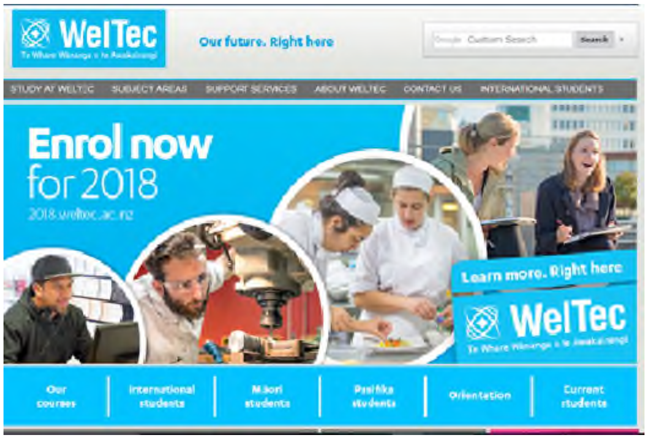


Figure 2. No public images in home page.

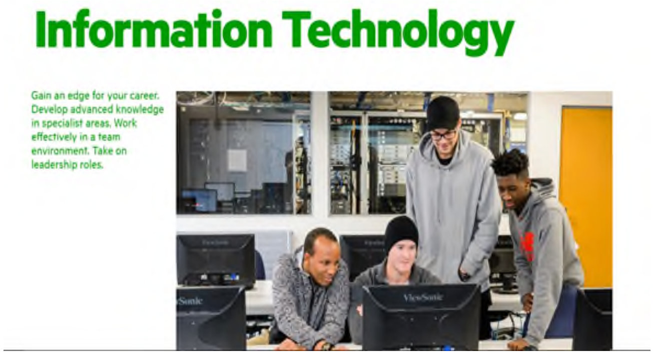
Masculinity: The study shown low score for Masculinity which means Asian students of Weltec prefer femininity in website design. TABLE III shows the failure of Weltec website.

TABLE III. CULTURAL WEB COMPONENTS FOR LOW MASCULINITY

Components of cultural factor	Identified in Weltec website
1. Vivid color scheme	No (Fail)
2. Presence of female themed images	No (Fail)

For example, when students open the information technology page they wish to see different color schemes with female oriented themes in Weltec website instead of pictures of only male students. The Figure 3 shows there is no female image in homepage of School of IT at Weltec. Another design issue, in the picture below we can see only one font color (Green) is used instead of multiple colors.

Collectivism: From the survey result, collectivism has high score so the Weltec website should reflect images of group achievement and group learning activities. Table IV shows the failures.



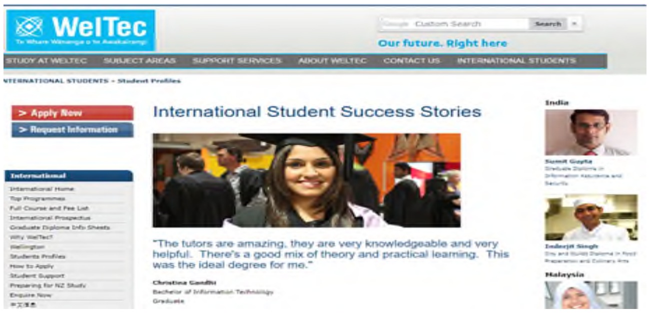
Information Technology

Gain an edge for your career. Develop advanced knowledge in specialist areas. Work effectively in a team environment. Take on leadership roles.

Figure 3. No female Theme in weltec website.

TABLE IV. CULTURAL WEB COMPONENTS FOR HIGH COLLECTIVISM

Components of cultural factor	Identified in Weltec website
Images of group learning activities	No (Fail)
Images of group achievements	No (Fail)
Images of institutional success	Yes (Pass)



International Student Success Stories

"The tutors are amazing, they are very knowledgeable and very helpful. There's a good mix of theory and practical learning. This was the ideal degree for me."

Christina Gandhi
Bachelor of Information Technology
Graduate

Figure 4. No group success stories.

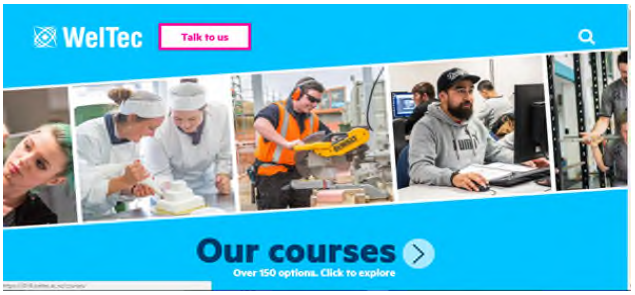


Figure 5. No group learning activities.

The Weltec website displayed individual success stories and pictures instead of group success stories and group achievement images (shown in Figures 4 & 5). When we observed Weltec website we were not able to find images with a group of students sitting together chatting or learning. In the courses web page, among the four photos only one is displaying image of two students and rest of all photos are presenting a single person (shown in figure 5).

Uncertainty Avoidance: From survey results, it is very clear that Asian students of Weltec prefer high uncertainty avoidance which means students expect assistance facilities. TABLE V describes the website characters to meet high uncertainty.

TABLE V. CULTURAL WEB COMPONENTS FOR HIGH UNCERTAINITY

Components of cultural Factor	Identified in Weltec website
Simple, Clear, prominent and limited choices	Yes (Pass)
Navigation stated with strict rules	No (Fail)
Presence of site map	No (Fail)
Messages, contents & visuals with direct meaning	No (Fail)
Presence of Search engine and Institutional calendar	Yes (Pass)

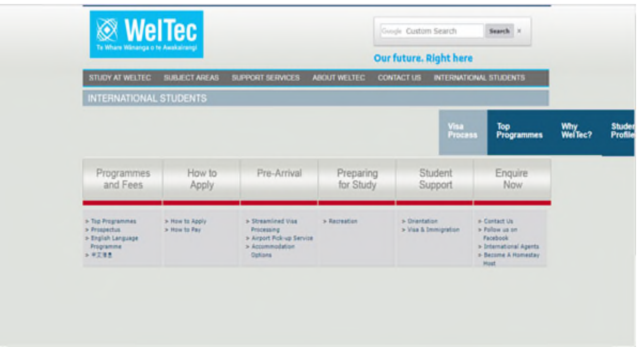
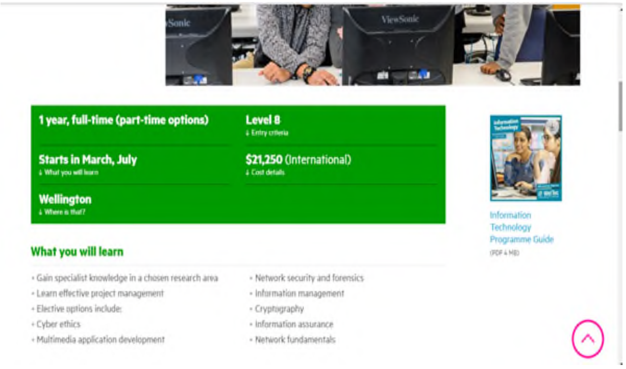


Figure 6. Navigation stated without strict rules.



What you will learn

- Gain specialist knowledge in a chosen research area
- Learn effective project management
- Elective options include:
- Cyber ethics
- Multimedia application development
- Network security and forensics
- Information management
- Cryptography
- Information assurance
- Network fundamentals

Figure 7. No clear description about the course.

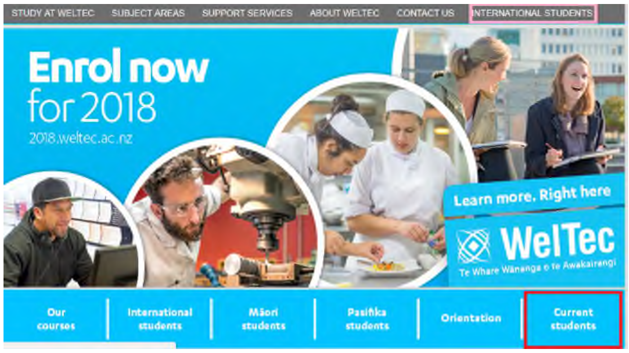


Figure 8. No direct Meaning.

When we investigated Weltec website for high UA features, we couldn't find the strict rules for navigation. Refer to Figure 6 when user clicks on the international tab, it navigates to the screen displayed in the Figure 7 and then when user clicks the top program tab it again takes user back to home page (Figure 8). In this navigation, user will be expecting to view list of programs offered by Weltec but there are no strict navigation rules. The Figure 7 displays the details about post graduate diploma in Information Technology. That page provides details regarding course fee and intake but course subject details and prerequisite of course are missing which is very important for a student to choose a program. In Home page there is a tab named "Current Student" but when clicked on the tab thinking it will be displaying details about current students but it takes user to a page where rules and regulations for newly enrolled students are listed. Moreover, there is no site map in Weltec website to provide support in uncertain situation.

Short-term Orientation: The survey response for short term orientation score is high. Table VI shows the failure of Weltec website for components of this attribute.

TABLE VI. CULTURAL WEB COMPONENTS FOR HIGH SHORT TERM ORIENTATION

Components of cultural factors	Identified in weltec website
Daily routine indicator of weltec	No (Fail)
Allow students to complete task quickly	No (Fail)
Daily activities or current events of weltec	No (Fail)
Presence of short term goals of weltec	No (Fail)

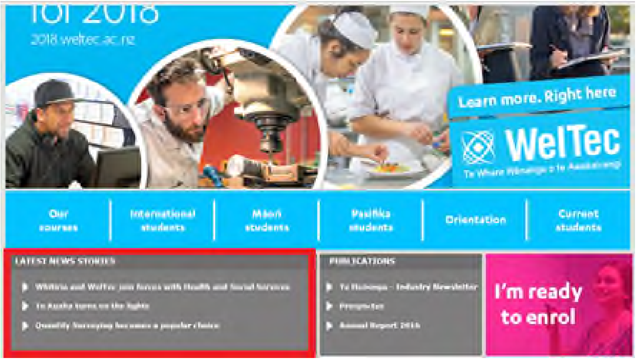


Figure 9. No daily activities.

Daily routine indicator of Weltec was not displayed in home page. Daily activities or current events of Weltec were not listed. Short-term goal of Weltec was not given. The most important one, enrolling into a course is a long process taking the students into many long pages that need scrolling and its time consuming. In Figure 9, we can see the latest stories but not the daily activities or current events. The short –term goals of Weltec is not displayed in home page of Weltec.

VI. CONCLUSION

Based on the discussion of Hofstede's cultural dimensions, this article explored how academic institutions might make their websites design more usable and attractive fpr Asian students. The important question which is brought up in this research is 'What is the approach to design a website which would appeal to diverse cultural backgrounds?'. This paper represented an approach aiming to enhance academic institution website interface attractiveness via an attending majority of Asian visitors' culture. We proposed an approach for website design with focusing on cultural factors of website visitors. Based on our survey, it was proved that the Weltec website is not fully Asian culture oriented and does not fully reflecting the cultural requirements of Asian students. Weltec website designed according to Cultural factors will increase usability and attract more students to Weltec. Weltec should consider cultural factors of target audience to achieve more desirable outcomes.

REFERENCES

[1]G. Ford and . P. Kotzé, "Designing Usable Interfaces with Cultural Dimensions," South Africa.

[2]G. Hofstede, Cultures and Organizations: Software of the Mind, New York: McGraw-Hill, 1997.

[3]E. G. Blanchard and D. Allard, "Modeling a User's Culture," in IGI global, switzerland, 2010.

[4] (. (. 1.-1. ©. P. C. w. T. M. M. International Journal of Basic and Applied Sciences, Walayat Hussain and A. Ahmed, "The Importance of Higher Education Websites and its Usability," [Online]. Available: https://www.researchgate.net/publication/274470588_The_importance_of_Higher_Education_Website_and_its_Usability.

[5]S. D. B. Eristi, "CULTURAL FACTORS in WEB DESIGN," Journal of Theoretical and Applied Information Technology , 2005-2009.

[6]"13 Advantages of Having a Website For Your Business," Dogulin digital, [Online]. Available: https://www.dogulindigital.com.au/advantages-benefits-website-for-business/.

[7]R. Rimondi, "Intercultural aspects of Web Design:," PsychNology Journal, vol. 13, no. 1, pp. 101-120, 2015.

[8]D. L. F. T. M. Smith A., "A Process Model for Developing Usable Cross-Cultural Websites," in Interacting with Computers, 2004, pp. 63-91.

[9]A. Marcus, "Cultural Dimensions and Global Web User-Interface Design: What? So What? Now What?," January 2000. [Online]. Available: https://www.researchgate.net/publication/249779007_Cultural_Dimensions_and_Global_Web_User-Interface_Design_What_So_What_Now_What.

[10]P.and. B. S. Russo, " How fluent is your interface? Designing for international," in Proceedings of 4th Conference, Boston: Addison-Wesley., 1993.

[11]W and. B. A. Barber, "Culturability: The Merging of Culture and Usability.," in Proceedings of the 4th Conference on Human Factors and Usability, 2005.

[12]B. P. C. Shneiderman, "Designing the user interface:Strategies for effective human-computer," Reading, Mass.: Addison, Wesley, 2010.

[13]A. G. E. Marcus, " Cultural dimensions andglobal web user-interface design: What? so what? now what?," in Proceedings of the 6th Conference on Human factors and web, Texas, 2000.

[14]A. H. S. E. M. J. Hurst, " Dynamic detection of novice vs. skilled use without a task model," the Human factors in Computing Systems, 2007.

[15]D. Victor, International Business Communications, Harper

Collins , 1992.

[16]E. Hall, The Silent Language, Doubleday, 1959.

[17]F. Trompenaars, Riding the Waves of Culture, London: Nicholas Brealey Publishing, 1993.

[18]G. Hofstede, Culture’s consequences (2nd ed.), Sage publishers, 2001.

[19]J. W. Bing, "Hofstede's consequences: The impact of his work on consulting and business practices," ITAP International, 2004. [Online]. Available: http://www.itapintl.com/about-us/articles/hofstedes-consequences.

[20]A. Emilie W. Gould, "Cultural Dimensions and Global Web Design," [Online].

[21]G. Hofstede, "The 6 dimensions of national culture," hofstede's insight, [Online]. Available: https://www.hofstede-insights.com/models/national-culture/.

[22]G. J. M. M. Hofstede, Cultures and Organizations: Software of the Mind, USA: McGraw-Hill, 2010.

[23]C. Smit, "What is uncertainity?," Culture matters, 21 October 2016. [Online]. Available: https://culturematters.com/what-is-uncertainty-avoidance/.

[24]G. Hofstede, Dimensionalizing cultures: The Hofstede model in context., 2011.

[25]E. Callahan, "Cultural Similarities and Differences in the Design of University Web sites," November 2005. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2006.tb00312.x/full.

[26]S. F. M. A. Mohammadi Akheela Khanum1, "Arabic Interface Analysis Based on Cultural Markers," Riyadh.

[27]E. A.-S. E. A.-A. Gaye Karaçay-Aydın, "Cultural Variability in Web Content: A Comparative Analysis of American and Turkey websites," International business research, Turkey, 2010.

[28]F. M. Ivorra, "Cultural Values and Interpersonality in Spanish and British University Websites," EPiC Series in Language and Linguistics, Spain, 2017.

[29]H. M. ,. P. Tanveer Ahmed, "Website Design and Localisation: A Comparison of Malaysia and Britain," International Journal of Cyber Society and Education, London, 2008.

Heart Rate Monitoring and Activity Recognition using Wearables

Toon De Pessemier, Enias Cailliau, and Luc Martens

imec - WAVES - Ghent University
Technologiepark-Zwijnaarde 15
9052 Ghent, Belgium
Email: toon.depessemier@ugent.be enias.cailliau@ugent.be luc1.martens@ugent.be

Abstract—**Wearables enable measuring physical activities and heart rate using accelerometer and heart rate sensor. However, the output of these sensors is often aggregated into a general activity status without detailed analysis and the link between activity recognition and heart rate measurement is often missing in health apps. This paper compares heart rate measurements during physical activity of three wearable devices: a specialized sports device with chest strap, a fitness tracker, and a smart watch. Due to unintended shifts of the device with respect to the wrist, the fitness tracker and smart watch have difficulties to measure sudden variations in heart rate. During the physical activity, movements of the user's wrist were measured using the accelerometer of the wearable. Correlations in the data patterns of the heart rate sensor and accelerometer are identified. Both sensors are used as input for personal recommendations for physical activities with a rule based filter. These recommendations are tailored to the user's physical capabilities and preferences by matching them to a user profile that is learned from the user's data. Combining the insights from heart rate sensor and accelerometer may allow to improve the accuracy of detecting physical activities, estimate the intensity of an activity, and generate more accurate recommendations.**

Keywords–Activity Recognition; Health Information; Recommendation; Personalization.

I. Introduction

The last decade, more formal and informal health information has become available, with the perspective of a new generation of well-informed, healthy individuals. This phenomenon turns users into health information producers and consumers by offering a multitude of health information services and data [1][2].

To cope with the problem of information overload incurred by this growing availability of data, recommender systems are used as an effective information filter and at the same time as a tool for providing personal suggestions [3][4]. These recommenders may suggest a specific fitness activity or a running trail out of the many available physical activities. But good recommendations should match the physical capabilities of each individual.

To assess the physical load of an activity for a user, measuring the user's physical movements (e.g., using a pedometer) is insufficient, since this neglects the user's effort with respect to his/her physical capacities. The user's physical limits and the intensity of an activity for a user can be estimated by the combination of heart rate measurements and motion sensors [5]. Recent wearable devices are often equipped with accelerometers for measuring movements and heart rate sensors. However, the accuracy of heart rate measurements using these devices is still unclear.

For heart rate monitoring, various methods exist. For this study with wearables, the two most important methods are electrocardiography and photoplethysmography. Electrocardiography (ECG) is the process of recording the electrical activity of the heart using electrodes placed on the skin [6]. These electrodes detect the small electrical changes on the skin that arise from the heart muscle's electrophysiologic pattern of depolarizing during each heartbeat. For medical purposes, e.g., in hospitals, this technique is applied with 10 electrodes, placed on the patient's limbs and on the surface of the chest.

Photoplethysmography (PPG), also known as optical heart rate sensing, is monitoring heart rate using photo diodes and LEDs [7]. Green light is absorbed by blood, hence its red color. When a light source is covered by a body part (e.g., the wrist in case of a wearable), the light is partially absorbed by the blood and partially reflected. The photo diode captures the reflected light. During a heart beat, more light is absorbed and the photo diode detects a reduction in green light intensity. Although a green LED provides the most accurate results, an infrared LED is often used since this consumes less energy. PPG is a cheap method for measuring heart rate, often used in wearables, but has some disadvantages. Motion artifacts can reduce the accuracy during exercises and free living conditions. Person-dependent variations may also influence the measurements, e.g., a different blood perfusion induces a different absorption of light. This paper discusses the use of PPG in wearables for heart rate measuring (Section IV).

Besides heart rate measurements, wearables can perform activity recognition based on the motion detected by the accelerometer. This typically results in a few statistics about the user's physical activity, such as the number of steps taken or the average speed of a running session; but the recognition of specific physical exercises is often still missing. More advanced solutions for activity recognition are often relying on multiple sensors placed on different parts of the body, e.g., on the chest and on the hip, composing a body sensor network [8]. However, this is often considered too intrusive for daily activities. Therefore, this study investigates activity recognition using popular wearable devices (Section V).

The goal of this study is to investigate the accuracy of heart rate measurements obtained with different wearables, and to analyze if measurements of heart rate sensor and accelerometer can be combined for an accurate activity recognition. According to our knowledge, this is one of the first studies that compares wearables worn around the wrist and a sports device with a chest strap for heart rate measurements during a physical activity with a lot of movement of the wrist. These measurement data are the input of recommender systems, which can improve human-web interaction by personalizing interfaces of web applications with tailored suggestions for physical activities. This study presents a rule based filter as recommender system.

The remainder of this paper is organized as follows. Section II refers to interesting related work. An overview of the wearable devices used in this study is provided in Section III. The next sections discuss the measurements of the wearables: the heart rate measurements are discussed in Section IV, activity recognition is the topic of Section V, and the usage of the combination of both is covered in Section VI. Section VII is about the rule based filter to generate personalized recommendations. Section VIII draws conclusions and points to future work.

II. Related Work

The rising interest in health-related data and applications strengthens the need to monitor heart rate and automatically recognize physical activities on a daily basis. Although the commercial sports devices and wearables are equipped with the necessary hardware to accomplish this challenging task, their accuracy is still unclear.

For commercially available breast belt measuring devices, detailed evaluations of the accuracy have been performed [9]. But for recent wearable devices, only a limited number of studies investigated the accuracy of heart rate data, often in specific conditions. In non-moving conditions, heart rate monitoring using a wrist-worn personal fitness tracker has been evaluated with patients in an intensive care unit [10]. The measured values were slightly lower than those derived from continuous electrocardiographic monitoring, i.e., the medical method for heart rate monitoring. The authors concluded that further evaluation is required to investigate if personal fitness trackers can be used in hospitals, e.g., as early warning systems. Another very related study has investigated the accuracy of step counts and heart rate monitoring with wearables [11]. Test subjects were asked to walk a specific number of steps during the measurements. The accuracy of the heart rate measurements with the tested wearable devices showed to be very high. Our paper contributes to the domain of health monitoring with wearables by studying the accuracy of heart rate measurements during intensive physical activities, and with various types of wearable devices.

In the domain of activity recognition with wearables, the focus is often on the classification of movement or transportation types. Hidden Markov models have been proposed [12] to recognize different physical activities, such as driving a car, riding a bicycle, walking, or standing still. In recent Android versions, similar activity recognition functionality is available through Google's activity recognition API [13]. In contrast, our research targets activities that cannot be classified based on the movement speed, but are characterized by specific hand or arm movements, such as Dumbbell Biceps Curl exercises. Our focus is on recognizing the number of repetitions in view of tracking the physical load, rather than on classifying the activities.

