TOWARDS OPEN DATA IN MACEDONIA: CRIME MAP BASED ON MINISTRY OF INTERNAL AFFAIRS' BULLETINS

Damjan Temelkovski, Milos Jovanovik, Igor Mishkovski, Dimitar Trajanov
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, Republic of Macedonia

ABSTRACT

Today, many organizations and institutions have vast collections of datasets and databases filled with information that, in general, can turn out to be very useful for individuals and for the society [1]. The Police is one such institution which is obliged by law to keep an archive of all the information it deals with on a daily basis. All that information about the criminal events that have occurred in the past can be used in many ways. In this work, we present a research project focused on crime analysis using the concept of crime map, which can be very useful for the law enforcement agencies as well as for the citizens. This research project resulted in an effective crime map of the Republic of Macedonia, which consists of about 1800 events in total. With this project we hope to encourage and motivate the Ministry of Internal Affairs, as well as other institutions, to continue liberating valuable data and bring the benefits of the Open Data idea to its citizens.

I. INTRODUCTION

The laws of the legal state were created to protect the individual as well as the entire society. Among these laws are the copyright and patent laws. However, today we can say that the world feels imprisoned behind its own regulations and is in need of liberation [2]. The "Open" philosophy, mainly Open Source, but also Open Content, Open Access, etc., is not a new concept. It is based on the idea of opening up the content, so that everyone can use it and alter it in his/hers own matter. The concept of Open Data is part of that philosophy and has been for some time now, but only with the rise of the Internet and the World Wide Web it gained in popularity as a distinct term [3]. All of the data archives in the government bodies are managed from the taxpayers' money, so the citizens have the right to get, use and share that data freely [4].

One of the fields that can benefit from the Open Data concept is crime analysis. Crime analysis is a law enforcement function that involves systematic analysis for identifying, and analysing patterns and trends in crime and disorder. A key component of the field of crime analysis is crime mapping, which gives a visual summarization of all the criminal events that undergo in a certain area, together with information about the location and severity of the crimes.

In June 2011, the Ministry of Internal Affairs of the Republic of Macedonia (MOI) started issuing a bulletin on their official website where they publish a selection of the criminal events of the past day. This was one step closer to the idea of opening this type of data to the general public, taking into consideration the privacy laws and the protection of the identity of the persons involved.

Using the official MOI's bulletin and with the help of the Open Data concept in this paper we produced a crime map of the Republic of Macedonia. However, because the official MOI bulletins were written in plain, natural language, in order to automate the mapping process (map an event to a specific location) we had to work with the basis of a popular field of artificial intelligence and linguistics, called natural language processing (NLP). Thus, this paper is organized as follows. In Section 2 we give the related work which inspired this paper. In Section 3 we describe the NLP used for mapping an event to a specific location, and the difficulties which we encountered in doing this. Here, we also describe the process of geocoding the locations on a map. Section 4 concludes the paper and gives directions for future work.

II. RELATED WORK

Many law enforcement agencies from countries around the world have created publicly available on-line crime maps. A fine example is the UK Police crime analysis section [5]. It is worth noting that they do not show actual locations in order to protect the privacy of the individual. Instead, they show a summary of events that occurred near a public building, such as a museum, a theatre, or a park. An example of the crime map for the city of Bristol is shown on Fig.1.

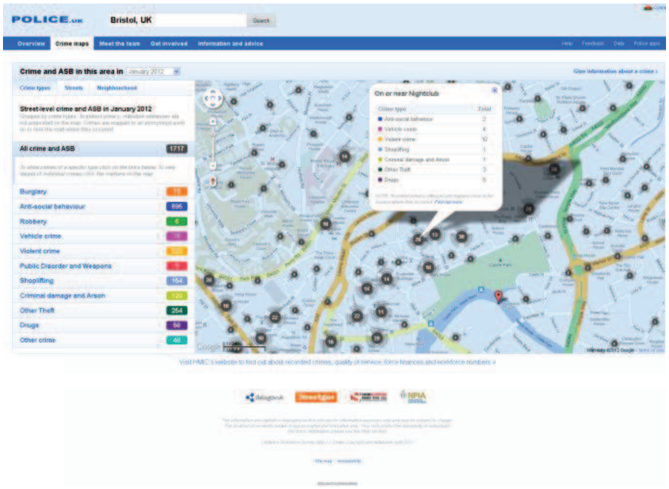


Figure 1: An example of the crime map of the UK Police for the city of Bristol.