The growing availability of these health data on the World Wide Web has brought the problem of information overload [14] to the ehealth domain. For instance, too many sports schedules are available in online databases, but only a minority is matching the physical capabilities and preferences of an individual. This emphasizes the need to personalize health information and services, i.e., "adapting the content, with the aid of computers, to the specific characteristics of a particular person" [15]. Personalized recommendations, tailored messages, and customized information have shown to be far more effective than the non-personalized alternative [3][4]. Unfortunately, many of these recommender systems rely on the manual input of users reporting their performed exercises. Our solution combines automatic activity recognition and heart rate measurements, which are used as input for a rule-based recommender system.

III. Wearable Devices

For measuring heart rate, three types of wearables were used: a smart watch, a fitness tracker, and a specialized device.

A. *Smart Watch*

Smart watches are equipped with various sensors but are not medically approved. The smart watch is a general purpose, fashionable device with features such as tracking physical activities and informing users. From a commercial viewpoint, the target group of customers is not limited to sports people, but includes also a broader group of people who like the design or the extra features of the gadget. Smart watches often have hardware capabilities allowing to extend their functionality with additional apps. In this study, the Huawei Watch was used as smart watch for the measurements because of its popularity and typical smart watch characteristics (e.g., Android Wear). Heart rate measurements are based on photoplethysmography. To capture heart rate data in real time, a special Android Wear app was developed for the Huawei Watch. This app communicates with our developed Android app running on a smartphone through the Wearable Data Layer API.

B. *Fitness Tracker*

These devices, typically worn around the wrist, measure movements and behavior, such as the number of steps taken, sleeping patterns, and sports activities, e.g., a light jog or a mad sprint. As with smart watches, fitness trackers are seldom approved for medical purposes. They are equipped with multiple sensors, such as a 3-axis accelerometer to monitor movement in every direction, an altimeter to measure altitude and keep track of the traveled height, and sometimes a gyroscope to measure orientation and rotation. Compared to smart watches, fitness trackers are more focused on tracking physical activities. In this study, the Microsoft Band 2 was chosen as fitness tracker because of two reasons. It allows real time analysis of sensor data (heart rate data using photoplethysmography and movement data through the accelerometer) and Microsoft provides a comprehensive API. The API offers functionality, such as aggregating the results of a query, thereby shifting the computational load to the Microsoft servers.

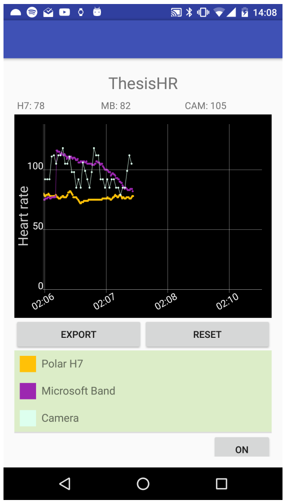
C. *Specialized Device*

The main purpose of this type of devices is measuring heart rate. Typical examples are pulse-oximeters, blood pressure monitors, and heart rate chest straps. These often have only a limited number of sensors and a limited number of features.

In this study, the Polar H7 was used as specialized sports device. This is a popular heart rate chest strap, which produces very accurate measurements (correlation of 0.97 with true heart rate [16]). Heart rate is measured using electrodes in the chest strap that detect heart pulse via an electronic signal.

IV. Heart Rate Measurements

To gather, store, and analyze heart rate measurements of these three device types, an Android app was developed and deployed on a Google Nexus 6P smartphone. Figure 1 shows a screenshot of this app. The wearable devices have a Bluetooth communication link with this smartphone and the app has a separate service running for each device to transfer the raw data to the smartphone.



H7: 78 MB: 82 CAM: 105

Figure 1. Screenshot of the Android app for gathering heart rate data.

A. In Rest Condition

To evaluate the accuracy of heart rate measurements of the three device types, these heart rate measurements were compared with the measurements of a specialized device that is approved for medical purposes, i.e., the Omrom M6 Comfort [17]. The Omrom M6 is a blood pressure monitor, which has to be attached around the upper arm for measuring the heart rate. Heart rate was measured for two persons, in a rest condition, in a home environment, at two different times. The first test subject (male) had a low natural heart rate, whereas the second test subject (female) had a rather high heart rate in rest condition. Table I shows for each device the mean, standard deviation, and median, indicating that all devices provide consistent results. The mean values and small standard deviation show that in rest condition, heart rate measurements obtained with these devices can be considered as reliable. The measurements of the Omrom M6, which is medically approved, are considered as the correct heart rate. The measurements of the Polar H7 are the most similar to the measurements of the Omrom M6. Since a blood pressure monitor is rather expensive and not practical during sports activities, the Omrom was not suitable to measure heart rate during physical activities. Therefore, the Polar H7 was considered as the reference device during physical activities.

B. During Physical Activity

Figure 2 shows the heart rate measurements obtained with the different devices during physical activity, more specifically Dumbbell Biceps Curl. These exercises for bicep muscles were performed in a fitness room by two people. Similar results are obtained for both persons (results are shown for only one person). During physical activities, such as Dumbbell Biceps Curl, measuring heart rate cannot be performed with the blood pressure monitor due to body movements and the non-wearable characteristic of the Omrom M6. For the three wearable devices, a significantly different signal of the heart rate measurements can be witnessed during this physical activity.

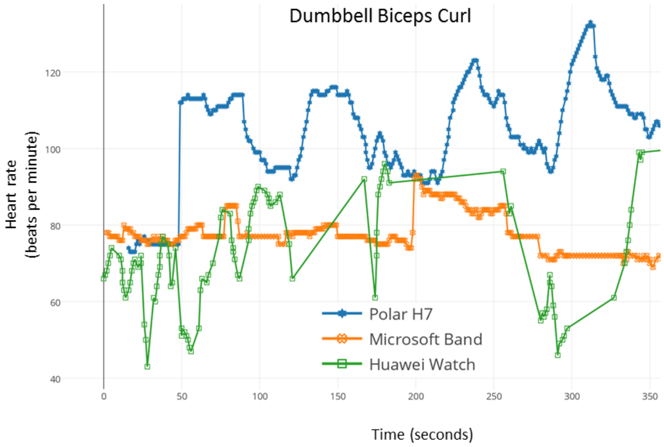


Figure 2. Heart rate measurements during Dumbbell Biceps Curl.

The heart rate signal produced by the *Polar H7* clearly shows a repetitive pattern that corresponds to the repetitions of the Dumbbell Biceps Curl exercise. The accurate measurements can be explained by the use of the chest strap, which is less influenced by movements than the devices worn around the wrist.

The heart rate registered by the *Microsoft Band 2* is consistently lower than the values measured by the Polar device. Moreover, rapidly varying heart rates due to periods of intensive physical activity are difficult to detect. As a result, the subsequent repetitions of the physical exercise are not clearly visible in the graph of the Microsoft Band in Figure 2.

With the *Huawei Watch*, less measurement samples are obtained compared to the Polar H7 and the Microsoft Band. Movements of this device, which is worn around the wrist, cause interruptions in the measurement process. Changes in the device's position relative to the wrist induce a sensor recalibration and can be noticed in Figure 2 as the time periods without measurement data from the Huawei Watch. Periods of intensive physical activities are noticeable by the variations in the data of the heart rate measurements. But the interruptions in the measurement data might be a problem for detailed heart rate monitoring during physical activities.

V. Activity Recognition

In order to monitor the proper execution of physical exercises by users, wearables can be used to register specific physical movements. The Dumbbell Biceps Curl exercise is a typical activity that allows detection of repetitions of this

TABLE I. MEAN $\bar{x}$, STANDARD DEVIATION σ, AND MEDIAN $\tilde{x}$ OF THE HEART RATE IN REST CONDITION WITH TWO USERS AT TWO TIMES

	User 1 - Test 1		User 1 - Test 2		User 2 - Test 1		User 2 - Test 2	
Device	$\bar{x} \pm \sigma$	$\tilde{x}$	$\bar{x} \pm \sigma$	$\tilde{x}$	$\bar{x} \pm \sigma$	$\tilde{x}$	$\bar{x} \pm \sigma$	$\tilde{x}$
Smart Watch (Huawei Watch)	55±2.0	55	55±2.0	56	73±3.3	73	72±3.2	71
Fitness Tracker (Microsoft Band)	50±2.9	50	64±6.0	64	75±3.3	75	76±1.7	76
Specialized Sports Device (Polar H7)	56±1.7	56	59±1.4	59	77±3.0	76	80±3.7	79
Specialized Blood Pressure Monitor (Omrom M6)	55±2.8	55	58±2.9	58	76±2.5	76	84±4.2	84

exercise by using data of the accelerometer of the wearable worn around the wrist. Figure 3 shows the pattern of the accelerometer data, gathered with the fitness tracker around the wrist, during the execution of this exercise. Although the execution speed of the activity and the body characteristics of the user may have an influence on the absolute values of the data of the accelerometer, the typical pattern consisting of local minima and maxima can be witnessed for every repetition.

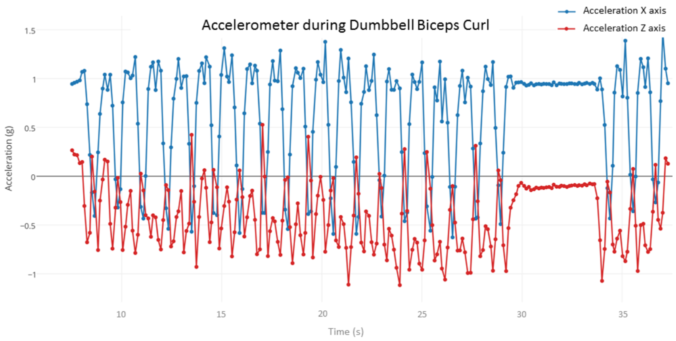


Figure 3. Measurements of the accelerometer of a wearable during Dumbbell Biceps Curl.

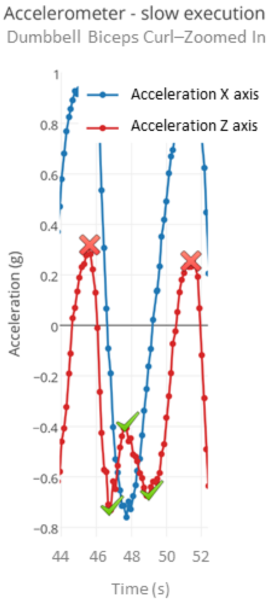


Figure 4. Detailed view on the local optima in the accelerometer data during Dumbbell Biceps Curl.

For each repetition of the Dumbbell Curl activity, 5 local optima on the Z-axis and 3 on the X-axis can be witnessed as visible in Figure 4: 1) a maximum on the Z-axis co-occuring with a maximum on the X-axis, 2) a minimum on the Z-axis, 3) a maximum on the Z-axis co-occuring with a minimum on the X-axis, 4) a minimum on the Z-axis, and 5) a maximum on the Z-axis co-occuring with a maximum on the X-axis. The red cross marks in Figure 4 denote the beginning and end of a repetition of the exercise, the green check marks indicate the intermediate optima. Recognizing the Dumbbell Biceps Curl execution based on the identification of a sequence of these 5 events has some benefits. The recognition process requires limited processing power, allowing real-time recognition (e.g., for e-coaching purposes) and making it usable on devices with limited processing power, such as wearables. Moreover, the detection of local optima makes the recognition method directly usable for different variations of the Dumbbell Curl, such as Concentration Curl, Hammer Curl, and Barbell Curl.

VI. HEART RATE AND ACTIVITY RECOGNITION COMBINED

Monitoring heart rate and simultaneously recognizing repetitions of an activity with the accelerometer allow a better health monitoring and e-coaching during workouts. Since raw data streams of both sources (heart rate sensor and accelerometer) are suffering from inaccuracies, the combination of both can improve health monitoring. For example, the intensity of a physical activity for an individual can be estimated based on the heart rate data. But in case of measurement interruptions in the heart rate data, accelerometer data can be used to estimate the performed physical activities.

For e-coaching purposes, our Android app uses both data sources to instruct the user during physical exercises thereby maintaining a healthy heart rate. Repetitions of an exercise are recognized and through text-to-speech techniques the repetitions are counted aloud or shown on the screen of the wearable. Each physical activity has a target range of the heart rate that can be expected during the performance. If the measured heart rate is out of this range, the user is warned by a clear indication on the screen of the wearable. After performing an activity, the app evaluates the intensity of the physical exercise as “too intensive”, “to easy”, or “just good”.

VII. USER PROFILING AND RECOMMENDATIONS

The physical exercises measured with the accelerometer, the heart rate, and the characteristics of the exercises are stored online in a user profile. Users can access their profile using a web application to analyze their history of physical activities. Moreover, this user profile is used for personalization of suggestions for new activities in our Android app, such as a set of Dumbell Biceps Curl exercises, a running track, a cycling track, etc.

To match the user’s preferences and physical capabilities to the physical activities and select the most suitable ones as recommendations, the activities of each type are processed by a specialized rule based filter. This rule based filter makes a selection of the activities based on characteristics of that type of activity, e.g., the distance for a running activity or the weight and number of repetitions for Dumbbell Biceps Curl. For each type of activity, a separate rule based filter is used in order to take into account the user's experience level for each activity individually. For example, suppose a user is an excellent runner. Recommendations for intensive running activities will be the most suitable, given the user's physical capabilities and history. Now, suppose that this user visits the gym for the first time with the goal of training the arm muscles. The user's body is not used to intensive Dumbell Biceps Curl activities. Recommendations at the level of starting users might be appropriate here. Therefore, a separate rule based filter is assigned to each activity type to handle these differences in training level for users. In future work, explanations about the recommendations can be added in order to further convince the user to adopt one of the offered recommendations [18]. These explanations can be expressed in terms of (the progress of) the physical capabilities of the user.

The rule based functionality is implemented based on Drools [19]. Drools is a business rules management system with business rules engine that is scalable and extendible through the use of drl files containing the rules. The goal of these rules is to filter the available activities in order to come up with the most suitable activity for the user taking into account the conditions/context at the moment of the recommendation.

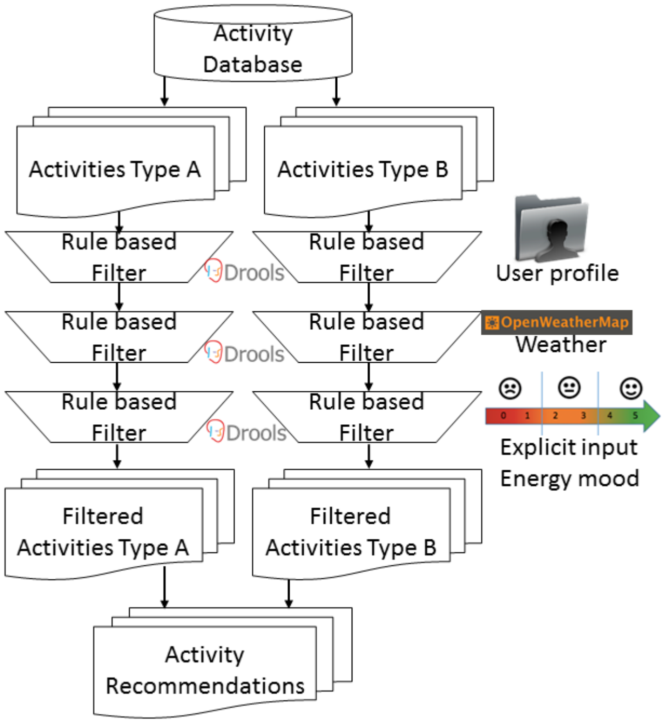


Figure 5. Graphical overview of the rule based filtering of the activities.

Figure 5 gives a graphical overview of the rule based filtering that is applied to the activities. The rules check the following conditions: 1) *User profile*: Do the length and intensity of the cycling or running track match the user's physical capabilities and habits? The target length of a track is similar to the length of the tracks in the user's history, but a small difference is tolerated. The intensity of the track is estimated based on the difference in altitude meters. For gym exercises, the intensity is estimated based on the weight or resistance of the fitness equipment and the number of repetitions. 2) *Weather*: Does the activity match the current weather conditions? For example, outdoor running activities are not recommended when it rains. To retrieve weather data at the user's location, the OpenWeatherMap.org REST API [20] is used. 3) *Energy Mood*: Does the activity match the user's energy level of the moment? The energy mood is a value, ranging from 0 to 5, that users can specify in the Android app to express their current feeling, e.g., energetic, tired, or something in between.

VIII. Conclusion

This study discussed the usage of wearables for heart rate measurements and the automatic recognition of physical activities. Measurements with a fitness tracker and a smart watch showed to be very accurate in case of limited physical movement, e.g., in a state of rest. In contrast, a discrepancy in the measurements of the wearables is witnessed during intensive physical activities (Dumbbell Biceps Curl). Shifts of the wearable with respect to the position of the wrist induce inaccuracies or even interruptions in the measurement process thereby hindering the monitoring of heart rate variations. Specialized sports devices, using a sensor with chest strap, produce more accurate heart rate measurements, even during intensive physical activities, and enable recognizing subsequent repetitions of a physical activity based on the periodic peaks in the heart rate. Therefore, our advise is to use a device with a chest strap for heart rate measurements in case of physical activities that involve a lot of movement of the wrist.

Besides, raw data produced by the accelerometer of wearables can be used to recognize repetitions of physical exercises with characteristic movements of wrist/hand/arm. E.g., the Dumbbell Biceps Curl exercise can be recognized based on a specific pattern with 5 local optima on the X and Z-axis of accelerometer data. Both raw data streams (heart rate and accelerometer data) can be combined for further analysis, but also to assist the user in coaching tasks, such as counting the number of times an exercise is performed, or instructing to decrease or increase the intensity of the exercise. Automatic activity recognition can help the user by reducing the burden of providing input about the performed activities in digital health services or fitness apps. Moreover, recognized activities can be stored in a user profile, which can be used as an indicator for the user's physical capabilities and habits. Based on this profile, the current weather, and the user's mood, personalized recommendations are generated using a set of rule based filters. In future research, we will investigate the recognition of other physical exercises and relate the resulting accelerometer data to heart rate data more in depth.

References

[1] G. Eysenbach, "Consumer health informatics," British medical journal, vol. 320, no. 7251, 2000, p. 1713.

[2] R. J. Cline and K. M. Haynes, "Consumer health information seeking on the internet: the state of the art," Health education research, vol. 16, no. 6, 2001, pp. 671–692.

[3] V. J. Strecher et al., "The effects of computer-tailored smoking cessation messages in family practice settings," Journal of Family Practice, vol. 39, no. 3, 1994, pp. 262–270.

[4] M. K. Campbell et al., “Improving dietary behavior: the effectiveness of tailored messages in primary care settings.” American journal of public health, vol. 84, no. 5, 1994, pp. 783–787.

[5] P. S. Freedson and K. Miller, “Objective monitoring of physical activity using motion sensors and heart rate,” Research quarterly for exercise and sport, vol. 71, no. sup2, 2000, pp. 21–29.

[6] L. N. Katz and A. Pick, Clinical electrocardiography. Lea & Febiger, 1956.

[7] A. Reisner, P. A. Shaltis, D. McCombie, and H. H. Asada, “Utility of the photoplethysmogram in circulatory monitoring,” The Journal of the American Society of Anesthesiologists, vol. 108, no. 5, 2008, pp. 950–958.

[8] G. Plasqui, A. Bonomi, and K. Westerterp, “Daily physical activity assessment with accelerometers: new insights and validation studies,” Obesity Reviews, vol. 14, no. 6, 2013, pp. 451–462.

[9] M. Weippert et al., “Comparison of three mobile devices for measuring r–r intervals and heart rate variability: Polar s810i, suunto t6 and an ambulatory ecg system,” European journal of applied physiology, vol. 109, no. 4, 2010, pp. 779–786.

[10] R. R. Kroll, J. G. Boyd, and D. M. Maslove, “Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: A prospective observational study,” Journal of medical Internet research, vol. 18, no. 9, 2016, p. 253.

[11] F. El-Amrawy and M. I. Nounou, “Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?” Healthcare informatics research, vol. 21, no. 4, 2015, pp. 315–320.

[12] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, “A hybrid discriminative/generative approach for modeling human activities,” in Proceedings of the 19th international joint conference on Artificial intelligence. Morgan Kaufmann Publishers Inc., 2005, pp. 766–772.

[13] Google, “Activity Recognition API,” 2018, [Online] Retrieved March, 2018, from https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionApi.

[14] L. Fernandez-Luque, R. Karlsen, and L. Vognild, “Challenges and opportunities of using recommender systems for personalized health education.” Studies in health technology and informatics, vol. 150, 2008, pp. 903–907.

[15] H. de Vries and J. Brug, “Computer-tailored interventions motivating people to adopt health promoting behaviours: Introduction to a new approach,” Patient Education and Counseling, vol. 36, no. 2, 1999, pp. 99–105.

[16] M. Altini, “Heart Rate Variability for Training,” 2013, [Online] Retrieved March, 2018, from http://www.marcoaltini.com/blog/heart-rate-variability.

[17] Omron, “Automatic Blood Pressure Monitor - Model M6 Comfort IT,” 2015, [Online] Retrieved March, 2018, from http://www.omron-healthcare.com/en/support/manuals/download/m6-comfort-it-hem-7322u-e-en.

[18] N. Karacapilidis, S. Malefaki, and A. Charissiadis, “A novel framework for augmenting the quality of explanations in recommender systems,” Intelligent Decision Technologies, vol. 11, no. 2, 2017, pp. 187–197.

[19] J. Community, “Drools - Business rules management System,” 2018, [Online] Retrieved March, 2018, from http://www.drools.org/.

[20] OpenWeatherMap, “Current weather and forecast,” 2018, [Online] Retrieved March, 2018, from http://openweathermap.org/.