Other examples are the crime maps from The Omega Group, a company that supports US law enforcements (see Fig.2) [6], and the US non-profit organization The Reinvestment Fund [7].

Our idea is to make a similar crime map of Macedonia, despite the notion that Macedonia and its police department are allegedly far behind those of the UK and the USA. This project has proven just the opposite; at least as far as crime mapping is concerned.

A. *Open Data*

The Open Data initiative was encouraged by government projects such as the 2009 US government project called Data.gov [8]. It is considered responsible for a chain of similar projects that followed in the UK [9], Norway [10], Russia [11], and many others [12]. The goal of these websites is to increase public access to high value, machine readable datasets generated by governmental organizations. Thereby, this concept is connected with the idea of Open Government, in which the citizens have the right to access all of the documents of the government, in order to provide an effective public insight.

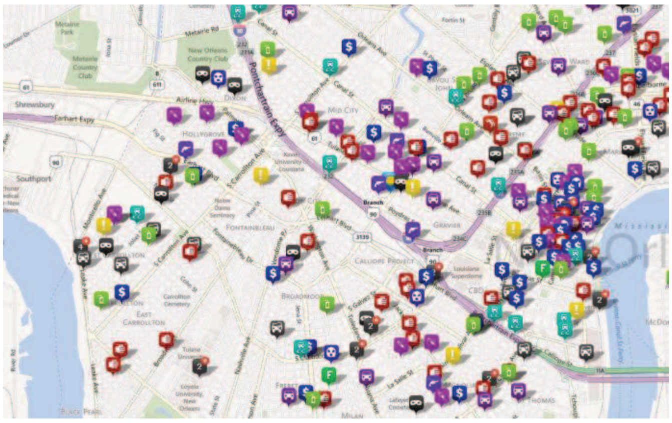
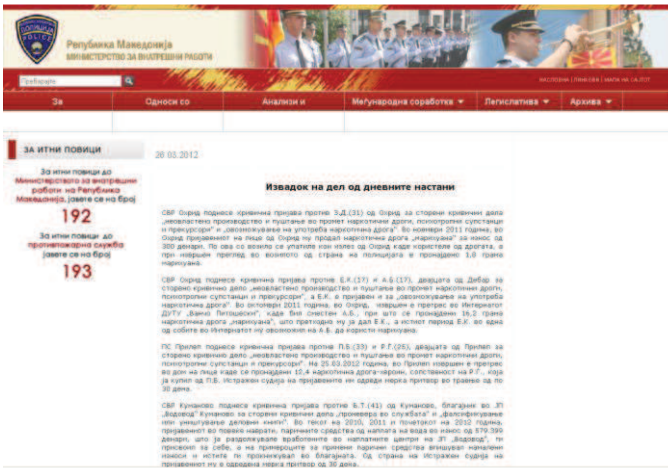


Figure 2: An example of the crime map of the US Police for Saint Louis, LA.

The Republic of Macedonia in the past several years has shown interest and made some efforts in making the Macedonian government more open. It is currently in the process of joining the Open Government Partnership [13]. We hope that by generating momentum in the Open Data field, we will be able to give contribution in this process.



ЗА ИТНИ ПОВИЦИ

За итни повици до Министерството за внатрешни работи на Република Македонија, јавете се на број

192

За итни повици до противпожарна служба јавете се на број

193

26.03.2012

Изводок на дел од дневните настани

СВР Охрид поднесе кривична пријава против З.Д.(31) од Охрид за сторени кривични дела „неовластено производство и пуштање во промет наркотични дроги, психотропни супстанци и прекурсори" и „овозможување на употреба наркотична дрога". Во ноември 2011 година, во Охрид пријавениот на лице од Охрид му продал наркотична дрога „марихуана" за износ од 300 денари. По ова со возило се упатиле кон излез од Охрид каде користеле од дрогата, а при извршен преглед во возилото од страна на полицијата е пронајдено 1,8 грама марихуана.