Potential Big Data Future Challenges for the GoodTurn System

Katy Snyder, Kevin Daimi , Jacob Vanassche
Computer Science and Software Engineering
University of Detroit Mercy
Detroit, USA
email: {daimikj, snyderke, vanassje}@udmercy.edu

***Abstract*—GoodTurn is an application designed and implemented by the University of Detroit Mercy through a grant from Ford Motor Company. It is a goods-moving system. In a manner similar to Uber, the application is aimed at facilitating and managing Ford employees' donation of their time and vehicles to assist the community by moving their goods and resources. The stakeholders—drivers, donors, and nonprofit/nongovernment organizations (NPO/NGO)— will use their iPhones and Android-based phones to connect to the application free of any charge. Increasing amounts of data are currently being generated. It is anticipated that the data will exceed one terabyte in the next few years. To address the benefits of the availability of future big data, this paper will present potential future challenges of the resulting big data's analytics. In particular, these challenges will address the decision-making needs of Ford, NPOs/NGOs, drivers, and donors. The paper will not address the implementation of any actual big data analytics using tools as the data is not yet completely developed. However, once the anticipated big data is generated, appropriate tools will be employed to obtain the needed knowledge and uncover the value of the stored data.**

Keywords—GoodTurn System; Big Data Analytics; Big Data Lifecycle; Prediction; Classification; Clustering

I. Introduction

GoodTurn is a system developed by the University of Detroit Mercy with a grant from Ford Motor Company to facilitate the work of nonprofit/nongovernment organizations, NPOs/NGOs, when dealing with donors. Ford's employees volunteer their vehicles and time to move goods and resources donated by people to the NPOs/NGOs designated locations. GoodTurn currently runs on iOS-based devices, with Android and web-based versions being developed. It is anticipated that this application will generate big data in the near future.

Currently, we are experiencing an explosion in the rate of the quantity of big data being generated due to the ever-increasing amount of data resulting from Web applications, networks, log files, vehicle performance, social media tweets, mobile applications, transactional applications, and sensing devices. These big data include massive potential hidden knowledge that can add business value to a variety of fields including healthcare, biological systems, transportation, online advertising, crash reporting, performance monitoring, energy management, student registration and financial services [1]-[3]. Innovations with big data vow to transform the way we live, work, and think by empowering process optimization, facilitating insight discovery and improving decision making [4]. With the evolution and improvement of web and other technologies, the enormous amount of data of different types is briskly generated and the amount of knowledge multiplies drastically [5][6]. Users are saturated with data in this big data time, however, identifying valuable data to obtain worthwhile information and knowledge has never been an easy task. Uncovering valuable information from the titanic amount of data is becoming more important, and many countries and enterprises are devoting time and money to acquirement and analysis of data [7].

Big data is generally defined by the three Vs; Volume, Velocity, Variety, and it has been very vital and constructive in achieving treasurable values with regards to supporting decision making, illuminating new insights, and process optimization [8]. The bulk of data created is constantly growing and taking the form of a variety of structures, and can be in motion and at rest. For example, Google receives over one billion queries per day, Twitter gets more than two hundred and fifty million tweets on a daily basis [9], Facebook goes through more than eight hundred million updates per day, and YouTube causes more than four billion views per day. The data generated is estimated in the order of zeta bytes at the present time, and it is intensifying at a rate around 40% per year [10].

Big data analytics involves applying advanced analytic techniques to exceedingly large and diverse datasets with the possibility of including structured, semi-structured, and unstructured data to explore new capabilities and insights [11]. If big data analytics techniques are deployed in a timely manner, the outcome can produce actionable insights that add significant value to organizations and help them improve the decision-making process, and create various opportunities for business improvements and success [12]-[13].

E. Žunić, A. Djedović, and D. Đonko [14] indicated that assorted types of mobile communication devices are more frequently used to access applications. They added that with mobile devices, anyone can clearly use or even develop a mobile application adding to the further explosion of applications and data. Mobile communication networks grant an immense range of communication services producing a substantial amount of network data [15]. The current innovations of wireless technologies in various forms and the constantly increasing mobile applications have turned mobile cellular networks into both generators and carriers of massive data [16].

The GoodTurn system was developed by the University of Detroit Mercy with a grant from Ford Motor Company. The goal of the application is to facilitate the moving of goods and resources donated by individuals to various NPOs/NGOs. To this extent, Ford employees (drivers) volunteer their vehicles and time to move these goods from donors' locations to the NPOs/NGO's locations. The GoodTurn application runs on iPhones and will soon run on Android-based phones. As stated above, mobile applications generate massive data. It is anticipated that GoodTurn will produce a terabyte within the next few years. This paper discusses a number of possible big data analytics applications when the big data matures. The outcomes of these applications will furnish actionable insight to Ford Motor Company, NPOs/NGOs, donors, and drivers. The rest of the paper is organized as follows: Section II will provide a brief description of the GoodTurn system. Section III will introduce the evolution of the GoodTurn big data. The GoodTurn big data lifecycle is presented in Section IV. Section V highlights the potential analytics of the future big data. Finally, the paper is concluded in Section VI.

II. GoodTurn system description

To get the flavor of the GoodTurn system, sample requirements and an overview of the system architecture with sample interface will be presented. Details of the GoodTurn software design are introduced in [17]. Security of the GoodTurn system is discussed in [18].

A. Functional requirements

The GoodTurn system has been developed using the client-server methodology. The clients will be accessing the system using iPhones, and Android-based phones. The GoodTurn system's functional requirements were gathered from Ford employees, non-profit organizations (NPOs), non-government organizations (NGOs), and the public. Samples of these requirements are shown below. Here, "requesters" represent NPOs or NGOs.

- The login screen must contain an option to save the user's email.
- The system must allow the requestor to reject a specific driver in the future.
- The system must allow the driver to reject a specific requestor/organization in the future.
- The system must allow the requester to verify a job was completed.
- The system should allow the user to register if they do not already have an account.
- The system must allow rating of users that were involved in a job.
- The system must allow users that were involved in a job to provide feedback.
- The system must provide a list of available jobs.
- The system should allow the driver to accept a job.
- The system must allow drivers to cancel accepted jobs.
- The system must allow the requester to start a new job.

B. Nonfunctional requirements

The constraints on the GoodTurn's functional requirements include performance, usability, security, privacy, reliability, and maintainability features. A small sample of these nonfunctional requirements will be presented below.

- The system should allow drivers and requesters to sign in within 5 seconds.
- Displaying blacklisted drivers for a specific requester should take no more than 5 seconds.
- Drivers and requesters should be able to use the system without any training.
- The system should provide messages to guide the users when invalid information is entered.
- Drivers, requesters, and system administrators should be authenticated
- Messages exchanged between all parties (drivers, requesters, system administrators) should be confidential.
- The system should not disclose requester information to non-drivers.
- The system should not disclose a driver information to non-requesters.
- The system must detect, isolate, and report faults.
- Backup copies must be stored at a different location specified by the NPO/NGO.
- Errors should be easily corrected using effective documentation.
- Additional features should be added without considerable changes to the design.

C. System architecture

An architecture embodies the high-level structures of a software system. By examining the architecture, one can conclude how multiple software components collaborate to accomplish their tasks. GoodTurn follows the three-tiered client-server architectural style. The GoodTurn system architecture is functionally decomposed into various functional components. To illustrate that, Figure 1 is used to demonstrate the top-level and first-level decompositions. The component, GoodTurn Startup, is further decomposed into second and third levels in Figure 2 below.

D. User Interface

To design a user interface for GoodTurn system, the humans that need to interact with the system need to be identified. These include drivers, NPOs/NGOs, donors, and system administrators. Later, settings for each way the user can communicate with the system need to be established. Samples of the User Interface are provided in Figures 3 and 4 below.

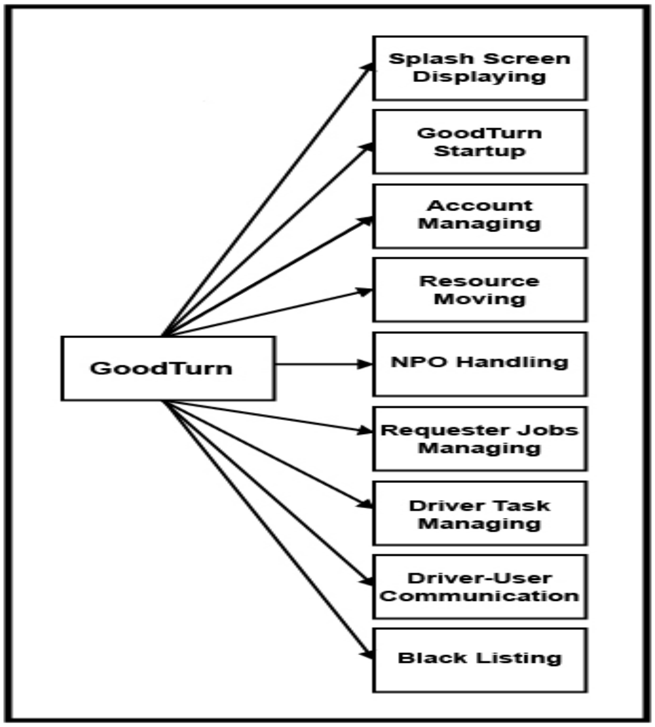


Figure 1. Top-level decomposition

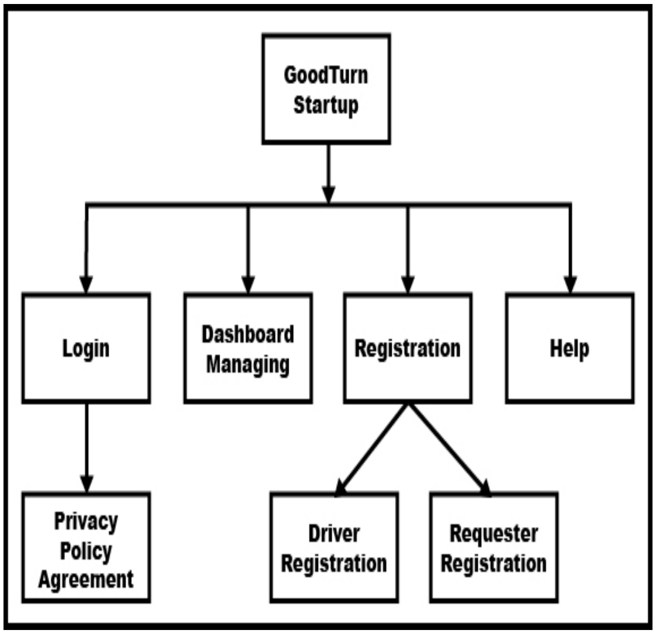
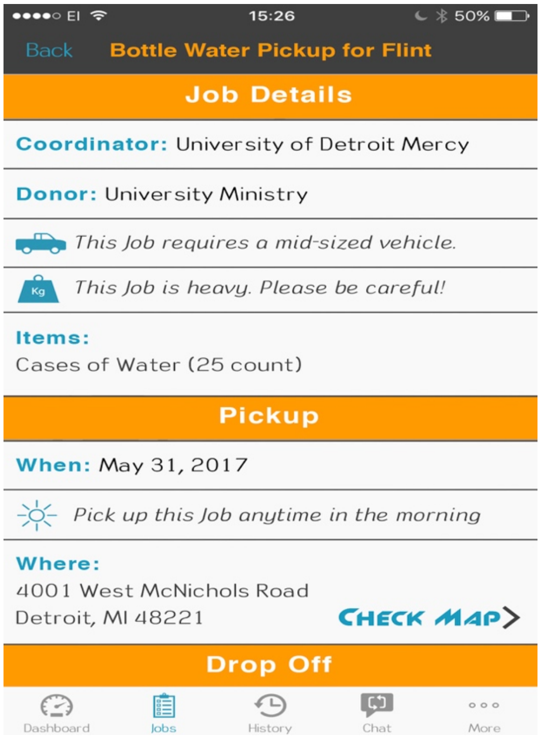


Figure 2. Second/Third levels decomposition



●●●●○ EI 15:26 50%

Back Bottle Water Pickup for Flint

Job Details

Coordinator: University of Detroit Mercy

Donor: University Ministry

This Job requires a mid-sized vehicle.

This Job is heavy. Please be careful!

Items:
Cases of Water (25 count)

Pickup

When: May 31, 2017

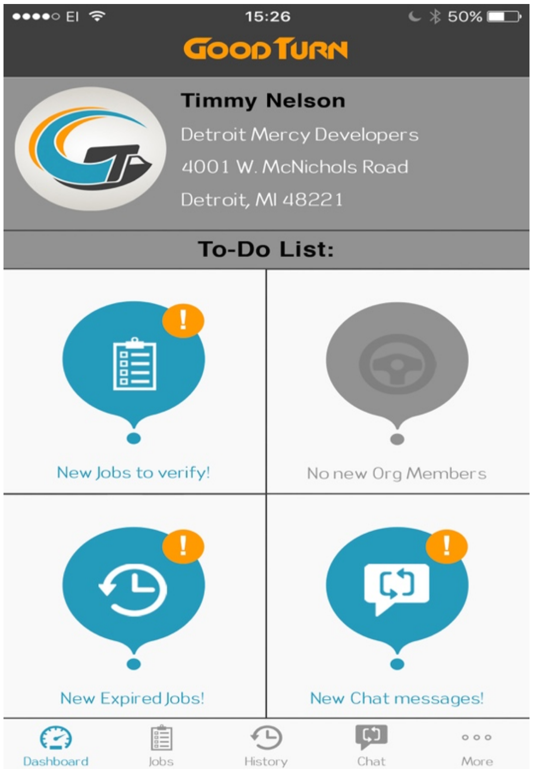
Pick up this Job anytime in the morning

Where:
4001 West McNichols Road
Detroit, MI 48221

CHECK MAP>

Drop Off

Figure 3. Available jobs for drivers



●●●●○ EI 15:26 50%

GOOD TURN

Timmy Nelson
Detroit Mercy Developers
4001 W. McNichols Road
Detroit, MI 48221

To-Do List:

New Jobs to verify!

No new Org Members

New Expired Jobs!

New Chat messages!

Figure 4. Requester’s new job

III. EVOLUTION OF GOODTURN BIG DATA

The GoodTurn system data currently has 404 bytes for each of the stakeholders; NPO, drivers, and donors. In addition, there is a total of 76 bytes for the metadata. This will give a total of 480 bytes per a stakeholder. This represent structured data only. The above data size does not include the huge unstructured data resulting from various feedbacks fields. Details of the data are depicted in Tables I and II below.

There are 75,437 nonprofit organizations (NPOs) in Michigan State [1]. Ford Motor Company's workforce is over 201,000 employees worldwide [2][3]. Michigan has total of 9,928,300 residents [4]. They contribute the equivalent of almost $5 billion to charity each year [5]. This implies that many of those residents donate at least once a year.

The exact figures of NPOs, donors and drivers will be determined when GoodTurn is implemented. These numbers of donors, drivers, and NPOs will continue to grow. As an example, the total number of NPOs in 2013 was 42,886 [6]. Comparing this to 75,437 nonprofit organizations in 2017, it can be concluded it is almost doubled. This implies there will be even more data in the future. Based on these figures, it is anticipated that the GoodTurn application's big data will develop in the near future ahead

TABLE I. INDIVIDUAL STAKEHOLDER DATA TOTALS

Category	Total Bytes
Account	081
Job	260
Chat	005
Dashboard	028
Misc.	030
Total	404

TABLE II. METADATA TOTALS

Category	Total Bytes
Account Metadata	28
Job Metadata	28
Chat Metadata	08
Dashboard Metadata	04
Misc. Metadata	08
Total	76

IV. GOODTURN BIG DATA LIFE CYCLE

The life cycle of GoodTurn big data analytics starts with the generation of big data. Data will continue to accumulate and reach more than a terabyte within few years. A decision should be made whether to store this big data on the cloud or on the GoodTurn server. Both approaches involve security and cost issues. The data needed for big data analytics will be selected and filtered. The filtering will encompass removing the data items that are not desirable for the analytics, and handling missing data values. Once filtering has been taken care of, the needed techniques, tools, and algorithms will be applied to achieve the analytics. As discussed in Section V below, the potential analytics will concentrate on predictions, classification, and clustering. The outcomes of these analytics would be actionable knowledge and insights beneficial to various stakeholders. Once the insights are available, the stakeholders; Ford, NPOs/NGOs, drivers, and donors, will be able to use these insights to inform decisions. The next cycle will start at big data generation with more data generated. This cycle is depicted in Figure 5 below.

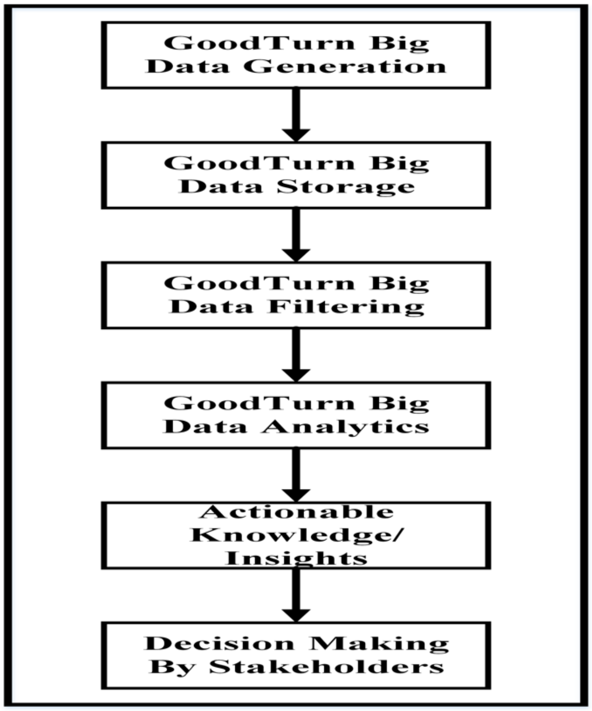

GoodTurn Big Data Generation
GoodTurn Big Data Storage
GoodTurn Big Data Filtering
GoodTurn Big Data Analytics
Actionable Knowledge/ Insights
Decision Making By Stakeholders

Figure 5. GoodTurn big data lifecycle

V. POTENTIAL ANALYTICS

Potential future analytics when the big data matures will be discussed. These proposed potential analytics will be linked to the GoodTurn's stakeholders. In other words, each stakeholder can only own the analytics with which they are concerned.

A. Ford Motor Company

The quality of vehicles and quantity of sales are important factors in automotive industry. The analytics below should address these concerns. Ford Motor Company also values and promotes philanthropic activities for its employees.

Note that analytics in bullets; 2,3, 4, and 5 will only be possible if the GoodTurn system allows comments regarding the selected type/model of vehicles in the Feedback field. Currently, this is not the case, but it is anticipated that will be the case in the future.

- *Predicting the demand for specific Ford vehicles*: NPOs/NGOs select the type of vehicle for moving the donor's goods/resources. Within the NPOs/NGOs, individuals decide the size of vehicle. Selecting the size of the vehicle will imply selecting the type/model of the vehicle because these data are already available for those individuals. If some vehicles are frequently selected, then those vehicles reflect the taste of users and could indicate these are preferred or on demand. Therefore, the *characteristics* of those *individuals* will reflect the type of people who would possibly buy such vehicles. Ford can continue to promote these vehicles and improve their future models. For vehicles that are infrequently requested, Ford will study the reasons behind that and improve these vehicles. If non-Ford employees are allowed to be drivers in the future, *competitive* vehicles will be involved. If some non-Ford vehicles are frequently used, then the *characteristics* of those individuals will guide Ford to investigate all the *design* aspects of the frequently competing demanded vehicles and make decisions on Ford competing vehicles.
- *Determining the characteristics of individuals who will complain about certain vehicles*: Complaining about a vehicle should be interpreted as complaining about the type and model of the vehicle. NPOs/NGOs and donors provide feedback after the job is completed. They can also provide feedback if they have to cancel the job while in transit. The feedback containing complains about the vehicles can be analyzed to find out from this sample the type of individuals that might be complaining about certain type/model of vehicles in the future. The resulting analytics could lead to improving those vehicles. In other words, those complains would be interpreted as *buyers'* needs. Therefore, people with similar *characteristics* can be targeted using the improved vehicles. If the *complains* are about competitor's vehicles, Ford can approach buyers by promoting the features/*characteristics* their vehicles have as compared to others.
- *Predicting the characteristics and features of vehicles that individuals might complain about*: Based on various complains about the vehicles used in moving goods, various features and specifications of those vehicles will be analyzed to conclude the features that might have caused these *complains*. Further testing and investigation will be carried out to isolate the focal *features*. Having done that, those features will either be avoided or improved in other models that share the same characteristics in the future. If the *complains* are about *competitor's* vehicles, Ford will check if the predicted features causing the complains exist in their vehicles and improve them.
- *Predicting whether a new driver (volunteer) will undergo complains*: By *analyzing* the feedback about drivers, Ford can use the *characteristics* of the drivers undergoing complains and classify if a new volunteer (driver) might encounter complains. This outcome can be used to effectively screen future drivers.
- *Isolating the characteristics and features of vehicles that will experience frequent problems*: Vehicles in transit can experience problems. These problems will be documented in the feedback. By analyzing those vehicles' features, Ford can determine the features causing the problems and issue the necessary recalls to fix these problems. They can further decide to illuminate some features from future vehicles if fixing/replacing them becomes costly or displeasing.
- *Identify clusters of potential buyers for certain vehicles*: Taking the possible analytics above, and provided the NPOs/NGOs and donors are from specific geographic location, Ford can determine the characteristics of such a geographic location to identify other geographic locations that have similar characteristics to market on-demand vehicles in those areas.

B. NPO/NGO

Nonprofit and nongovernment organizations are interested in getting good number of donors, certain types of goods, concentrating on geographical areas that have frequent donors or many donors, ensuring the selected drivers are reliable.

- *Identifying geographic locations of possible frequent donors*: By analyzing the geographic locations of current donors, an NPO/NGO can select other geographic locations with similar characteristics and properties that could possibly provide frequent donors.
- *Detecting geographic locations of maximum number of donors*: Using the current geographic locations containing the maximum number of donors (not necessarily frequent donors), an NPO/NGO can use the qualities and features of these locations to find out similar geographic locations to target them for possible high volume of donors.
- *Anticipation of a needed type of goods*: This potential analytic can be achieved in two ways. First, a similar analysis to the geographic locations above could be carried out to determine the potential geographic locations that may donate the needed goods and resources based on the traits of current locations providing the needed type of goods. Second, an NPO/NGO can use the characteristics of individuals who donate such needed type of goods to focus on individuals with the same characteristics.

- *Association of a needed type of goods*: NPOs/GPOs can look for features of individuals who donate goods that complement each other (table and chairs for example), and then use these features to aim at possible donors who may donate goods which coexist with each other to be useful.
- *Expectation of the number of certain goods/resources needed for disasters such as hurricane, tornado, and earthquakes*: This can be achieved by targeting individuals (donors) in certain geographic locations, or individuals in disperse locations. Using the traits of individuals who normally donate goods/resources that are useful when a disaster takes place, and geographic locations with the maximum number of disaster-needed goods donations, the right targets would be determined.
- *Predicting potential donors who would default*: NPOs/NGOs spent time and money in calling donors and in following up calls. Donors that prove they are not reliable should be avoided in the future. To accomplish that, characteristics of defaulting donors will be identified and used to predict donors who would default in the future to avoid them.
- *Foreseeing drivers who would default or rejects a request after accepting it*: Centered around the details of those drivers who either do not show up or change their minds after accepting a delivery, a future (new) driver can be classified as either reliable or unreliable. This information can be used to improve screening of future drivers.
- *Predicting drivers who would never reject a request*: another approach for predicting the reliability of drivers is to explore the details of those drivers who have never rejected a request for delivery, and to use the outcome of this exploration to conclude if a new (future) driver would never reject any request.
- *Foretelling drivers who would be willing to drive long distance*: a number of drivers (volunteers) would not go long distance. This will result in delaying the delivery and annoying the donor. Investigating the qualities and characteristics of drivers who carried out long distance tasks will help NPOs/NGOs to determine if a certain driver would be willing to accept a long-distance delivery.

C. Driver

In general drivers need to see NPOs/NGOs do not default and provide accurate details on distance, and type, weight, and size of goods. Furthermore, they will be looking for reliable donors at pickup locations.

- *Predicting the NPOs/NGO's who might cancel their request at the last moment*: It is important for a driver to know if an NPO/NGO would change their mind to avoid wasting time driving to the pickup location and not taking care of their other personal obligations. This could be fulfilled by investigating the characteristics of NPOs/NGOs who have cancelled their request at the last moment and avoid accepting requests from NPOs/NGOs who reveal the same descriptions.
- *Identifying NPOs/NGOs who will not provide precise details about their request*: Drivers can anticipate those NPOs/NGOs who will not provide the accurate details about the request, such as distance, size, type of goods, etc., by scrutinizing the NPOs/NGOs who already did that. If an NPO/NGO is classified as one of these, the driver will make sure they will ask for all the details before accepting the job.
- *Anticipating donors who are most likely going to default*: Donors who have defaulted before will be aimed at to generate an understanding of their aspects. Knowing that will help drivers to avoid accepting requests from donors with those aspects.
- *Projecting donors who will change the request's location after the driver arrives at the original location:* Drivers can guess whether a donor will change the location of pickup upon arrival by probing the details of donors who have done that before. To avoid wasting time, the designated driver will be in contact with that donor to ensure the address is not changed or to obtain the new location before making the selection.
- *Expectation of the actual size of goods*: Speculating the size of goods/resources to be picked up is important for drivers. Improper size might not fit in the selected vehicle. This will result in wasting driver's time. This could be a result of the NPO/NGO not recognizing the right size (entering a wrong size), or the donor changing the size upon arrival of the driver. Both the NPO/NGO and donor causing such problem need to have their features inspected. Once their features are identified, an NPO/NGO or a donor could be classified as possible providers of the wrong size. Hence, the driver can contact either one or both to circumvent wasting time and frustration.
- *Estimation of the actual distance*: It is possible for NPOs/NGOs to make mistakes regarding the actual distance or even the nearest estimate. A driver might accept a job thinking it is a short distance, but it turns out to be a long distance once they hit the road. To avoid such situations, drivers need to seek out those NPOs/NGOs who would possibly provide the inaccurate distance. Analyzing their traits will reveal which NPO/NGO need to be consulted regarding the distance before accepting the job.
- *Predicting the jobs/requests that require heavy lifting*: Some drivers might have back, shoulder, or neck problems. They are only allowed to lift certain maximum weight. Such drivers realized that goods/resources are heavier than what they can manage when arriving at the pickup location. Dissecting the NPOs/NGOs that made drivers go through such situations helps to conclude their features. This will help drivers to elude NPOs/NGOs with similar features.

D. Donor

Donors are concerned about reliability of vehicle and driver. In addition, the facilitating of picking up the goods is a priority for them.

- *Predicting drivers that are not reliable*: Based on the behaviors of drivers who either default, or do not arrive within the estimated time, donors can predict if a driver will not show up or arrive late. They then make their decision regarding the driver using the results of this analysis.
- *Foretelling NPO/NGO who most likely will delay picking up the donations for some time*: Donors are eager to have their donation of goods/resources be taken care off as quickly as possible. Some NPOs/NGOs might delay assigning a driver to pick up these goods/resources. By studying the peculiarities of such NPOs/NGOs, donors can guesstimate which NPOs/NGOs are most likely to reflect such a behavior.
- *Anticipating NPO/NGO who will most likely default on providing a receipt*: Receipts of donations estimated monetary value are important for donors for taxing purposes. Delaying those receipts or not providing them will upset donors. Donors will not be interested in dealing with such NPOs/NGOs in the future. Conjecturing the individualities of such NPO/NGO will help drivers to avoid dealing with those with similar characteristics.
- *Predicting NPOs/NGOs that might send the wrong type of vehicle for the size of the donated goods*: Selecting the wrong vehicle for picking up goods/resources will be annoying for both drivers and donors and a waste of time. Donors need to have their goods moved as soon as possible. Assessment of traits will help identifying those who would possibly behave in a similar fashion.

VI. CONCLUSION

With the notion of big data analytics as the practice of exploring huge and diverse datasets with sophisticated analytic approaches and methods to reveal hidden patterns, unfamiliar correlations, and other valuable knowledge to make abreast decisions, this paper contributed by analyzing the possible potential future analytics of the GoodTurn system to allow the stakeholders; Ford, NPOs/NGOs, drivers, and donors, to make informed decisions that improve their way of doing things and save them time and money. Various potential predictions, classifications, and clustering are suggested based on the future GoodTurn big data when matured. As soon as the big data become fully established, big data analytics tools will be adopted to get the desirable knowledge and insights for decision making.

REFERENCES

[1] W. Raghupathi and V. Raghupathi, “Big Data Analytics in Healthcare: Promise and Potential,” Health Information Science and Systems, vol. 2, no. 1, pp. 110-119, 2014.

[2] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, “Efficient Machine Learning for Big Data: A Review,” Big Data Research, vol. 2, no. 3, pp. 87-93, 2015.

[3] H. C. Chen, R. H. Chiang, and V. C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact,” MIS Quartely, vol. 36, no. 4, pp. 1165-1188, 2012.

[4] A. L'Heureux, K. Golinger, H. F. Elyamany, and M. A. M. Capretz, “Machine Learning with Big Data: Challenges and Approaches,” IEEE Access, vol. 5, pp. 7776-7797, 2017.

[5] D. T. Nguyen and J. E. Jung, “Real-Time Event Detection for Online Behavioral Analysis of Big Social Data,” Future Generation Computing Systems, vol. 66, pp. 137-145, 2016.

[6] P. Campbell, “Editorial on Special Issue on Big Data: Community Cleverness Required,” Nature, vol. 455, no. 7209, pp. 1-1, 2008.

[7] S. Choi, J. Seo, M. Kim, S. Kang, and S. Han, “Chorological Big Data Curation: A Study on the Enhanced Information Retrieval System,” IEEE Access, vol. 5, pp. 11269-11277, 2017.

[8] G. Bello-Orgaz, J. J. Jung, and D. Camacho, “Social Big Data: Recent Achievements and New Challenges,” Information Fusion, vol. 28, pp. 45-59, 2016.

[9] J. Lin and D. Ryaboy, “Scaling Big Data Mining Infrastructure: The Twitter Experience,” SIGKDD Explorations, vol. 14, no. 2, pp. 6-19, 2103.

[10] W. Fan and W. Bifet, "Mining Big Data: Current Status and Forecast to the Future," SIGKDD Exploration, vol. 14, no. 2, pp. 1-5, 2013.

[11] X. Amatriain, “Mining Large Streams of User Data for Personlized Recommendations,” SIGKDD Explorations, vol. 14, no. 2, pp. 37-48, 2012.

[12] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, “Protection of Big Data Privacy,” IEEE Access, vol. 4, pp. 1821-1834, 2016.

[13] B. Matturdi, X. Zhou, S. Li, and F. Lin, “Big Data Security and Privacy: A review,” China Communications, vol. 11, no. 14, pp. 135-145, 2014.

[14] E. Žunić, A. Djedović, and D. Đonko, “Application of Big Data and Text Mining Methods and Technologies in Modern Business Analyzing Social Networks Data about Traffic Tracking,” in Proc. The 2016 XI International Symposium on Telecommunications (BIHTEL), Sarajevo, Bosnia-Herzegovina, 2016, pp.1-6.

[15] I. Chih-lin, L. Yunlu, S. Han, W. Sihai, and G. Liu, “On Big Data Analytics for Greener and Softer RAN,” IEEE Access, vol. 3, pp. 3068-3075, 2015.

[16] J. Liu, F. Liu, and N. Ansari, ``Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," IEEE Network., vol. 28, no. 4, pp. 32-39, 2014.

[17] K. Snyder and K. Daimi, “The Application of Software Engineering to Moving Goods Mobile App,” in Proc. The 2017 International Conference on Software Engineering Research and Practice (SERP’ 17), Las Vegas, USA, 2017, pp. 88-97.

[18] K. Snyder and K. Daimi, “Securing Ford Mobility System,” in Proc. The Thirteenth International Conference on Networking and Services (ICNS 2017), Barcelona, Spain, 2017, pp. 1-7.

[19] Tax Excempt World, “NonProfit Organizations in Michigan by County,” https://www.taxexemptworld.com/organizations/michigan-counties.asp, September, 2017, [retrieved: March, 2018].

[20] Ford Motors on Forbes Global 2000 Lists, “https://www.forbes.com/companies/ford-motor/,” 2017, [retrieved: March, 2018].

[21] Wikipedia, “Ford Motor Company – Number of Employees, https://en.wikipedia.org/wiki/Ford_Motor_Company, [retrieved: March, 2018].

[22] Detroit Free Press, “Michigan's population increased for the fifth straight year in 2016,” http://www.freep.com/story/news/local/michigan/2016/12/20/michigan-population-increase-census/95631726/, [retrieved: March, 2018].

[23] The Nonprofit Sector in Michigan, “Economic Impact of Michigan’s Nonprofit Sector,” https://www.independentsector.org/wp-content/uploads/2016/12/Michigan-1.pdf, [retrieved: March, 2018].

[24] Center on Nonprofits and Philanthropy, “The Nonprofit Sector in Brief 2015,” October 2015, https://www.urban.org/sites/default/files/publication/72536/2000497-The-Nonprofit-Sector-in-Brief-2015-Public-Charities-Giving-and-Volunteering.pdf, [retrieved: March, 2018].

A Web-based Collaborative Tool for Land Use Computation on Satellite Images

Maria Grazia Albanesi, Riccardo Amadeo
Dept. of Electrical, Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
e-mail: mariagrazia.albanesi@unipv.it, riccardo.amadeo@gmail.com

***Abstract*— In this paper, we present the prototype of a new tool for computing land use on satellite images. Its main feature is the possibility of being used in a collaborative Web-based environment, where multiple expert users cooperate in validating the territory classification of different areas, according to the presence or absence of anthropic activity. For land use estimation, we use a well-known indicator in literature, namely the Anthropentropy Factor (AF). The novelty of the approach is the use of open software libraries, based on Keyhole Markup Language (KML), and Google Application Programming Interfaces (API) for the AF computation; the conceptual approach has been compared to other previously implemented solutions, namely the use of proprietary, ad hoc software tool and the use of Geographic Information System software on European Corine Land Cover data sets. A preliminary result of the work in progress is presented and commented.**

Keywords- Land use; satellite images; anthropentropy factor; Web-based application.

I. INTRODUCTION

The present paper describes a work in progress of a research project related to a great critical issue in environment preservation: land use estimation. Preserving lands, i.e., the wild, natural territory from the negative consequences of an inappropriate, out of control urban expansion, is one of the most important target for biodiversity preservation [1]. The percentage of land that annually is subtracted to natural, wild eco-systems for human activities expansion, has been estimated in the 36 European countries about 112000 ha/year in the period 2000-2006, reaching the percentage of 9% in the most urbanized countries [2].

Land use is not only due to urbanization, i.e., the expansion of rural and urban settlements, but also to the creation of industrial, intensive farming and touristic sites, roads and communication lines. We refer to all these activities as *anthropic activities*, and the consequent land occupation as *anthropic places*. Therefore, in the analysis of the territory we considered an anthropic area if it is occupied by signs of human presence, such as buildings, paved roads and railways, places of intensive agriculture, and industrial settlements, both for production and services.

Despite the great importance of land use estimation, its computation is far to be simple and fast: obtaining data about the territory and validating the area distribution of anthropic places is a long-time consuming process, even if the modern image satellite acquisition and processing have partially simplified the task. In fact, satellite images have to be processed in order to classify each pixel to anthropic places or not, and this is far to be completely automatized by low-level image processing techniques. For this reason, the project here described aims at giving a possible answer to the problem of a fast and easy-to-update process for validating the territory occupation and computing the land use indicator.

The paper is organized as follows: in Section II, a discussion about the choice of land use indicator is given. In Section III, a synthetic definition of AF helps the reader to understand the computational steps realized in the Web-based prototype. In Section IV, a discussion about related works on AF computation is given, in order to distinguish the novelties of the Web-based collaborative solution. In Section V, the details of the prototype, the computer technologies (languages and tools), and data formats are discussed. Conclusion and hints about future work end the paper.

II. THE CHOICE OF LAND USE INDICATOR

In this research, we choose the indicator of land use named *Anthropentropy Factor* (AF in the following) [3]; this neologism is derived from the Greek term *Anthropos* (*Ἀνθρωπος*) = man, and *entropy* and express, in a quantitative way, the “disorder” introduced in a natural, wild eco-system by the presence of human beings and their related anthropic activities. The peculiarity of this indicator is that it expresses the anthropic impact on land use not only by computing the simple percentage of soil occupied by human activities and urban expansion, but it takes into consideration also the *shape* of the anthropic areas subtracted to nature. In this way, the indicator gives also information about another aspect for land use, i.e., land fragmentation, which is considered by the UN Convention on Biological Diversity [4] as the major threat on species biodiversity preservation, because it limits the wandering and spreading of animals. In fact, after the first step of territory analysis for defining the anthropic areas, information about their shapes and contiguity is taken into consideration by expanding their boundaries in the two dimensions and, consequently, by increasing their size. The extension has two main purposes: first of all, it takes into account the negative effects due to noise and pollution close to the boundaries of the anthropic places. Secondly, if anthropic areas are sufficiently close each other, the enlargement causes an effect of filling small holes, which produces wider areas to be considered in the final computation of the land use indicator. At time of writing, the AF is the only indicator in literature which take into account the fragmentation of an anthropic place instead only the simple numeric value of its area.

III. THE ANTHROPENTROPY FACTOR

In this Section, we recall briefly the AF definition, in order to understand how it is computed in the core of the Web based collaborative tool here described. The procedure of the computation of the *AF* consists of the following five steps:

1) Let us consider a generic geographic region bounded by recognized borders; in this project, and in all the previous research activities, we considered as target territory the municipality, because in Italy the municipality is the administrative body in charge of deciding policies for land destination and preservation. Let define as S the area (in squared kilometres) of a target municipality under investigation.

2) Within the target municipality, the satellite image of this territory is analyzed to classify each pixel if belonging to either an anthropic area or to a wild natural area. This step is performed in a semiautomatic way by analysing the satellite images. After an initial automatic pre-processing for boundaries delimitation and road extraction, the classification is performed and validated by human experts.

3) We define a neutral sub-region (Neutral Zone) as the part of target territory containing at inland water (lakes, rivers) extending more than two squared kilometres and/or areas located more than 3,000 m above sea level. Let define NA as the area (in squared kilometres) of the Neutral Zone.

4) Each area occupied by anthropic places is enlarged along its boundaries with a buffer of 50 meters. The reason of this numerical choice is fully discussed in our previous works [3,5] and is here omitted for brevity. The enlargement is conceptually equivalent to the morphological image processing operation of dilation [6] with a circle of radius of 50 meters. We define the union of all the anthropic enlarged areas as Death Zone of the region, i.e., the zone where natural wildness is completely lost (dead) for the human anthropic influence. Let define DA as the area (in squared kilometres) of the Death Zone.

5) We define the Anthropentropy Factor *AF* as the ratio:

$$AF = DA / (S - NA) \qquad (1)$$

The *AF* is a real number in the range [0-1]. The higher is AF, the more critical is the situation for what concern land use and environmental preservation of wild ecosystems. For sake of completeness, the special case of *NA* = *S* is not considered, as it would mean that the entire target territory is occupied by water or it is located above 3,000 m above the sea, thus it is not suitable to land use and the computation of FA becomes meaningless.

IV. EXISTING APPROACHES AND NOVELTY OF WEB-BASED SOLUTION

At time of writing, no related works are present in literature about Web-based approaches for AF computation. Moreover, contributions are presents for other interesting computations about land use, but they refer to particular aspects of land data analysis and processing, e.g., hydrological change impact assessment [7], or the issue of sharing and integrating different geo-analysis models across an open web environment [8]. For this reason, in order to appreciate the novelties of the work in progress here presented, we chose to compare it to the different approaches used in our previous research activities involving AF indicator [3,5]. The relation between our current work and the previous ones is that the main goal of the research in [3,5] was to define the new indicator and to prove its efficacy in expressing the real situation of land exploitation; on the other hand, in the present paper we focus our attention on some critical issues of AF computation and we propose, as a possible solution, a Web-based collaborative framework on Google Earth satellite maps.

The AF indicator was first proposed in the Italian National project called ACI project (*Antropentropia Comuni Italiani*, i.e., Anthropentropy of Italian Municipalities) [3]. The initial, ambitious goal was to map the entire complete Italian territory in such a way that, for each of the 8092 Italian municipalities, the AF is computed according to (1). However, this goal has been disregarded, as the mapping of land use was possible only for seven of the twenty regions of Italy. In fact, the main problem is data availability, i.e., a suitable description of the territory in order to determine the position and extension of the anthropic places. At this purpose, the ACI project adopted two possible approaches. The first one refers to the data-set obtained by the Corine Land Cover (CLC) project [9]; the definition of CLC data-set started in 1990 and periodically, data are updated. Unfortunately, this approach has two big critical issues: first, Corine data set was not available for all the Italian territory, but only for 7 regions (over a total of 20 regions), namely for the 40,9% of municipalities (3311 over 8092). The second critical issue of this approach is the slowness of data update. As the validation of CLC data is a time-consuming process, the CLC data-sets refer to a description of the land use of several years before. For example, in 2018 the most recent CLC data-set of Italian territory refers to year 2012. The CLC data-set have been processed using standard Geographic Information System (GIS) software to implement the procedure to compute the values of AF according to (1). We called this first solution AF Computation based on *CLC/GIS approach*.

The second approach uses open data (Google Earth maps) and ask to users of a social network to generate the images of Death and Neutral Zones of a territory of a municipality under investigation. The images has been collected at the Computational Sustainability Unit at the Department of Electrical, Computer and Biomedical Engineering of Pavia University, were a software based on Matlab framework has been developed for AF computation. We have called this approach AF computation based on User Generated Content and open data (*UGC/Open data*). The main drawback of this approach is, as every solution based on crowdsourcing, the fact that its success is related to the degree of participation of the user community.

Unfortunately, this was not confirmed in the course of the project, and with this approach only the small percentage of 0,5% of the Italian municipalities was covered.

The novelty of the project here reported is to combine the strength of the two previously described approaches and to overcome their main drawbacks. In fact, the project is based on open, very fast updated data of Google Earth, but their processing is performed in a Web environment by trusted users, i.e., experts able to use a Web based framework to run the software for Death Zone generation and AF computation. Moreover, this software has been completely rewritten in java, and the dependency on proprietary suite of Matlab has been abandoned. Therefore, this solution seems to overcome the main drawback of the first CLC/GIS approach, i.e., the slowness of data update, as Maps in Google Earth are updated annually, as well of the second approach (UGC/Open data), because of the adoption of open source software used by a limited community of experts allow a very fast AF computation, whose efficacy is not related to a social community participation.

The present project uses the Web platform to share data and open software procedures in order to settle a collaborative environment for AF computation. Experts access maps of the territory of all the Italian Municipalities and can share intermediate results or work in collaborative form. For example, more than one expert user can analyse the territory of the same municipality to classify the anthropic areas and apply the dilation operation to generate the boundaries of the Death Zone.

V. THE WEB-BASED COLLABORATIVE PROTOTYPE

In this work in progress, we present here the general framework of the system and a preliminary result of AF computation on an entire municipality.

A. Technologies and tools

Several existing technologies and data sources are combined to realize the Web-based tool for AF computation on Google Earth images. First of all, we use an archive of shape files, downloaded from the official Italian National Institute of Statistics, to code territory boundaries information of all the 8092 Italian municipalities [10].

Then we used the free and open source QGIS software [11] to convert shape files into KML files. KML is an XML-based language for the management of three-dimensional geospatial data and it is used in several popular software, such as Google Earth and Google Maps. Successively, we developed Java code to subdivide the original KML file in multiple files, one for each municipality, in order to load them separately, by the Web tool, in a typical Google map interface. After these preliminary steps of converting and importing files in the proper environment, the software development used Google Maps API's to add drawing layers on a simple map object using the Drawing tools and the library Turfjs [12], which makes available JavaScript functions for advanced geospatial analysis in browsers. The code refers also to the GeoJSON open standard format, designed for representing simple geographical features, along with their non-spatial attributes.

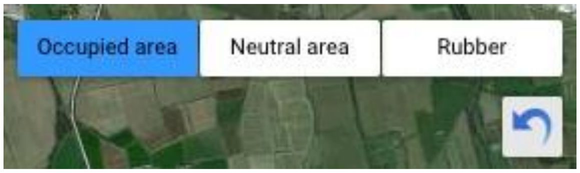


Figure 1. Graphical controls for AF computation.