СВР Охрид поднесе кривична пријава против Е.К.(17) и А.Б.(17), двајцата од Дебар за сторено кривично дело „неовластено производство и пуштање во промет наркотични дроги, психотропни супстанци и прекурсори", а Е.К. е пријавен и за „овозможување на употреба наркотична дрога". Во октомври 2011 година, во Охрид, извршен е претрес во Интернатот ДУТУ „Ванчо Питошески", каде бил сместен А.Б., при што се пронајдени 16,2 грама наркотична дрога „марихуана", што претходно му ја дал Е.К., а истиот период Е.К. во една од собите во Интернатот му овозможил на А.Б. да користи марихуана.

ПС Прилеп поднесе кривична пријава против П.Б.(33) и Р.Г.(25), двајцата од Прилеп за сторено кривично дело „неовластено производство и пуштање во промет наркотични дроги, психотропни супстанци и прекурсори". На 25.03.2012 година, во Прилеп извршен е претрес во дом на лица каде се пронајдени 12,4 наркотична дрога-хероин, сопственост на Р.Г., која ја купил од П.Б. Истражен судија на пријавените им одреди мерка притвор во траење од по 30 дена.

СВР Куманово поднесе кривична пријава против Б.Т.(41) од Куманово, благајник во ЈП „Водовод" Куманово за сторени кривични дела „проневера во службата" и „фалсификување или уништување деловни книги". Во текот на 2010, 2011 и почетокот на 2012 година, пријавениот во повеќе наврати, паричните средства од наплата на вода во износ од 579.399 денари, што ја раздолжувале вработените во наплатните центри на ЈП „Водовод", ги присвоил за себе, а на примероците за примени парични средства впишувал намалени износи и истите ги прокнижувал во благајната. Од страна на Истражен судија на пријавениот му е одредена мерка притвор од 30 дена.

Figure 3: The Macedonian MOI's website and the daily bulletin (for 26.03.2012).

The Ministry of Internal Affairs of the Republic of Macedonia (MOI), as of June 2011, has been issuing an e-bulletin with a selection of the daily criminal events, as shown in Fig.3. This bulletin is in a plan text format, written in Macedonian language. If the location and the type of event can be extracted from this bulletin, then it would be simple enough to create a national crime map.

B. *Crime mapping*

A crime map can be regarded as a map of the criminal activities in the neighborhood used for one's personal safety and precaution. However, the scope of crime mapping is bigger than this. Using geographic information systems (GIS), crime analysts can overlay other datasets, such as census demographics, location of certain buildings of interest - such as schools, night clubs, etc. - to better understand the causes of crime and help law enforcement administrators devise strategies to deal with the problem and make the streets safer. Using crime mapping, they help the law enforcement management make better decisions, target resources, and formulate strategies. These maps can be also used for tactical analysis - crime forecasting, geographic profiling, etc. [14].

Computer-based crime mapping started developing in the 1980's by the US National Institute of Justice (NIJ). They started a project in the police department of Chicago which explored crime mapping as a way to help community policing. The success of this project paved the way for other ones, such as "CompStat" by the police department of New York City [15].

The Republic of Macedonia has not yet published an open crime map; however, they most likely have a certain internal crime mapping system. All information regarding this is kept private and closed, but we think it would be more beneficial for the citizens if it is open and available for use.

III. Crime Map for the Republic of Macedonia

A. *Using NLP for Text Analysis*

Natural language processing (NLP) is regarded a field of artificial intelligence and linguistics which deals with the computers and humans in the human native, natural language. The term is normally used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language [16]. The reason why NLP is of interest in our project is because the information that the MOI publishes on their website is written in a narrative and natural manner. The computer works well with data in a structured form, so some text processing was necessary in order to structure the data. There are many ways and algorithms to process natural language, many of the more modern and sophisticated ones use the paradigm of machine learning [17].

Addressing our problem from a simpler point of view, we used a basic keyword tagging method [18], where we established a group of repeating keywords in the texts, and we draw conclusions based on their presence or absence, and the phrases that follow or precede them.