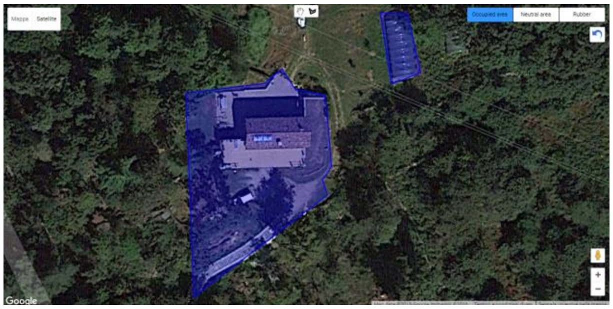


Figure 2. Two anthropic areas (in blue) inside a forest.

By combining graphical primitives with image processing functions, we coded and implemented the Web-based prototype for AF computation.

B. Funcionality and preliminary results

First of all, expert and trusted users are authorized to access the data of the projects by a simple login/password procedure in a browser. The user can choose the municipality for which the AF indicator has to be computed, by starting a new project on its territory, or edit an existing one. The projects can be public and shared with other users.

The graphical tools implemented in the project allow to draw the boundaries of the anthropic areas in the target territory, or edit, cancel or modify them. Anthropic areas are automatically coloured by the graphical tools in blue. Neutral areas are coloured in yellow. By using Google Maps API's we created drawing layers on a simple map object by embedding drawing tools for inserting circles, polygons, polylines, rectangles and by defining figures using WGS84 coordinates, directly attaching them on the map. Moreover, the developed software provides the definition of several polygon attributes, such as fill colour, border colour, and opacity. The information (coordinates, boundaries) related to each anthropic area is added to the map object in a custom data layer. Four graphical controls (see Figure 1) are added to Google Earth maps: *Occupied area* (polygon denoting urbanized, anthropic areas) *Neutral area* (polygon denoting stretches of water, such as lakes, lagoons, whose area is higher than 2 square km or lands with an altitude higher than 3000 m), *Rubber* (polygon that clears the inside area) and Undo (to undo the last graphical operation on the map). Figure 2 shows an example of drawing two polygons, in order to define two anthropic areas of houses inside a forest.

Other implemented functions manage polygon operation, such as intersection, sum, subtraction and expansion. In fact, anthropic areas have to be checked to verify if they do not exit the boundaries of the municipality: in this case the polygon area is reduced to fit the boundaries.

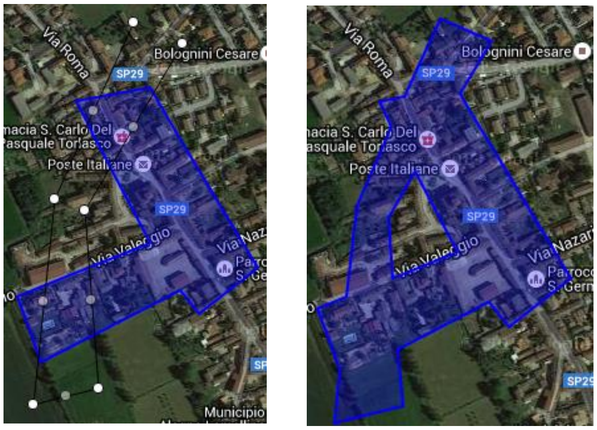


Figure 3. Drawing overlapping polygons. On the left: an exisitng polygon (in bue) and a new area, delimited by its corners (white dots). On the right: the result of polygon union.

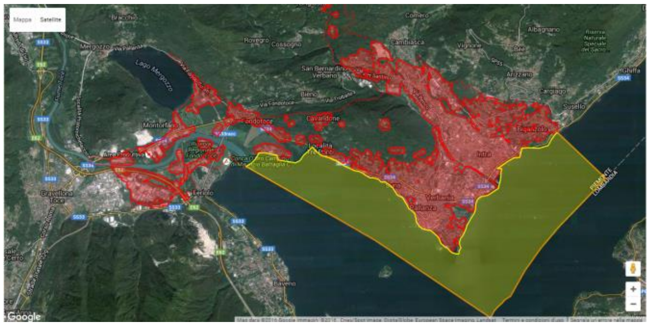


Figure 4. AF computation on Google Earth map for the municipaliaty of Verbania, Italy.

Moreover, whenever the user add a new polygon, the software verifies if it intersects other polygons already present in the map. If it does, the two polygons are substituted by their union (see Figure 3). This step may last several days and intermediate results may be saved on a server. In a successive session, the analysis may be restarted at the point of the last save operation, by the same user or by other users, in a full collaborative and shared manner. Once the territory has been completely analyzed, the blue areas are ready for the dilation and the definition of the Death Zone. The Java routines enlarge the blue areas according to the dilation operation, in order to generate automatically the Death Zone. Moreover, the software reads the values of S from the internal database and computes the values of DA, NA for the final computation of AF indicator, according to (1). This value can be saved in the project for a further check or saved in the internal database, to associate to the municipality under investigation the value of AF indicator and the year the map refers to. In Figure 4, the first example of the map analysis and computation of the land use indicator on an entire municipality is shown: it refers to the municipality of Verbania (North Italy, Lat. 45°55′16″ N, Long. 8°33′06″ E). The yellow area refer to the lake Lago Maggiore, the red area to the computed Death Zone after dilation. For this municipality, the AF computation is equal to 0.5654, showing a worrying situation: more than half of the territory has been completely anthropized.

IV. Conclusion and Future Work

As pointed out at the beginning of this paper, this is the description of a work in progress. The first results are very encouraging because the collaborative Web tool has proved to be an agile, simple and efficient framework to easily analyze up-to-date maps of the Italian territory and to compute in a fully automatic way the AF indicator.

The project is being under development to validate the tool on a significant number of municipalities. Once we populated our database of a certain number of classified maps, future work will be related to the use of these data to train a machine learning algorithm to automatically classify unknown maps, without the human supervision.

References

[1] LUCAS: Land Use and Land Cover survey, European Commission on Environment, http://ec.europa.eu/eurostat/statistics-explained/index.php/LUCAS_-_Land_use_and_land_cover_survey [retrieved: March, 2018].

[2] New guidelines to reduce soil sealing, European Commission Fact Sheet of 2017, http://ec.europa.eu/environment/soil/sealing_guidelines.htm, [retrieved: March, 2018].

[3] M. G. Albanesi and R. Albanesi, “A New Approach Based on Computer Vision and Collaborative Social Networking for Environmental Preservation: Theory, Tools and Results of Italian ACI Project”, Proceedings of The Eighth International Multi-Conference on Computing in the Global Information Technology, Nice (France), 21-26 July 2013, pp. 16-21, Copyright (©) IARIA, 2013

[4] Fragmentation of ecosystems and habitats by transport infrastructure, http://www.eea.europa.eu/data-and-maps/indicators/fragmentation-of-land-and-forests/eu-ac-fragmentation, [retrieved: March, 2018].

[5] M. G. Albanesi amd R. Albanesi, “A Decision-making Support System for Land Use Estimation Based on a New Anthropentropy Predictive Model for Environmental Preservation – Theory, Model and Web–based Implementation”, International Journal On Advances in Intelligent Systems, v 7 n 1&2, pp. 85-102, ISBN-1942-2679, Copyright (©) IARIA, 2014.

[6] R. C. Gonzales and R. E. Woods, Digital Image Processing, 3d efition, Pearson Prentice Hall, 2008, Chapter 9, “Morphological Image Processing”.

[7] J.-Y. Choi, B. A. Engel, L. Theller, J. Harbor, “Utilizing Web-based GISand SDSS for Hydrological Land Use Change Impact Assessment”, Trans. of the ASAE. Vol. 48(2), pp.815-822 , 2005.

[8] S.Yue, M. Chen, Y. Wen, G. Lu, “Service-oriented model-encapsulation strategy for sharing and integrating heterogeneous geo-analysis models in an open web environment”, ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier, vol. 114, pp. 258-273, April 2016.

[9] Commission of the European Communities: Corine Land Cover Project, http://www.eea.europa.eu/publications/COR0-landcover [retrieved: March, 2018].

[10] ISTAT On Line Municipality Database: https://www.istat.it/it/archivio/6789 [retrieved: March, 2018].

[11] QGIS software: http://www.qgis.org/, [retrieved: March, 2018].

[12] Turfjs home page, http://turfjs.org/ [retrieved: March, 2018].

Assessing the Accuracy of Crowdsourced POI Names

Abdulelah A. Abuabat, Mohammed A. Aldosari, Hassan A. Karimi
School of Computing and Information
University of Pittsburgh, Pittsburgh, USA
e-mail:{aaa185, maa303, hkarimi} @ pitt.edu

***Abstract*—The purpose of this paper is to assess the accuracy of the Volunteered Geographic Information (VGI), specifically naming point of interests (POIs), in OpenStreetMap (OSM). For this, we compared, from a lexical perspective, the similarity of POI names in the OSM dataset with their corresponding names in a reference dataset. The overall similarity is 80.62% suggesting that POI names in OSM have potential to be an accurate and reliable source.**

Keywords- Crowdsourcing; Volunteered Geographic Information (VGI); text similarity analysis; point of interest (POI); OpenStreetMap (OSM).

I. INTRODUCTION

The debut of Web 2.0 has led to the emergence of many new applications, models, tools, and projects, among other things. One of these projects is crowdsourcing. Although the crowdsourced data occasionally is redundant and noisy [5], crowdsourced data often shows decent quality [4], flexibility [4], reliability [6], and economy [6]. One of the well-known crowdsourcing projects is OpenStreetMap (OSM), a popular map-based Volunteered Geographic Information (VGI) [16] project designed to crowdsource geospatial data to build a freely accessible world map. Contributors can easily create a new POI, modify an existing POI, or extract data from a region of their interest. They collaboratively contribute to OSM through the OSM official website, desktop or mobile based applications.

Generally, VGI-based projects are known to be effective since contributors are not restricted to add or edit data/information. Contributors can report a change that might occur faster than commercial geographical information providers. For instance, contributors could report a road closure that occurs due to a natural disaster, and the change would be reflected immediately [9]. Considering that contributors do not follow specific standards to contribute new data, it is imperative to pay attention to quality of VGI data [7] [8]. Of possible VGI data errors, those related to naming POIs are focused on this paper. The contribution of this paper is to check the reliability of POI names in OSM, as a representative collaborative mapping project. The remainder of this article is structured as follows. Section II discusses related work. Section III discusses the methods used to measure POI naming accuracy. The analysis performed and its results are discussed in Section IV. Section V concludes the work.

II. RELATED WORK

OSM is a collaborative mapping project where anyone can contribute geospatial data and it is intended to be widely available and used by others without any restrictions [8] [13]. As OSM has become popular and widely used, attentions have been paid to its data quality [3] [8] [14]. Assessing the quality of data/information collected by the crowd in OSM is of great importance to the products and services that are based on the OSM data and maps. The two approaches in assessing VGI data quality are quality measure and quality indicator. In the quality measure approach, VGI data are compared with a reference dataset. In the quality indicator approach, VGI data are evaluated through intrinsic methods. In these methods, VGI data quality is evaluated by means other than a reference dataset [2]. For instance, contributors' behaviors are analyzed to estimate the overall data quality. VGI data quality, in terms of basic data quality measures, such as completeness, attribute accuracy, and semantic accuracy, have been extensively studied. It is argued in [15] that in OSM and similar projects, attribute names may be highly inaccurate due to lack of standards and clear naming conventions. In [12], answering the question regarding the number of contributions needed to map an area accurately was focused. It was found out that five contributions would result in an acceptable level of positional accuracy. In a recent work in [3], the authors assessed a subset of the OSM dataset with a reference dataset by analyzing the POIs that have changed frequently in terms of their names and positions. In their work, they focused on only one POI type, subway station, since it has a frequent number of changes regarding its name and position. They proposed an approach for identifying whether two POIs are homologous based on three measures: position, name, and amenity type. They manually evaluated these points and found that the majority of them are similar; 328 out of 329 POIs correctly matched their corresponding OSM POIs. Different from this work, we consider all amenity types in an OSM dataset and evaluate their similarity in terms of POI names. Furthermore, our work also evaluates names as they are edited and reviewed by other contributors.

III. METHODOLOGY

In this section, we will give a general overview of OSM, the methods we followed in our analysis to assess the quality of OSM POI names, and the measure we used to calculate the similarity between pairs of corresponding names in the OSM dataset and the reference dataset.

A. Overview

In OSM, the data model is classified into three types: nodes, ways, or relations. Nodes represent POIs which are objects or entities, e.g., a school or a restaurant. Ways represent groups of interrelated nodes, such as a group of

POIs in a building. Relations represent the relationships between nodes, ways or other relations. Contributors provide information, to the best of their knowledge, about POIs. They create new POIs and add some information about each. Other contributors may update a POI content by correcting errors and/or adding further information. While creating or editing a POI, contributors may choose among agreed-upon information from which they can make a selection, such as amenity types. They may also include some other information about a POI, for example, name and address. Our work in this paper is focused on assessing the overall quality of POI names in OSM. Currently, there is a void in the literature about text verification methods to check the correctness of the written words. For instance, if a contributor writes "Universty of Pittsbrg" instead of "University of Pittsburgh", OSM allows the contributor to save the incorrect name. Moreover, contributors may follow different naming conventions while creating or editing a POI. For instance, a contributor may write a street name as "Fifth Ave.", while another contributor may edit it to "Fifth Avenue". As it is stated by [10], most OSM contributors are amateur and have diverse backgrounds, education, and cartographic knowledge. In addition to the OSM dataset that was extracted from the OSM world history file, we have obtained a reference dataset as a ground truth data. The reference dataset is provided by Placesdatabases.com, a commercial vendor of spatial data. As it is mentioned in [11], the ground truth data is also susceptible to errors, and the assumption that the ground truth data is fully reliable is not valid. For instance, ground truth data may be outdated or may not be updated regularly as new data is added to the map compared to the VGI data which may be updated as soon as new data comes in. Placesdatabases.com claims that their data is completely refreshed every three months.

B. Methods

In this work, we use the following three methods to measure the similarity between the POI names in the OSM dataset with their corresponding POI names in the reference dataset.

Method 1. In this method, the overall similarity between the POI names in the last version of the OSM dataset and their corresponding POI names in the reference dataset is measured. Since POIs in OSM are usually updated frequently through a set of revisions, we assume that the latest version contains the most accurate POI names. Contributors may update POI names as they recognize errors, and POI names may evolve over time to be accurate and reflect the real names. However, POI names may not be correct if contributors have different views as to which is the correct name of a POI.

Method 2. In this method, we measure the overall similarity between the POI names in the last version of OSM dataset and its earlier version and consider only those OSM POI names that perfectly match (100%) their corresponding POI names in the reference dataset. The objective is to analyze whether or not the OSM POI names have been edited and revised frequently.

Method 3. In this method, we measure the average percentage of edits needed for an OSM POI name to match perfectly (100%) its corresponding name in the reference dataset. The objective is to realize how many edits on average are needed for POI names in OSM to be accurate and perfectly match their corresponding POI names in the reference dataset.

C. Similarity Measure

String similarity analysis is considered a significant tool in different applications, such as text mining, text classification, document analysis and clustering, and information retrieval. Two strings can be similar semantically or lexically. String similarity measure can be divided into two main categories: term-based and character-based. Since our work is focused on similarity measure between pairs of POI names, we compare string pairs lexically by taking the character-based approach. We use the Levenshtein Distance Strings Metric algorithm, which is character-based and calculates the minimum number of single character edits, i.e., deletion, substitution, and insertion, for the comparison. Table I shows an example of this algorithm that is used to compare two strings.

To compare two POI names, we consider the location, represented as latitude and longitude, of each POI in the OSM dataset. Next, we search the selected OSM POI with the nearest two POIs in the reference dataset, using the Euclidean distance. After finding the nearest two POIs, we check the POI names, by using the Levenshtein Distance Strings Metric algorithm, to see which one has the highest names similarity.

Our approach of matching the POIs in the OSM dataset and the reference dataset may produce inaccurate results because of two main issues. First, the nearest POI in the reference dataset may not be the correct corresponding POI. This issue might occur due to location accuracy [1] [2]. Second, multiple POI locations may overlap, in other words, POIs inside a POI. For example, two POIs might overlap within the same boundary like McDonald's as a restaurant and Walmart as a supermarket, as in Figure 1. To address these issues, we set specific conditions to improve the matching quality.

TABLE I. AN EXAMPLE SHOWING THE RESULTS OF THE LEVENSHTEIN DISTANCE STRINGS METRIC ALGORITHM

1st String	2nd String	Similarity
University of Pittsburgh	University of Pittsburgh	100%
	University Pittsburgh	96%
	University of Pitt	86%
	Pittsburgh	59%
	School of Computing and Information	20%
	NA	0%

These conditions were derived by conducting an analysis in the city of Pittsburgh, as described below. We determined a distance threshold to reduce matching errors through an analysis where we checked the locations of the two nearest Starbucks branches. We found that they are approximately 400 meters away from each other. We used 400 meters as a threshold for the maximum distance between these two POIs. Thus, an OSM POI will not be incorrectly matched with a similar but not corresponding POI in the referenced dataset. For instance, one of the Subway branches, in Figure 2, may be incorrectly matched with the other branches in the reference dataset although their POI names may be similar. The reason why we did not consider a larger or smaller distance is because some places are very large, like a university campus or a shopping center, and some are very small like Starbucks. Therefore, by using a threshold like the one here, we can ensure that a large POI, which may contain other small POIs within its boundary, will be included in the process, see Figure 1. Additionally, in the matching process, we include both POIs so that the most similar POIs, in terms of names, are considered [3]. To address the second issue, which is POIs overlapping, we examine several names for the same POI, especially names with abbreviations, such as “Saint → St.”, “Fifth → 5th”, “Avenue → Ave.”, and state abbreviation “New York → NY”, to find a minimum similarity percentage. We found that 40% is reasonable as the minimum similarity percentage. Table II shows an example of this test.

TABLE II. SIMILARITY RESULTS FOR NAMES AND THEIR POTENTIAL EQUIVALENT NAMES.

1st String	2nd String	Similarity
Saint Louis	St. Louis	80%
Fifth Avenue Station	5th Ave.	43%
New York	NY	40%
Starbucks	Subway	27%
Walmart	McDonald's	24%

IV. DISCUSSION AND RESULTS

The number of POIs, which have names in the OSM dataset in Pennsylvania is 89207. Of these, 17136 POIs (19.2%) have 100% similarity with the reference dataset. In the next two sub-sections, we will discuss the results of each method and the obstacles we faced.

A. Results

By applying Method 1, we found 80.62% overall similarity between the POI names in the OSM dataset and the POI names in the POI names in the reference dataset. By applying Method 2, we found 98.74% match between the POI names in the latest version of OSM dataset and the earlier version of OSM dataset. This means that if the POI name in OSM is entered accurately the first time, there is a high probability that it will remain to be accurate and unchanged in subsequent versions. By applying Method 3, we found that after 3.9% of the number of edits, OSM POI names will match the corresponding names in the reference dataset correctly. For instance, if a POI name is edited 100 times, it is likely that the accurate name remains the same after the fourth edit. As we can see, the percentage of edits needed to ensure accurate POI names is relatively low. This means that if a POI name is accurate the first time it is entered into the OSM dataset, chances are low that it will be edited in subsequent versions. In other words, most often the contributors tend to enter the correct names of POIs in the first place.

B. Limitations

As the goal of this work is to assess the quality of OSM POI names by comparing them against the names in the reference dataset, there are several considerations, related to quality standards which are mentioned in [12], that are worth mentioning. For instance, contributors may follow different approaches to identify and specify the location (latitude, longitude) of a POI on a map where each approach may result in a different location. This issue might also be found in the reference dataset. For example, matching McDonald’s in Figure 1 would find a closest POI in the reference dataset

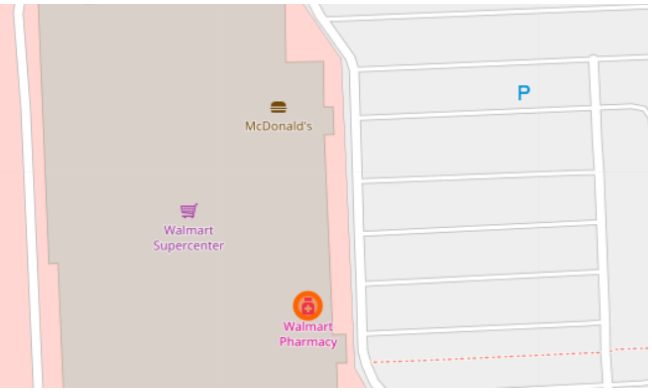


Figure 1. Example of POIs located inside a POI. Walmart Pharmacy and McDonald’s inside Walmart supermarket [17].

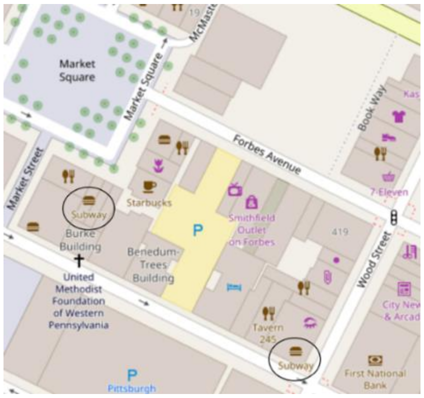


Figure 2. Example of same POIs located close to each other within a distance below the threshold [17].

where the distance between the locations in OSM and in the reference dataset is very small, thus considered overlapped; one scenario is that the McDonalds's in OSM is matched with the Walmart Supercenter in the reference dataset. In situations like this, the similarity threshold, discussed above, is used to preclude those comparisons where names are significantly different.

In addition to the issue of matching the OSM POI names with their corresponding names in the reference dataset, there is an issue of semantic similarity. Contributors may use different words or symbols interchangeably while they mean the same thing. For instance, a contributor may write a POI's name as "School of Computing & Information" instead of "School of Computing and Information". In such situations, our proposed approach of similarity measure may not produce 100% match, despite the fact that both names are semantically the same. One way to address this issue is by reminding the contributor of the common naming conventions used during the process of naming POIs. Also, in OSM we observed that contributors interchangeably write names in short forms, e.g., "5th Ave." instead of "Fifth Avenue". In such situations, while both names are semantically the same, the similarity percentage will be low. However, adhering to a naming convention is one way to address the semantic similarity issue, but there still remains the problem of different naming standards in different countries.

V. CONCLUSION

In this paper, we focused on assessing the accuracy of VGI in naming POIs. We implemented three methods to measure accuracy of POI names: the overall similarity between the OSM dataset and the reference dataset, the similarity between the last version and an earlier version of the OSM dataset, and the average number of edits needed to have OSM POI names to be 100% similar to their equivalent in the reference dataset. We focused on the lexical perspective of the names, rather than the semantic view of the names, and found that most POI names in OSM are accurate. This work introduces new research questions: How can the accuracy of POI names be improved? Can there be a unified style for naming POIs? Can there be an algorithm that helps contributors by suggesting names?

ACKNOWLEDGMENTS

While each author contributed equally, we would like to thank the School of Computing and Information at the University of Pittsburgh for the encouragements and the support.

REFERENCES

[1] W. H. Gomaa and A. A. Fahmy. "A survey of text similarity approaches," International Journal of Computer Applications, no. 13, pp.68, 2013.

[2] C. Barron, P. Neis, and A. Zipf, "A comprehensive framework for intrinsic OpenStreetMap quality analysis," Transactions in GIS, vol. 18, no. 6, pp. 877-895, 2014.

[3] G. Touya, V. Antoniou, A. Olteanu-Raimond, and M. Van Damme. "Assessing Crowdsourced POI Quality: Combining Methods Based on Reference Data, History, and Spatial Relations," ISPRS International Journal of Geo-Information 6, no. 3. pp. 80, 2017.

[4] M. Van Exel, E. Dias, and S. Fruijtier. "The impact of crowdsourcing on spatial data quality indicators," In Proceedings of the 6th GIScience international conference on geographic information science, pp. 213, 2010.

[5] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu. "Maximizing benefits from crowdsourced data," Computational and Mathematical Organization Theory 18, no. 3, pp. 257-279, 2012.

[6] O. Alonso, and S. Mizzaro. "Using crowdsourcing for TREC relevance assessment," Information Processing & Management 48, no. 6. pp. 1053-1066, 2012.

[7] A. J. Flanagin, and M. J. Metzger. "The credibility of volunteered geographic information," GeoJournal 72, no. 3-4, pp. 137-148, 2008.

[8] L. A. Ali, F. Schmid, R. Al-Salman, and T. Kauppinen. "Ambiguity and plausibility: managing classification quality in volunteered geographic information," In Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 143-152. ACM, 2014.

[9] M. Zook, M. Graham, T. Shelton, and S. Gorman. "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake," World Medical & Health Policy 2, no. 2, pp. 7-33, 2010.

[10] H. Senaratne, A. Mobasheri, A. L. Ali, C. Capineri, and M. Haklay. "A review of volunteered geographic information quality assessment methods," International Journal of Geographical Information Science 31, no. 1, pp. 139-167, 2017.

[11] D. Jonietz, and A. Zipf. "Defining fitness-for-use for crowdsourced points of interest (POI)," ISPRS International Journal of Geo-Information 5, no. 9, pp. 149, 2016.

[12] M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. "How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information," The Cartographic Journal 47, no. 4, pp. 315-322, 2010.

[13] S. S. Sehra, J. Singh, and H. S. Rai. "Assessing OpenStreetMap Data Using Intrinsic Quality Indicators: An Extension to the QGIS Processing Toolbox," Future Internet 9, no. 2, pp. 15, 2017.

[14] R. Karam and M. Melchiori. "A crowdsourcing-based framework for improving geo-spatial open data," In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, pp. 468-473 , 2013.

[15] J. Girres and G. Touya. "Quality assessment of the French OpenStreetMap dataset," Transactions in GIS 14, no. 4, pp. 435-459, 2010.

[16] M. F. Goodchild. "Citizens as sensors: the world of volunteered geography," GeoJournal 69, no. 4, pp. 211-221, 2007.

[17] OpenStreetMap. www.openstreetmap.org/. Accessed 20th November 2017.

Value Annotation of Web Resources: the ValueML Language

Massimo Romanin, Elio Toppano
Dipartimento di Scienze Matematiche, Informatiche e Fisiche (DMIF)
Università di Udine,
Udine, Italy
e-mail: elio.toppano@uniud.it

***Abstract*—In the multimedia design field, we have recently witnessed a shift of focus from products and the user' s experience to social effects of technologies and the quality of life. In this context, values play an important role. They may be inscribed within an artifact as symbolic meanings or as a built-in use consequence. In spite of their growing relevance, there is not yet a markup language for value annotation. This paper describes a proposal for filling this gap. After a brief review of various perspectives on the concept of value and relevant taxonomies, we discuss the syntax and semantics of a preliminary version of the ValueML language together with an example of annotation of a commercial video clip.**

Keywords-value; annotation; semantic web; markup languages.

I. INTRODUCTION

In spite of the growing relevance of values in information technology [1], computer systems [2] [3], human computer interaction [4], multimedia, and game design [5] [6], there is still no common interchange language for the analysis and annotation of web resources that deal with this kind of abstract constructs. This paper describes a proposal for filling this gap.

The need for value annotation of communicative artifacts is present in several application domains. In the fields of Marketing and Brand Communication, for instance, values constitute an important component of web sites, commercial videos and advergames. They may refer to the advertised product or service (i.e., a value proposition) or, more generally, to a company's brand identity (i.e., the brand core values) and brand world (i.e., the brand world ethos). Political parties, religious communities, as well as social activists focus on values as one of the fundamental content and theme of their messages in designing web sites and blogs. Sometimes, values are explicitly communicated. L'Oreal, for example, started out its campaign "Beauty for all" by explaining the deep values (e.g., passion, innovation, entrepeneurial spirit, open-mindedness, excellence, and responsibility) that were at the base of its messages with an explicit document published on the web [7]. Most often, values are implicitly inscribed within digital discourses - namely, written texts, visual advertisements, commercial videos, web sites, games - by an appropriate selection and composition of content (e.g., denotative, connotative, and narrative meanings) and expression (e.g., plastic features of visual and auditory signs). The project Values at Play (VAP), for example, is an initiative aimed at investigating the role of social, moral, and political values in digital games [8]. It builds on the premise that games, like other computer and information systems, may embody values in their architecture, interaction paradigms, and mechanics. In the same vein, Value Sensitive Design [9], Value Centered Design [10], Design for Subjective Well-Being [11], and Design for Sustainability [12] explore conceptualizations and methods for facilitating values conscious design, while Generative Semiotics studies values in narrative products [13]. In addition, values are an important component of cultures. Therefore, independently of explicit design intentions, values are inevitably inscribed within communicative artifacts as a reflection of the culture of their clients, designers and developers (Culture *in* Design). Alternatively, communicative artifacts can be intentionally designed to adapt to the culture of target users (Design *for* Culture). This is at the base of the localization of web resources, a challenging issue addressed by several approaches in the field of cross-cultural design. In all the above cases, it seems important to be able: i) to identify the values that are embedded in products (Which values?); ii) to associate values with design choices (How values are communicated?), and iii) to explicate the goals and intentions that are at the base of the selection of those particular values and their expression in the considered artifact (Why?). The annotation of a communicative artifact with its inscribed values could be exploited for several tasks ranging from resource filtering and retrieval, to content repurposing or reuse. Notice that assuming that computer systems and web resources express or embody values means assuming that they are not morally neutral and that it is possible to identify *tendencies* in them to promote or demote particular moral values, and norms [14]. Such tendencies are embedded in the sense that they can be identified and studied largely or wholly independently of actual uses of the artifact, although they manifest themselves in a variety of uses of the system (not necessarily in all uses!).

The paper is organized as follows. In Section II we discuss the concept of value from different perspectives and we illustrate available taxonomies and vocabularies of values. Next, in Section III we state the scope and the aim of the study by focusing on the values that are inscribed within an artifact during its development stage. The main requirements of a language for value annotation are then introduced in Section IV together with a possible solution, i.e., the ValueML. A simple example of analysis and annotation of a video commercial using the proposed language is presented in Section V to show the effectiveness of the approach. Finally, in Section VI some conclusions are reported.

II. State of the Art

A. The concept of value

The concept of value has several meanings according to the specific perspective from which it is considered [1] [15] [16] [17]. Looking at existing literature, the term "value" is interpreted as:

- an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end state of existence (i.e., value as enduring belief system);
- the monetary sacrifice people are willing to make for a product (i.e., value as exchange);
- the utility of the physical properties of the product, which is realized only upon its use (i.e., value as perceived utility);
- an indicator of how much one desires a product or fears of loosing it (i.e., value as attachment);
- sign or meaning, e.g., an index of social status, lifestyle, modernity (i.e., value as meaning);
- an indicator of how the interaction with a product is aesthetically, cognitively or affectively worth to be made (i.e., value as good experience).

In addition, there is the need to disambiguate among different concepts that are in some way correlated such as values, needs, desires, preferences, and goals (see for example [1]). For some scholars values are abstract, desirable trans-situational goals; for others they are relatively stable individual preferences that reflect socialization; yet others consider values as cognitive representation of needs. In his Value Theory [17], Schwartz, defines values as "desirable, trans-situational goals, varying in importance, that serve as guiding principles in people's lives". Most importantly, he identifies five main features of the conception of value that are implicit in the works of many theorists and researchers:

- values are beliefs tied inextricably to emotions;
- values are a motivational construct. They refer to desirable goals people strive to attain;
- values transcend specific actions and situations;
- values serve as standards or criteria to guide selection or evaluation of action, policies, people or events;
- values are ordered by importance relative to one another.

We argue that a clear understanding of the meaning of value is an important step toward the development of a *value ontology* for applications in the Semantic Web field.

B. Value taxonomies

Several efforts have been made, in the past, to classify values and propose appropriate (with respect to specific criteria) value taxonomies. Schwartz, for instance, proposed a set of ten basic values each one described in terms of its motivational goal. They are: self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence, and universalism [17]. These values are structured on a circular pattern where congruent values (e.g., achievement and power) are located on adjacent positions while conflicting values (e.g., achievement and benevolence) on opposite sides. Boztepe, focusing on user's values, proposed a classification including nineteen different values clustered into four main categories namely, utility, social significance, emotional, and spiritual [15]. Value Sensitive Design focuses on values in computer systems such as privacy, freedom from bias, informed consent, accountability, property rights, to name only a few [9]. Specific taxonomies have been proposed in marketing [18] and in game design [8]. Floch, for example, distinguishes four types of product values - namely, practical, critical, utopian, and ludic values - that are at the base of the main marketing strategies. Friedman et al. [9] uses a set of seventeen values (e.g., diversity, justice, inclusion, equality, environmentalism, creativity, trust, etc.) for the analysis and design of digital games. Recent research in Positive Design focuses on hedonic values (e.g., pleasure) and eudaimonia (e.g., personal flourishing) [11]. What emerges from the comparison of current literature is that available proposals are very different in terms of: i) number of values considered; ii) level of generality; iii) granularity of proposed distinctions, iv) types of values considered, and v) terms used to denote the values. Some values (e.g., autonomy, self actualization) are common to various approaches while others (e.g., informed consent, humor) are present only in some vocabularies. Some vocabularies are more heterogeneous than others merging general and specific values or values having different nature such as hedonic values (e.g., pleasure) with ethical (e.g., morality, virtue), political (e.g., justice), and cultural values (e.g., life quality, happiness). Moreover, not all approaches accurately define the concepts represented by their vocabularies; so there are ambiguities (i.e., multiple interpretations of values) and "semantic confusion" within and across vocabularies. What is needed is a conceptualization that integrates existing proposals (or part of them) into a coherent and comprehensive framework that could be used as a guiding framework for media content annotation. An interesting step toward this goal is the work by Brey reported in [5]. The author illustrates an articulation of axiology that provides structure and overview to relevant values belonging to traditional theories (e.g., ethics, aesthetics, and politics) including cultural values of Theories of Good applied to new media.

III. Scope and aim of the study

A. The values inscribed within a product

Our study is intended to focus on values that are (intentionally or unintentionally) inscribed within a communicative artifact during design and system implementation. We call these values the "Values *in* the product" to distinguish them from other kinds of values such as, for example, the values of the stakeholders (e.g., the client's, designer's or user's internal conceptions of what is worth/important in life), the economic value of the product (i.e., its exchange value), or the value of the product as experienced by a user during its use (i.e., the perceived use value). The latter two interpretations - namely the exchange and perceived use values - can be referred to as the "Value *of* the product". The values inscribed in a product are strictly connected to the choices taken by designers and developers during production since these decisions are made on the base of criteria that are usually value-laden. For communicative artifacts, these choices may refer to several aspects such as, for example, i) the adopted conceptual model or meta-model of the artifact; ii) the articulation of content meaning; iii) meaning presentation or expression, and iv) technology. Here are some examples. The adoption of the Semantic Markup for Web Services (OWL-S) instead of the Web Service Modeling Ontology (WSMO) reflects different values and has different ethical implications as discussed in [19]. In a narrative commercial clip, ethical or moral values can be inscribed in the story (content) or expression (e.g., aesthetic values); in a digital game they can be embodied in game mechanics (i.e., in game rules), game dynamics or experience (i.e., hedonic pleasure). In Persuasive Technologies [20], Design for Sustainable Behavior [12], and in applications inspired to Nudge Theory [21], values are directly related to the intended behavior or state we want to be enabled, induced or fostered in users. According to the above discussion, we can distinguish the following two main cases:

- an artifact may have embedded values understood as special kind of built-in consequences. This conception (i.e., *causalist conception* of embedded values) relates values to causal capacities of an artifact to affect the environment. In other words, the artifact use causes a state of the world that realizes some kind of value;

- an artifact may be expressive of values (i.e., *expressive conception* of embedded values) in that it contains symbolic meanings that refer to values. These values may represent the values of designers, clients or users. This does not imply that it also functions to realize these values. It is conceivable that the values expressed in artifacts cause people to adopt these values and thus contribute to their realization. Whether this happens or not remains an open question.

B. *The problem addressed: value annotation*

Generally speaking, value taxonomies and vocabularies can be exploited in three different use cases:

- manual annotation of multimodal resources with inscribed values;

- automatic value detection and classification. The goal, here, is to model the means/ends relationships existing between measurable features of multimodal artifacts and abstract constructs such as value concepts;

- value generation, that is, simulation of specific values by an appropriate selection and composition of multimedia content and expression.

Our study focuses on the first use case. The problem we intend to address is thus the following: to design a general-purpose language for the manual annotation of values inscribed within a multimodal resource. The language should let the annotator to define the scope of a value annotation and to describe the value itself by referring to a specific and shared vocabulary. We envisage several possible ways in which the annotation could be used including:

- retrieval and selection/filtering of resources or part of them on the base of intended embodied values. It may be possible, for example, to annotate specific fragments of a multimodal resource with intended values and then retrieve the fragments using the values as key words;

- reuse a resource for new goals or contexts (i.e., repurposing). The identification and value annotation of multimodal fragments enables a designer to reuse the content of the fragment for new goals/objectives or in new communication contexts;

- exploitation of design knowledge embodied within an artifact for new products. Linking values to fragments is a way to explicitly represent how values are communicated in *that* artifact. This knowledge is design knowledge that may be used as inspirational for innovative products and design solutions;

- construction of a shared data base of value annotation resources that can be used as a ground base or training set for automatic recognition of values or for scientific research.

IV. The ValueML

The following is an initial unstructured list of requirements for the development of a language for value annotation. They are based on an understanding of the needs arising from concrete scenarios of manual annotation of multimodal texts. The desired language should allow the annotator to represent:

- the *intended* values a designer/author wants to be embodied in the multimodal resource under development. The focus is thus on annotation during the design and the development of the communication message rather than during its final use. This is not to deny that annotations made by the users are important. Simply, social tagging comes after the product has been developed and published and has different goals. It may be used, for example, to assess the effectiveness of intended value communication;

- values that have been expressed by different semiotic modalities (e.g., written texts, static and dynamic images, sound objects) and dispersed across content (e.g., narratives, denotative and connotative meanings) and expression (e.g., plastic features). As an example, brand values and brand ethos could be communicated by a story while aesthetic values by the product look and feel;

- non-economic values. We are interested in moral, ethical, political, aesthetic, and cultural values rather than the economic value of the product;

- values that belong to different taxonomies. The language should be sufficiently flexible so that the annotator is not constrained to use a specific vocabulary but can select among a set of available vocabularies the one that better satisfies his/her goals in the design situation at hand. The design of the language must be modular so that the appropriate vocabulary of descriptors for the target use can be chosen;

- values at different aggregation levels. In other words, it should be possible annotate the entire resource, as well as specific fragments or parts of it. Moreover, it should be possible to trace the time course of values in dynamic products such as audio-visuals;

- values associated to specific components in narrative texts. For example, it should be possible to associate specific values to the characters of a narrative told by a video and track the evolution of these values during story events. Alternatively, it should be possible to associate a value to the events of the story, or the consequences of these events;

- the relevance of an intended value with respect to other values inscribed within the same product and a measure of confidence in the attribution of that value to that fragment (i.e., a confidence of annotation accuracy);

- complex values i.e., combination of simple values occurring simultaneously in a particular segment of a resource;

- how a value is communicated in the product that is the modality (e.g., by expressive or narrative features);

The following section describes the main features of a draft language proposal intended to satisfy most of the above design requirements.

A. *Language syntax and semantic*

ValueML uses a XML-based syntax. Structurally, ValueML uses elements and attribute names to indicate the type of information being represented; attribute values provide actual information. The proposed language adheres to the following syntactic principles:

- the value annotation is self-contained within a 'value' element;

- all values belong to a specific controlled vocabulary;

- the annotating data is a value label; it is explicit from which vocabulary the value label is chosen. We draw on existing literature to propose a set of value categories;

- the link to the annotated material (i.e., the target) is realized by a reference using a URI and the reference has an explicit role. Two roles have being proposed namely: 'expressedBy' and 'signifiedBy';

- the modality of value expression or signification is specified, e.g., by storyline content, sound objects or visual plastic features;

- the target of annotation (scope) may be a block of text, an image or part of it, a segment of an audio or video asset, a node of a XML document (e.g., a SMIL presentation);

- a set of contextual elements can be used to describe the type of resource, its name, its web URL, etc.

V. CASE STUDY

We illustrate an example of annotation of a specific multimodal resource: a commercial clip. The aim is to assess the feasibility of value annotation with ValueML in a concrete case.

A. *A test bed: the Citroen BX ad campaign*

The clip selected for the analysis and annotation represents a well known ad campaign by Citroen [22]. It was produced in 1982 to advertise the BX model. Our choice is motivated by the availability of original script and critical essays about the considered video that provide the main source of annotation knowledge [18].

Annotation is preceded by an analysis of the clip that is driven by a semiotic meta-model of the video as discussed in [23]. The meta-model distinguishes four interrelated levels of analysis (Figure 1): i) the textual level representing the concrete/physical manifestation of the video content in terms of audio-visual features; ii) the discourse level referring to thematic, figurative, rhetorical aspects; iii) a shallow narrative level describing the story told by the video in terms of abstract roles (called actants) and narrative schemas (e.g., the narrative canonical schema), and, iv) a deep narrative level that uses a specific tool called semiotic square to articulate deep semantic meanings such as narrative values (axiology). Signification unfolds by crossing these levels from shallow features of the video to the most abstract and deep ones.

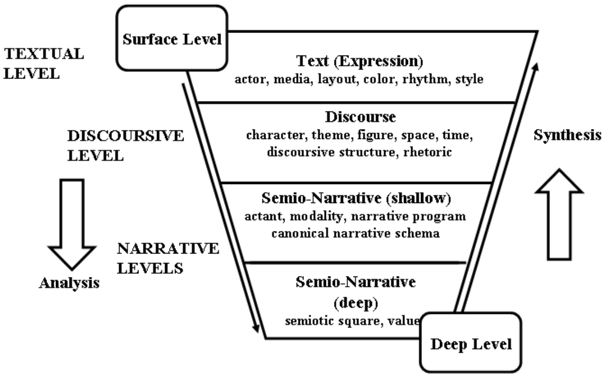

Surface Level
TEXTUAL LEVEL
Text (Expression)
actor, media, layout, color, rhythm, style
DISCOURSIVE LEVEL
Discourse
character, theme, figure, space, time, discoursive structure, rhetoric
Synthesis
Semio-Narrative (shallow)
actant, modality, narrative program canonical narrative schema
NARRATIVE LEVELS
Analysis
Semio-Narrative (deep)
semiotic square, value
Deep Level

Figure 1. A schematic view of the meta-model used for the analysis of the video commercial.

At the deep Semio-Narrative level, values have been classified according to Floch [18] into four classes namely, practical, critical, utopian and ludic values. They represent the vertices of a semiotic square. Practical values refer to utility, usefulness; critical values to convenience, performance, quality; utopian values to identity, reflection, social relations, and ludic values to surprise, madness, astonishment, irony and pleasure including aesthetic pleasure. The selection of specific values in the construction of a story allows the author to realize specific marketing strategies. In the considered video clip, values are communicated as follows:

- the first segment of the video (time interval: [0 s, 33 s], 14 shots) represents practical values (see Figure 2). A red car leaves Paris at midnight (Minuit, Paris ...) under the rain. After 8 hours it gets to the sea (.. 8 heures, la mer). The car is presented as a safe, confortable, and quick mean to escape from the everyday city life. Onboard a young lady, takes off her hat, smiling. An off screen voice (by Julien Clerc) sings: "J'aime, J'aime, J'aime";

- a following segment (time interval: [34 s, 40 s], 3 shots) is used to communicate ludic values (see Figure 3). The car suddenly dives in the sea without it could be possible to attribute this mad action to the driver that is never shown. The plunge, unexpected and irrational, represents the negation of practical values shown in the previous segment;

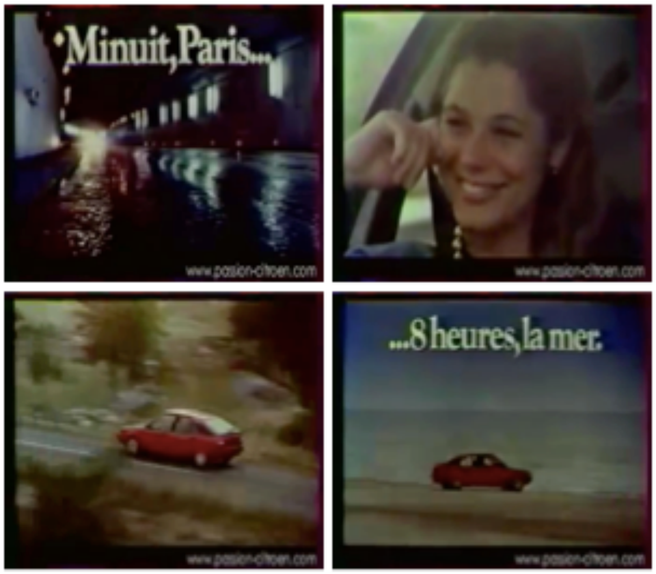


Figure 2. Four key frames of the first segment of the video clip representing practical values.