The processing of the plain text bulletins is as follows: first we look for the name of the city, which we do with a basic comparison with a list of all the cities in Macedonia, and the input text. However, in the following we show some of the difficulties we encountered, for which we had to use text processing and matching.

For detecting the address at which the criminal event occurred, we use the Macedonian equivalent for the "str." suffix as a keyword, which in our case is a prefix: "ул." (from улица – street). Although it seems quite straight-forward, we have a problem with the common prefix for "boulevard", which in Macedonian is "бул." (from булевар). This prefix contains "ул." in it, but should be distinguished, since a street and a boulevard with the same name aren't uncommon in Macedonia.

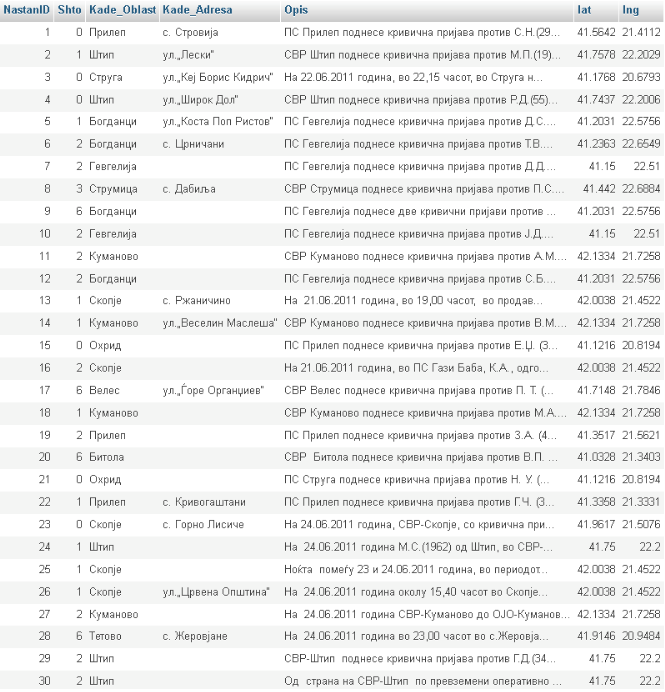
Furthermore, many events occur in the Macedonian countryside and MOI does not give street addresses for them, but only the name of the village. The problem here is that the bulletins do not have a conventional way of presenting names of villages, as it has for street names. Therefore, several keywords had to be taken in consideration, such as the prefix "с." or the entire word "село" preceding the name of the village (where село means village), or the definite article form: "селото" – meaning "the village". All three of those were usually followed by a „village name", as a way of presenting the name (with the Macedonian quotations marks: „"), but quite often represented as "village name" (with the English quotation marks: ""), or a mixture of the two, or no quotations at all. The fact that many villages have more than one word in their name made the idea of using a white-space delimiter inapplicable. Therefore, our decision tree grew bigger.

The type of the crime should also be extracted in a similar way and then categorized in order to pin a different icon-marker on the map. In order to do this, we used another example of open data in Macedonia. Namely, the equivalent of the US District Attorney (our Public Prosecutor) and his office have the entire code of criminal offences published on their official website [19]. From this code, we extracted all the possible official offences one can be charged for and then checked if there was a match with the plain-text input. It helped that some of the texts were written in a conventional way of writing the official criminal offence in quotes. However, this was not sufficient, so we used common descriptive words that appeared often as keywords, such as: fight (тепачка), shot (пукал), accident ((сообраќајна) несреќа), etc.

B. *The Database*

Using NLP as described previously, we managed to get a structured form of the data. We stored this data within a database. Our database is a table for all the events, consisting of a column for an ID number, a WHAT, a WHERE, a full description (the actual text from the bulletin), a latitude value and a longitude value gained by geocoding as described in the following part.

The WHAT value in the database is an integer which indicates the category in which the event is a part of. In the current stage of the project, we have 6 categories: armed offences, violent crimes, theft or burglaries, document-related crimes, drug-related crimes, and traffic accidents. A seventh category is used for labeling all other types of crimes.