Figure 3. Two key frames of the second segment of the video clip representing ludic values.

Figure 4. A key frame of the final segment of the video clip representing utopian values.

- a final segment (time interval: [41 s, 47 s], one shot) represents utopian values (see Figure 4). Here, the car (Citroen BX) is no more an instrument, it is a subject, it lives (Citroen BX. Elle vit.).

Figure 5 shows the semiotic square with the trajectory of values expressed by the Citroen BX clip during presentation. Practical and ludic values are communicated through the visual track, while the utopian valorization is explicitly expressed by a voice over.

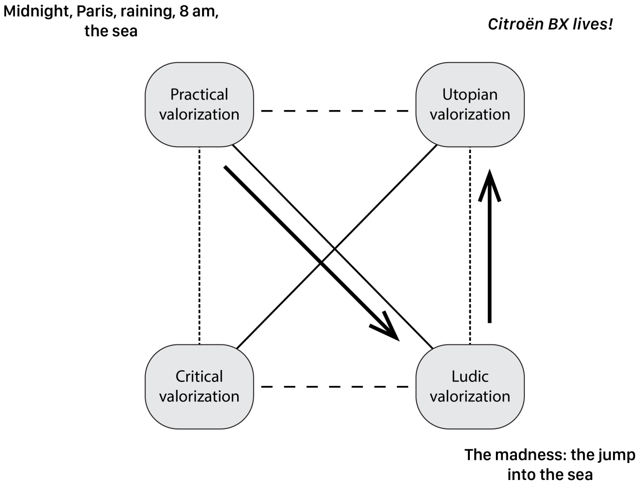
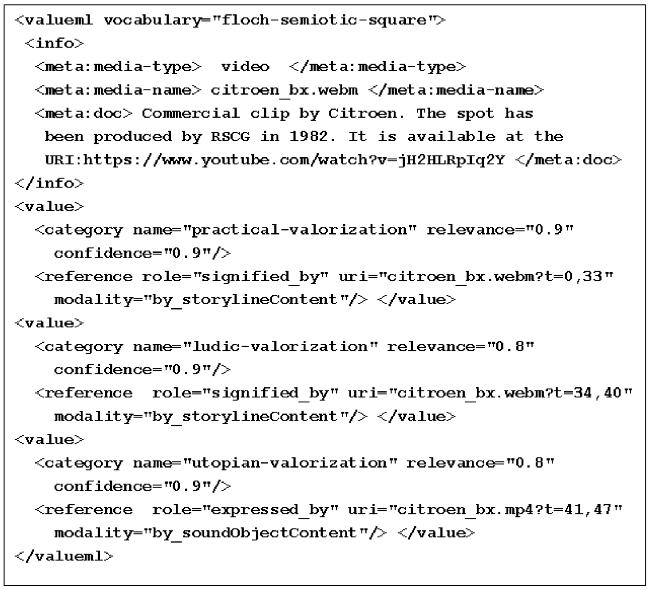


Figure 5. Semiotic square of consumer values

B. Value annotation

Figure 6 shows a simple example of annotation of the considered video clip using the ValueML.



```
<valueml vocabulary="floch-semiotic-square">
 <info>
  <meta:media-type>  video  </meta:media-type>
  <meta:media-name> citroen_bx.webm </meta:media-name>
  <meta:doc> Commercial clip by Citroen. The spot has
   been produced by RSCG in 1982. It is available at the
   URI:https://www.youtube.com/watch?v=jH2HLRpIq2Y </meta:doc>
</info>
<value>
  <category name="practical-valorization" relevance="0.9"
    confidence="0.9"/>
  <reference role="signified_by" uri="citroen_bx.webm?t=0,33"
    modality="by_storylineContent"/> </value>
<value>
  <category name="ludic-valorization" relevance="0.8"
    confidence="0.9"/>
  <reference  role="signified_by" uri="citroen_bx.webm?t=34,40"
    modality="by_storylineContent"/> </value>
<value>
  <category name="utopian-valorization" relevance="0.8"
    confidence="0.9"/>
  <reference  role="expressed_by" uri="citroen_bx.mp4?t=41,47"
    modality="by_soundObjectContent"/> </value>
</valueml>
```

Figure 6. Annotation of the video clip with ValueML.

The attribute 'vocabulary' specifies the set of values that are used for the annotation, i.e., the Floch's classification. Within the 'info' element various metadata are present to describe several contextual information such as resource type, name, and description. Value annotation starts with the 'value' element. 'Relevance' and 'confidence' are specified by a continuous unitless scale such as [0,1]. The expression of a value may be masked by another one, it may be inhibited, minimized or even exaggerated. Therefore, the human annotator needs to indicate the degree of importance and confidence that a certain attribution is correct. The 'role' and 'uri' attributes within the 'reference' element are used to associate values to video segments and to describe the type of relationship existing between annotated and annotating data. The 'modality' attribute specifies the features of the video that are charged with value meaning thus realizing a means/ends chain.

VI. Conclusions

To our knowledge, this is the first attempt to propose a language for value annotation of web resources. Analogous projects exist in the field of Affective Computing and Sentiment Analysis [24]. We refer, in particular, to the recent W3C initiative of EmotionML for the annotation of expressed emotions [25]. ValueML is inspired to such an effort; it may be seen as a complementary resource for describing experiential and socio-cultural aspects of artifacts.

Current experiments have confirmed the feasibility of the approach and the effectiveness of this preliminary version of the language. However, more analyses are required, before addressing formalization, in order to identify limitations and possible improvements. As multimedia designers, we expect to gain knowledge about how values can be, and are actually embodied in artifacts. A goal is, for example, to better understand the relationship existing between the axiological level and the narrative one. Which kind of properties should be possessed by the main components of a narrative (e.g., space, time, characters, relationships, passions, events) in order to effectively communicate a given set of values? This knowledge could be used to enrich the domain of possible values associated to the 'modality' attribute. Another open issue regards the specification of scales. Should they be continuous or discrete, unipolar or bipolar, etc. We have postponed a more detailed specification of scales after the acquisition of more knowledge. The final step will be the design and implementation of an ontology to define the terms of the ValueML language, to relate the terms to one another, and to define mappings between value vocabularies when possible.

References

[1] S. Kujala and K. V.V.Mattila, "Value of Information Systems and Products: Understanding the Users' Perspective and Values," Journal of Information Technology Theory and Application, 9:4, pp. 23-39, 2009.

[2] B. Friedman and H. Nissembaum, "Bias in computer systems," ACM Trans. on Information Systems, Vol. 14, No. 3, pp. 330-347, 1996.

[3] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: mapping the debate," Big Data & Society, July-December, Sage, pp. 1-21, 2016.

[4] B. Friedman and P. H. Kahn, "Human values, Ethics, and Design, " in Handbook of Human-Computer Interaction, J. Jacko, A. Sears, and N.J. Mahwah, Editors, Lawrence Erlbaum, pp. 1177-1201, 2003.

[5] P. Brey, "Theorizing the cultural quality of new media," Techné, 11:1, Fall 2007, pp. 2-18. URL: http://scholar.lib.vt.edu/ejournals/SPT/v11n1/brey.html [retrieved: April, 2018].

[6] M. Flanagan, J. Belman, H. Nissenbaum, and J. Diamond, "A method for discovering values in digital games," Proc. DiGRA Conference, pp. 752-760, 2007.

[7] L'Oreal Web Site URL: http://www.loreal.com/group/who-we-are/our-values-and-ethical-principles [retrieved: April, 2018].

[8] J. Belman, H. Nissenbaum, M. Flanagan, and J. Diamond, "Grow-A-Game: a tool for values conscious design and analysis of digital games," Proc. of DiGRA Conference, Vol. 6, pp. 14-17, 2011.

[9] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in Early engagement and new technologies: opening up the laboratory, pp. 55-95, Springer, Netherlands, 2013.

[10] G. Cockton, "A development framework for value-centered design," Proc. CHI EA '05, pp. 1292-1295, 2005.

[11] P. M. A. Desmet and A. E. Pohlmeyer, "Positive Design: an introduction to Design for Subjective Well-Being," International Journal of Design, Vol. 7, No. 3, pp. 5-19, 2013.

[12] F. Ceschin and I. Gaziulusoy, "Evolution of design for sustainability: from product design to design for system innovation and transition," Design Studies, 47, pp. 118-163, 2016.

[13] C. Bianchi, "Semiotic approaches to advertising texts and strategies: narrative, passion, marketing," Semiotica 183, 1/4, pp. 243-271, 2011.

[14] P. Brey, "Values in technology and disclosive computer ethics," The Cambridge Handbook of Information and Computer Ethics, Ed. L. Floridi, Cambridge University Press, pp. 41-58, 2009.

[15] S. Boztepe, "User Value: competing theories and models," International Journal of Design, Vol. 1, No.2, pp. 55-63, 2007.

[16] C. A. Le Dantec, E. S. Poole, and S. P. Wyche, "Values as lived experiences: evolving Value Sensitive Design in support of value discovery," Proc. CHI 2009, ACM, pp. 1141-1150, 2009.

[17] S. H. Schwartz, "An Overview of the Schwartz theory of Basic Values," Online Readings in Psychology and Culture, 2(1), Article 11, 2012, https://doi.org/10.9707/2307-0919.1116

[18] J. M. Floch, Semiotics, Marketing and Communication: beneath the signs, the strategies. Palgrave MacMillan, 2001.

[19] L. Anticoli and E. Toppano, "Technological mediation of ontologies: the need for tools to help designers in materializing ethics," International Journal of Philosophy Study, Vol. 1, Issue 3, pp. 23-31, 2013.

[20] B. Fogg. Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann, San Francisco, 2003.

[21] R. H. Thaler and C. S. Sunstein. NUDGE. Improving decisions about health, wealth, and happiness, Yale University Press, 2008.

[22] Citroen Bx Campaign, YouTube, URL: https://www.youtube.com/watch?v=jH2HLRpIq2Y [retrieved: April, 2018].

[23] E. Toppano and V. Roberto, "Semiotic annotation of narrative video commercials: bridging the gap between artifacts and ontologies," International Journal on Advances in Internet Technology, vol. 10, nr. 3&4, pp. 145-162, IARIA, 2017, http://www.iariajournals.org/internet_technology/ [retrieved: April, 2018].

[24] E. Cambria, "Affective Computing and Sentiment Analysis," IEEE Intelligent Systems, March/April, pp. 102-107, 2009.

[25] W3C. Emotion Markup Language (EmotionML), May 2014, URL: https://www.w3.org/TR/emotionml/[retrieved: April, 2018].

An Open Data Approach to Publish Relational Data

Miguel Bento Alves

ESTG, Instituto Politécnico de Viana do Castelo
4900-348 Viana do Castelo
Atlanta, Georgia 30332–0250
NOVA-LINCS, Universidade Nova de Lisboa
2829-516 Caparica, Portugal
Email: mba@estg.ipvc.pt

João Ferreira Nunes

ESTG, Instituto Politécnico de Viana do Castelo
4900-348 Viana do Castelo
Email: joao.nunes@estg.ipvc.pt

***Abstract*—Open Data is a principle that defines that data should be freely available to all, without any kind of restrictions from copyright, patents or other mechanisms of control. During this work, we've applied this concept into data that was modelled using a relational methodology. Thus, we've transformed the relational data into Open Data using a Semantic Web approach, namely RDF data. Furthermore, we've implemented a set of relational restrictions in the RDF data by means of semantic rules. These rules are used to guarantee the integrity of the Open Data repository. Tools were developed to manipulate the Open Data repository, ensuring the data integrity.**

***Index Terms*—Open Data; Semantic Web; Semantic Rules;**

I. Introduction

Open Data's principle [1] claims that data should be freely available, without any kind of restrictions from copyright, patents or other mechanisms of control. Another key concept that it is implicit to this ideal, is the interoperability, which refers to the capability of several systems and organizations in working together. In this specific case, it refers to the capability to combine - or inter-operate - different sets of data.

The concept of Open Data derives in a sense from Semantic Web [2], introduced by Berners-Lee [3]. Semantic Web is realized by assigning some meaning to the published content over the Internet in a way that it becomes discernible both to humans and computers. In this way, interoperability and cooperation between systems is enhanced. The meaning of the content is achieved by its classification and its relation with ontologies, which is a model that represents a set of concepts within a domain and the relationships between them.

Data publication in open format and its usage has been a hot topic within the scientific community over the past years. The Linked Open Data (LOD) project [4] is a good example of this practice. It aims to create structured and interconnected datasets, generating a data cloud. The LOD project contains more than 31 billion facts, linking more than 500 million facts to each other. A central dataset in the Linked Open Data project is the DBpedia [5], created from data extracted from Wikipedia containing over 1 billion facts. The LOD project proposes the publication of data using Web standards along with *links* to other data sources, giving a semantic context that allows easy access and easy interpretation of data. Linked Data also implies the use of standards, such as HTTP, RDF [6] or SPARQL [7], making it easier to use on the Web.

In our project, we created a data repository capable to publish information using the Open Data philosophy, so that it could be used externally either by humans and machines. One of the requirements imposed was that the project information should be centralized and consolidated through Semantic Web principles.

This project arises with the need to share and make public the data produced under the TREASURE project - a Research & Innovation Action financed by European Commission under the Horizon 2020 (grant agreement no. 634476). The aim of the project is to improve knowledge, skills and competences necessary to develop existing and create new sustainable pork chains based on European local pig genetic resources (local breeds). Initially, the information requirements were analyzed based on a relational model approach [8] to create a relational database. In order to enable the reuse of all the work produced during the initial phase of the project, it was decided to replicate the relational model for a Semantic Web approach. Therefore, the entire relational model was transformed into an RDF model. The challenge is to represent the relational structures, consisting of tables, fields and the stored data, in triple *Subject-Property-Subject* inherent to the RDF model. Although the transformation of the relational model into a RDF model is not a new topic (as can be seen in Section V), none of the approaches studied fulfilled our requirements for this project. In the best of our knowledge, we do not identify any work that follows the approach we have done in this work and which will be detailed throughout this document. We've also developed two tools to manipulate data in the open data repository. The first one allows select one local database and transfer the data to the Open Data repository. The second one is a SPARQL endpoint, that can be used either in a program or with a Web interface, that allows the execution of SPARQL commands in the central repository. Both tools guarantees the integrity of the data considering the relational constraints implemented.

This paper is organized as follows: in Section II, we describe the ontologies created in our approach and the data structure in datasets. In Section III the tools that were developed are described and, in Section IV, we present the created semantic rules to guarantee the integrity of the data. In Section V, some of the related work described is presented and compared to our approach. Conclusions and some ideas to be developed in the near future are presented in Section VI.

II. Ontologies and datasets

Although the focus of our work was to publish the produced data from the TREASURE project on an Open Data approach, the developed system that we designed is adaptable to any relational database. To accomplish this we proposed a three layers model, where at the upper-level the most important (or most used) concepts of the relational model are modeled; at the middle-level the meta-model of the relational database is modeled; and at the lower-level the database information is represented. With the two highest layers, we have all the knowledge on how the information in a database is organized, and from that we can extract information about what is modeled. This will also allows to support reasoning on the data model, as will be demonstrated throughout this work.

A. Relational Model Ontology

In the upper-level, we have created an ontology to represent the concepts of relational databases. Not all the concepts of relational databases were modeled, since we decided to model the most commonly used concepts of our project. The ontology was modeled in OWL [9] and it is represented in Figure 1.

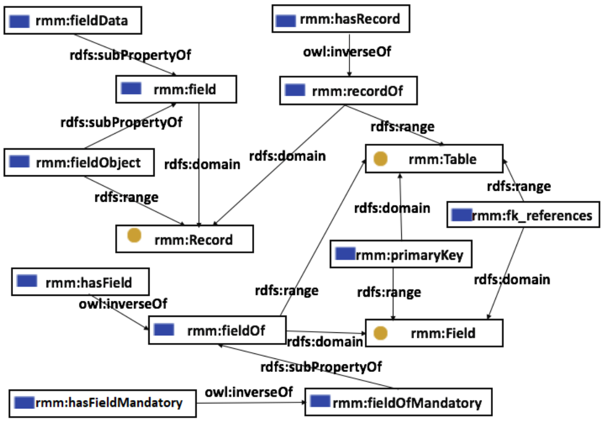


Fig. 1. Relational Model Ontology

B. Database meta-model

At the middle-level is represented the meta-model of a database, namely and as an example, which tables were created, which fields have each table, the primary keys of the tables and the foreign keys. In the Figure 2, a sub-model of the data model of this project is represented. In the list Listing 1 is encoded in OWL the sub-model shown in Figure 2 (due lack of space, we do not list all triples).

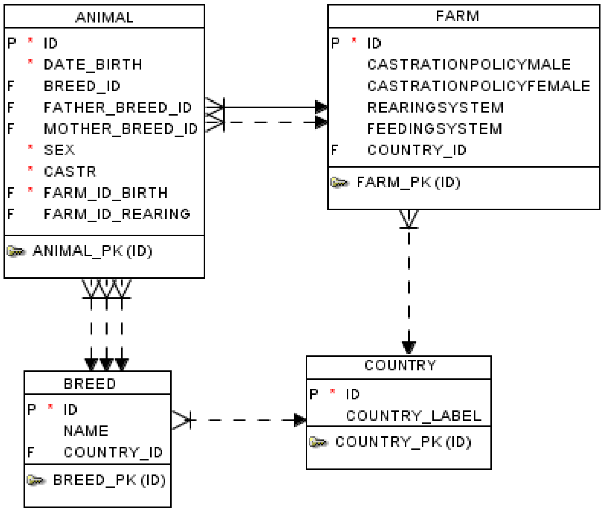


Fig. 2. A sub-model of the project Data Model

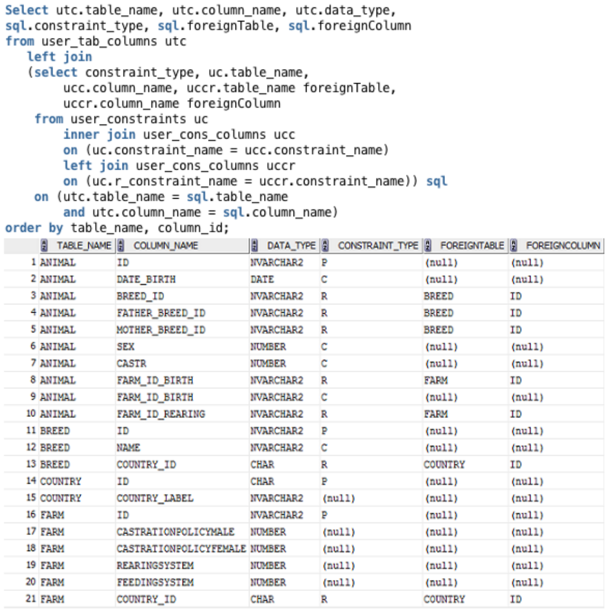
```
exa_mm:Country rdf:type rmm:Table ;
      rmm:primaryKey <http://www.example.com/exa_mm.ttl/Country#id>.
<http://www.example.com/exa_mm.ttl/Country#id>rmm:fieldOf exa_mm:Country ;
      rdfs:subPropertyOf rmm:fieldData .
<http://www.example.com/exa_mm.ttl/Country#country_label>
      rmm:fieldOf exa_mm:Country ;
      rdfs:subPropertyOf rmm:fieldData .

exa_mm:Breed rdf:type rmm:Table ;
      rmm:primaryKey <http://www.example.com/exa_mm.ttl/Breed#id>.
<http://www.example.com/exa_mm.ttl/Breed#id>rmm:fieldOf exa_mm:Breed ;
      rdfs:subPropertyOf rmm:fieldData .
<http://www.example.com/exa_mm.ttl/Breed#country_id>
      rmm:fieldOf exa_mm:Breed ;
      rdfs:subPropertyOf rmm:fieldObject ;
      rmm:fk_references exa_mm:Country .

exa_mm:Farm rdf:type rmm:Table ;
      rmm:primaryKey <http://www.example.com/exa_mm.ttl/Farm#id>.
<http://www.example.com/exa_mm.ttl/Farm#id>rmm:fieldOf exa_mm:Farm ;
      rdfs:subPropertyOf rmm:fieldData .
<http://www.example.com/exa_mm.ttl/Farm#castrationPolicyMale>
      rmm:fieldOf exa_mm:Farm ;
      rdfs:subPropertyOf rmm:fieldData .
<http://www.example.com/exa_mm.ttl/Farm#country_id>
      rmm:fieldOf exa_mm:Farm ;
      rdfs:subPropertyOf rmm:fieldObject ;
      rmm:fk_references exa_mm:Country .

exa_mm:Animal rdf:type rmm:Table ;
      rmm:primaryKey <http://www.example.com/exa_mm.ttl/Animal#id>.
<http://www.example.com/exa_mm.ttl/Animal#id>
      rmm:fieldOf exa_mm:Animal ;
      rdfs:subPropertyOf rmm:fieldData .
<http://www.example.com/exa_mm.ttl/Animal#breed_id>
      rmm:fieldOfMandatory exa_mm:Animal ;
      rdfs:subPropertyOf rmm:fieldObject ;
      rmm:fk_references exa_mm:Breed .
<http://www.example.com/exa_mm.ttl/Animal#father_breed_id>
      rmm:fieldOf exa_mm:Animal ;
      rdfs:subPropertyOf rmm:fieldObject ;
      rmm:fk_references exa_mm:Breed .
```

Listing 1 - Encoding of the relational sub-model presented in Figure 2

The middle-level layer that represents the objects of a given database contains the same information that can be found in the data dictionary of the database manager system. Although the catalogs (data dictionaries) of the different database management systems are not standardized, the essential information is available in all of them. In this way, we can create automatisms so that this middle-level layer can be created in an automatic way. In Figure 3 we can see a SQL command that

ISBN: 978-1-61208-636-1

extracts the metadata (in Oracle) from the relational sub-model represented in the Figure 2 and the result of this command. Comparing the result of the SQL command with the OWL encoding listed in Listing 1 makes it obvious that the OWL encoding can be done automatically.



```
Select utc.table_name, utc.column_name, utc.data_type,
sql.constraint_type, sql.foreignTable, sql.foreignColumn
from user_tab_columns utc
   left join
   (select constraint_type, uc.table_name,
        ucc.column_name, uccr.table_name foreignTable,
        uccr.column_name foreignColumn
    from user_constraints uc
        inner join user_cons_columns ucc
        on (uc.constraint_name = ucc.constraint_name)
        left join user_cons_columns uccr
        on (uc.r_constraint_name = uccr.constraint_name)) sql
    on (utc.table_name = sql.table_name
        and utc.column_name = sql.column_name)
order by table_name, column_id;
```

	TABLE_NAME	COLUMN_NAME	DATA_TYPE	CONSTRAINT_TYPE	FOREIGNTABLE	FOREIGNCOLUMN
1	ANIMAL	ID	NVARCHAR2	P	(null)	(null)
2	ANIMAL	DATE_BIRTH	DATE	C	(null)	(null)
3	ANIMAL	BREED_ID	NVARCHAR2	R	BREED	ID
4	ANIMAL	FATHER_BREED_ID	NVARCHAR2	R	BREED	ID
5	ANIMAL	MOTHER_BREED_ID	NVARCHAR2	R	BREED	ID
6	ANIMAL	SEX	NUMBER	C	(null)	(null)
7	ANIMAL	CASTR	NUMBER	C	(null)	(null)
8	ANIMAL	FARM_ID_BIRTH	NVARCHAR2	R	FARM	ID
9	ANIMAL	FARM_ID_BIRTH	NVARCHAR2	C	(null)	(null)
10	ANIMAL	FARM_ID_REARING	NVARCHAR2	R	FARM	ID
11	BREED	ID	NVARCHAR2	P	(null)	(null)
12	BREED	NAME	NVARCHAR2	C	(null)	(null)
13	BREED	COUNTRY_ID	CHAR	R	COUNTRY	ID
14	COUNTRY	ID	CHAR	P	(null)	(null)
15	COUNTRY	COUNTRY_LABEL	NVARCHAR2	(null)	(null)	(null)
16	FARM	ID	NVARCHAR2	P	(null)	(null)
17	FARM	CASTRATIONPOLICYMALE	NUMBER	(null)	(null)	(null)
18	FARM	CASTRATIONPOLICYFEMALE	NUMBER	(null)	(null)	(null)
19	FARM	REARINGSYSTEM	NUMBER	(null)	(null)	(null)
20	FARM	FEEDINGSYSTEM	NUMBER	(null)	(null)	(null)
21	FARM	COUNTRY_ID	CHAR	R	COUNTRY	ID

Fig. 3. Example of a SQL command that extracts the metadata from the relational sub-model

C. Dataset

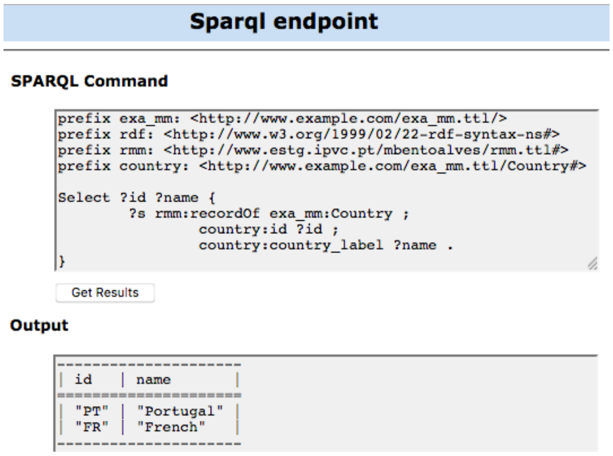
In the most specific layer, the data themselves are represented. In our project, the information produced by the TREASURE project, was initially stored in the database. In the list Listing 2 some example data is presented. Note that in the case of foreign keys, we decided to point to the record instead of storing a key value, as was implicit in the high-level ontological model.

```
country:pt rmm:recordOf exa_mm:Country ;
     country:id "PT" ;
     country:country_label "Portugal" .

exa_mm:Country rmm:hasRecord [
     country:id "FR" ;
     country:country_label "French"
].

breed:b2703 rmm:recordOf exa_mm:Breed ;
     breed:id "b2703" ;
     breed:name "Bisaro" ;
     breed:country_id country:pt .
```

Listing 2 - Example of data contained in the dataset

III. System developed

Our system was developed using the Jena Framework [10], a free and open source Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS [11], OWL, a query engine for SPARQL and it includes a rule-based inference engine. Jena is widely accepted for Semantic Web applications because it offers an "all-in-one" Java solution. Our system consists by two sub-systems. One of them allows to select one local database and transfer the data to the Open Data repository. The other one is a SPARQL endpoint, that can be used either in a program or with a Web interface, that allows the execution of SPARQL commands in the central repository. We decided to develop our own SPARQL endpoint instead of using Fuseki [12], the SPARQL server of Jena package, because we implemented several relational constraints over our central repository and we want to control the integrity of the data against the relational constraints. Every SPARQL command that change the data content must carry the central repository from an integrity state to other integrity state. The relational constraints implemented are detailed in Section IV. In the Figure 4 is showed the interface of our SPARQL endpoint.



Sparql endpoint

SPARQL Command

```
prefix exa_mm: <http://www.example.com/exa_mm.ttl/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rmm: <http://www.estg.ipvc.pt/mbentoalves/rmm.ttl#>
prefix country: <http://www.example.com/exa_mm.ttl/Country#>

Select ?id ?name {
        ?s rmm:recordOf exa_mm:Country ;
                country:id ?id ;
                country:country_label ?name .
}
```

Output

```
---------------------
| id   | name       |
=====================
| "PT" | "Portugal" |
| "FR" | "French"   |
---------------------
```

Fig. 4. Sparql Endpoint

IV. Rules to implement relational constraints

When we convert a relational database into an Open Data repository, the main concern is to ensure its integrity, taking into account the inherent constraints of the relational model. In OWL, we do not have all the necessary mechanisms to impose the constraints of the relational model. Therefore, it was necessary to implement relational constraints using semantic rules. A rule language is needed for several reasons, at least because of the limitations of OWL [13].

The process to ensure the integrity of the Open Data repository is as follows: each time a command that can change the data content of the repository is invoked, a SPARQL command is executed to verify if any rule or constraint has not been met; if this occurs, the reverse command is executed in order to reset the database to the last integrity state identified, and an error is issued. All invoked commands that might change the contents of the data repository go to a queue in order to run sequentially. In this approach, the integrity of the data repository with concurrent commands was not thought of.

In general, a row in a table represents a relationship among a set of values where each element is termed an attribute value or a field. Also, in a table, all its fields are distinct, i.e., each table cannot have the same field more than one time. The Rule 1 expresses this condition and if for some reason it occurs, an error message is returned.

```
(err:MultipleFieldError err:violation ?Msg) <-
      (?Record1 ?Field ?Value),
      (?Record1 rdf:type rmm:Record),
      (?Field rdf:type rmm:Field),
      (?Record1 ?Field ?Value2),
      (?Record1 rmm:recordOf ?Table),
      notEqual(?Value, ?Value2),
      strConcat(’Table -> ’, ?Table, ’;record -> ’, ?Record1, ’, field -> ’, ?Field,
?Msg1),
      strConcat(?Msg1, ’; values -> (’, ?Value, ’, ’, ?Value2, ’)’, ?Msg) .
```

Rule 1 - Rule to avoid repeated fields in a table

Each table should have a primary key, a key, or a set of keys, that identifies univocally a row. If two distinct rows have the same primary key, then its constraint is violated. The Rule 2 ensures that the primary key isn’t violated. We defined that all primary keys must be constituted by single fields and the Rule 2 assumes this assumption. Furthermore, in sub-section IV-A we discuss about composite keys.

```
(err:PrimaryKeyError err:violation ?Msg) <-
      (?Record ?Field ?Value),
      (?Record rdf:type rmm:Record),
      (?Field rdf:type rmm:Field),
      (?Table rmm:primaryKey ?Field),
      (?Record1 ?Field ?Value),
      notEqual(?Record, ?Record1),
      strConcat(’Table -> ’, ?Table, ’ ; records -> (’, ?Record, ’, ’, ?Record1, ’)’,
?Msg1),
      strConcat(?Msg1, ’; value -> ’, ?Value, ’)’, ?Msg) .
```

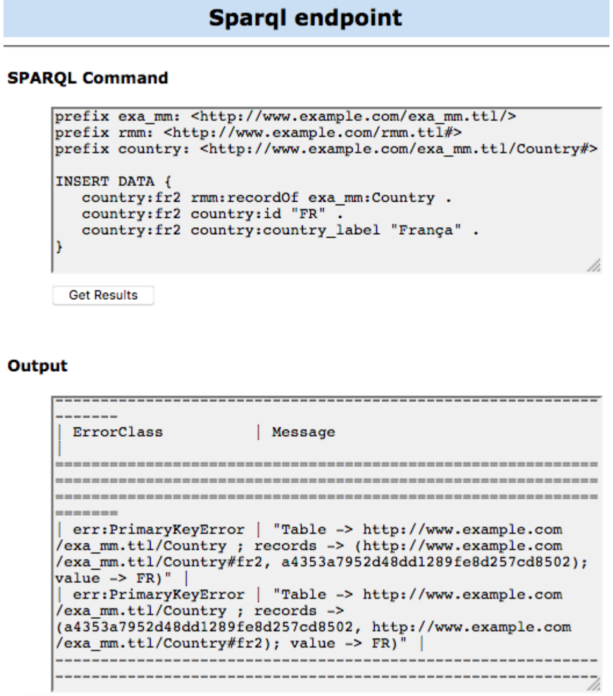
Rule 2 - Rule to avoid violation of primary key constraint

When a given table has a foreign key, then its value either is null or must exist in case it is a primary key. Our approach was to point into the record in the related table, i.e., where the key is primary. The Rule 3 ensures that a reference made in a foreign key exists in the table where the key is primary.

```
(err:ForeignKeyError err:violation ?Msg) <-
      (?Record ?Field ?Value),
      (?Record rdf:type rmm:Record),
      (?Field rmm:fk_references ?TablePK),
      (?TablePK rmm:primaryKey ?FieldPK),
      noValue(?Value rmm:recordOf ?TablePK),
      (?Record rmm:recordOf ?Table),
      strConcat(’Table -> ’, ?Table, ’ ; record -> ’, ?Record, ’ ; value -> ’, ?Value,
?Msg1),
      strConcat(?Msg1, ’ ; TablePK -> ’, ?TablePK, ?Msg) .
```

Rule 3 - Rule to avoid violation of foreign key constraint

In the ontological model of the relational constraints errors, the classes *err:MultipleFieldError*, *err:PrimaryKeyError* and *err:ForeignKeyError* are sub-classes of the class *err:Error*. So, after a SPARQL command that change the data of the repository, we look for all errors whose class is a sub-class of *err:Error*. In Figure 5 we give an example of a command that violates the primary key.



Sparql endpoint

SPARQL Command

```
prefix exa_mm: <http://www.example.com/exa_mm.ttl/>
prefix rmm: <http://www.example.com/rmm.ttl#>
prefix country: <http://www.example.com/exa_mm.ttl/Country#>

INSERT DATA {
   country:fr2 rmm:recordOf exa_mm:Country .
   country:fr2 country:id "FR" .
   country:fr2 country:country_label "França" .
}
```

Output

```
--------------------------------------------------------------
-------
| ErrorClass          | Message
|
==============================================================
==============================================================
==============================================================
=======
| err:PrimaryKeyError | "Table -> http://www.example.com
/exa_mm.ttl/Country ; records -> (http://www.example.com
/exa_mm.ttl/Country#fr2, a4353a7952d48dd1289fe8d257cd8502);
value -> FR)" |
| err:PrimaryKeyError | "Table -> http://www.example.com
/exa_mm.ttl/Country ; records ->
(a4353a7952d48dd1289fe8d257cd8502, http://www.example.com
/exa_mm.ttl/Country#fr2); value -> FR)" |
--------------------------------------------------------------
--------------------------------------------------------------
```

Fig. 5. Sparql endpoint integrity error

A. *Composite keys*

As previously mentioned, during the modeling process of the relational database we assumed that all keys, primary and foreign, are single keys. Composite keys were not in the scope of our project in this first approach. However, as future work we want to deal with composite keys. We can already introduce some solutions to deal with it: the primary key could be a field, when it is a single key, or a list of fields in case of composite keys. Considering this assumption, the Rule 4 defines that an error is returned whenever a primary composite key is violated.

```
(err:PrimaryKeyError err:violation ?Msg) <-
      (?Record rmm:recordOf ?Table),
      (?Record1 rmm:recordOf ?Table),
      notEqual(?Record, ?Record1),
      (?Table rmm:primaryKey ?ListFields),
      (err:PrimaryKeyError err:validate validatePK(?Record, ?Record1, ?ListFields)),
      strConcat(’Table -> ’, ?Table, ’ ; records -> (’, ?Record, ’, ’, ?Record1, ’)’,
?Msg) .

-> table(err:validate).

(err:PrimaryKeyError err:validate validatePK(?Record, ?Record1, rdf:nil)) <-
IsTrue().

(err:PrimaryKeyError err:validate validatePK(?Record, ?Record1, ?ListFields)) <-
      notEqual(?ListFields, rdf:nil),
      (?ListFields rdf:first ?Field),
      (?Record ?Field ?Value),
      (?Record1 ?Field ?Value),
      (?ListFields rdf:rest ?RestFields),
      (err:PrimaryKeyError err:validate validatePK(?Record, ?Record1, ?RestFields))
.
```

Rule 4 - Rule to avoid the violation of primary key constraint in composite keys

Considering our approach that a foreign key points to a record of the primary key, the problem of a composite key doesn't happen. However, as we want to foreseen in the future the two possibilities (point to a record or keep the value of the key), we need a rule with the same approach done to the Rule 4.
Another solution is construct a specific rule to a table with composite keys. In this case we loose the generality of the rule but we can create the rule in a automatic manner from a template and from the database catalog.

V. Related work

A W3C Recommendation that describes R2RML, a language for expressing customized mappings from relational databases to RDF datasets is presented in [14]. In [15], it is presented a transformation of relational model to RDF model, a two-step approach where is extracted the semantics of relational model and then it is transformed in RDF models. In general, the approach followed in this work is done by us when we create the middle-tier, the representation of the meta-model in RDF. We also refer that we can create this middle-tier automatically by the information extracted from the database catalog. In [16], it is presented a tool that transforms relational data into OWL2 and performs data validation to report errors in the data. This validation is accomplished through rules. In [17], it is presented a set of mappings that transforms relational data and schema to Semantic Web models. OWL and SWRL are used to express relational constraints satisfaction or validation conditions. SPARQL query is used to verify these constraints. This work can be considered close to our work in the relational constraints approach since they used rules in backward mode to verify possible violation of relational constraints. However, our rules are generalized to all kind of data models and are based on meta-model while in their work the rules are designed for a specific data model. In [18], it is performed a practical evaluation of existing approaches in automatic generation of ontology from data models. The purpose of this work is the evaluation of the availability of existing approaches for automatic or semi-automatic generation of ontology from data models, the evaluation of the tools according to their operability and the evaluation of the resulting ontologies to assess their quality in supporting semantic interoperability. In [19], it is performed a survey of the works about the creation of an ontology from an existing database instance and the discovery of mappings between an existing database instance and an existing ontology. It is also presented the motivation, benefits, challenges and solutions.
Some solutions that are presented in our work, could be developed using Shapes Constraint Language (SHACL) [20], which is a World Wide Web Consortium (W3C) specification for describing and validating RDF graphs data against a set of conditions. However, developing generalized SHACL rules for all models it is not possible, in the best of our knowledge. Furthermore, even for a specific data model it is not possible develop SHACL rules for constraints that includes composite fields.

VI. Conclusion

The main goal of this work was to produce a public repository of the information retrieved from the TREASURE project using the principles inherent to the Open Data philosophy. Initially, we modeled the information using a relational approach, having created a data model that would take into account the requirements of the project in terms of information. During the evolution process to an Open Data repository, and in order to avoid another analysis process, we decided to transform the relational information into RDF information. Afterwards, a Semantic Web approach was followed. The data repository is organized in a three layer schema. In a more generic layer, is the ontological model that represents the relational model and where its objects are characterized. In an intermediate layer is the meta-model of the database, where the data structure is described. Finally, in a more specific layer it is the data itself, which in our case, is the data generated by the TREASURE project. We created a system that allows the data repository to be automatically fed from a database, as well as a SPARQL endpoint enabling both updating and selecting data from the data repository. This SPARQL endpoint can be used either in a program or through a Web interface. Moreover we implemented a set of constraint mechanisms of the relational model by means of logical rules, ensuring the consistency of the data repository. Any change of data conducts the data repository from one *integrity state* to another *integrity state*. If any command does not respect the defined rules, a rollback of the operation is executed. Although this work was developed under TREASURE project, we have always been concerned with generalizing ontological models and the developed systems for any domain.

As future work we intend to implement more constraint mechanisms of the relational model for a more comprehensive use. In the way that our system is structured, this involves the development of more semantic rules. The semantic rules are kept in plain text file, therefore, we do not need any programming code change. As we refer before, composite keys in primary and foreign keys will be implemented. With regard to the TREASURE project, we intend to make an analysis using an OWL approach, instead of a relational approach, in order to have richer descriptions of the data. In addition, we intend to make mappings so that it can be framed as Linked Open Data, namely linking the main concepts to DBpedia.

References

[1] O. K. Foundation, Ed., *The Open Data Handbook / Das Open Data Handbuch*, 2012. [Online]. Available: http://opendatahandbook.org/

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web a new form of web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, pp. 28–37, May 2001.

[3] T. Berners-Lee, "Linked-data design issues," W3C design issue document, June 2009. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far." *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, March 2009.

[5] S. Auer et al., “Dbpedia: A nucleus for a web of open data,” *The Semantic Web*, pp. 722–735, November 2008.
[6] R. Cyganiak, D. Wood, and M. Lanthaler Resource description framework (rdf): Concepts and abstract syntax. February 2014. [Online]. Available: https://www.w3.org/TR/rdf11-concepts/
[7] E. Prud’hommeaux and A. Seaborne, “SPARQL Query Language for RDF,” W3C, Tech. Rep., March 2013. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/
[8] E. F. Codd, “A relational model of data for large shared data banks,” *Commun. ACM*, vol. 13, no. 6, pp. 377–387, Jun. 1970. [Online]. Available: http://doi.acm.org/10.1145/362384.362685
[9] S. Bechhofer et al., Owl web ontology language reference. February 2004. [Online]. Available: http://www.w3.org/TR/owl-ref/.
[10] A. Seaborne. Jena, a Semantic Web Framework. November 2010. [Online]. Available: http://wiki.apache.org/incubator/JenaProposal
[11] P. Hayes and P. Patel-Schneider. Rdf semantics. 2014 February. [Online]. Available: https://www.w3.org/TR/rdf11-mt/
[12] Jena fuseki. 2014 [Online]. Available: https://jena.apache.org/documentation/serving_data/
[13] B. Parsia et al., Cautiously approaching swrl. Preprint submitted to Elsevier Science. 2005. [Online]. Available: http://www.mindswap.org/papers/CautiousSWRL.pdf
[14] S. Das, S. Sundara, and R. Cyganiak. R2rml: Rdb to rdf mapping language. September 2012. [Online]. Available: https://www.w3.org/TR/r2rml/
[15] Y. Lv and Z. Ma, “Transformation of relational model to RDF model,” in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Singapore*, October 2008, pp. 506–511. [Online]. Available: https://doi.org/10.1109/ICSMC.2008.4811327
[16] A. Yajai and G. Sriharee, “Eertoowl2: A tool for transforming RDB data to OWL2 for data validation,” in *IEEE 24th International Conference on Tools with Artificial Intelligence, ICTAI 2012, Athens, Greece*, November 2012, pp. 970–975. [Online]. Available: https://doi.org/10.1109/ICTAI.2012.137
[17] X. Fan, P. Zhang, and J. Zhao, “Transformation of relational database schema to semantics web model,” in *Second International Conference on Communication Systems, Networks and Applications*. IEEE, June 2010. [Online]. Available: https://doi.org/10.1109/iccsna.2010.5588750
[18] B. E. Idrissi, S. Baïna, and K. Baïna, “Automatic generation of ontology from data models: A practical evaluation of existing approaches,” in *IEEE 7th International Conference on Research Challenges in Information Science, RCIS 2013, Paris, France*, May 2013, pp. 1–12. [Online]. Available: https://doi.org/10.1109/RCIS.2013.6577694
[19] D.-E. Spanos, P. Stavrou, and N. Mitrou, “Bringing relational databases into the semantic web: A survey,” *Semant. web*, vol. 3, no. 2, pp. 169–209, April 2012. [Online]. Available: http://dx.doi.org/10.3233/SW-2011-0055
[20] H. Knublauch and D. Kontokostas. Shapes constraint language (shacl). W3C. July 2017. [Online]. Available: https://www.w3.org/TR/shacl/