NastanID	Shto	Kade_Oblast	Kade_Adresa	Opis	lat	lng
1	0	Прилеп	с. Стровија	ПС Прилеп поднесе кривична пријава против С.Н.(29...	41.5642	21.4112
2	1	Штип	ул.„Лески"	СВР Штип поднесе кривична пријава против М.П.(19)...	41.7578	22.2029
3	0	Струга	ул.„Кеј Борис Кидрич"	На 22.06.2011 година, во 22,15 часот, во Струга н...	41.1768	20.6793
4	0	Штип	ул.„Широк Дол"	СВР Штип поднесе кривична пријава против Р.Д.(55)...	41.7437	22.2006
5	1	Богданци	ул.„Коста Поп Ристов"	ПС Гевгелија поднесе кривична пријава против Д.С....	41.2031	22.5756
6	2	Богданци	с. Црничани	ПС Гевгелија поднесе кривична пријава против Т.В....	41.2363	22.6549
7	2	Гевгелија		ПС Гевгелија поднесе кривична пријава против Д.Д....	41.15	22.51
8	3	Струмица	с. Дабиља	СВР Струмица поднесе кривична пријава против П.С....	41.442	22.6884
9	6	Богданци		ПС Гевгелија поднесе две кривични пријави против ...	41.2031	22.5756
10	2	Гевгелија		ПС Гевгелија поднесе кривична пријава против Ј.Д....	41.15	22.51
11	2	Куманово		СВР Куманово поднесе кривична пријава против А.М....	42.1334	21.7258
12	2	Богданци		ПС Гевгелија поднесе кривична пријава против С.Б....	41.2031	22.5756
13	1	Скопје	с. Ржаничино	На 21.06.2011 година, во 19,00 часот, во продав...	42.0038	21.4522
14	1	Куманово	ул.„Веселин Маслеша"	СВР Куманово поднесе кривична пријава против В.М....	42.1334	21.7258
15	0	Охрид		ПС Прилеп поднесе кривична пријава против Е.Џ. (3...	41.1216	20.8194
16	2	Скопје		На 21.06.2011 година, во ПС Гази Баба, К.А., одго...	42.0038	21.4522
17	6	Велес	ул.„Ѓоре Органџиев"	СВР Велес поднесе кривична пријава против П. Т. (...	41.7148	21.7846
18	1	Куманово		СВР Куманово поднесе кривична пријава против М.А....	42.1334	21.7258
19	2	Прилеп		ПС Прилеп поднесе кривична пријава против З.А. (4...	41.3517	21.5621
20	6	Битола		СВР Битола поднесе кривична пријава против В.П. ...	41.0328	21.3403
21	0	Охрид		ПС Струга поднесе кривична пријава против Н. У. (...	41.1216	20.8194
22	1	Прилеп	с. Кривогаштани	ПС Прилеп поднесе кривична пријава против Г.Ч. (3...	41.3358	21.3331
23	0	Скопје	с. Горно Лисиче	На 24.06.2011 година, СВР-Скопје, со кривична при...	41.9617	21.5076
24	1	Штип		На 24.06.2011 година М.С.(1962) од Штип, во СВР-...	41.75	22.2
25	1	Скопје		Ноќта помеѓу 23 и 24.06.2011 година, во периодот...	42.0038	21.4522
26	1	Скопје	ул.„Црвена Општина"	На 24.06.2011 година околу 15,40 часот во Скопје...	42.0038	21.4522
27	2	Куманово		На 24.06.2011 година СВР-Куманово до ОЈО-Куманов...	42.1334	21.7258
28	6	Тетово	с. Жеровјане	На 24.06.2011 година во 23,00 часот во с.Жеровја...	41.9146	20.9484
29	2	Штип		СВР-Штип поднесе кривична пријава против Г.Д.(34...	41.75	22.2
30	2	Штип		Од страна на СВР-Штип по превземени оперативно ...	41.75	22.2

Figure 4: A view of the database (text in Macedonian).

A sample of the database and the detected crime events can be seen on Fig.4. We used the WHERE, the latitude and the longitude for determining the location, the WHAT for the category of crime, and put the full description as a tooltip text on the marker on the map, as demonstrated on Fig.5, Fig.6 and Fig.7.

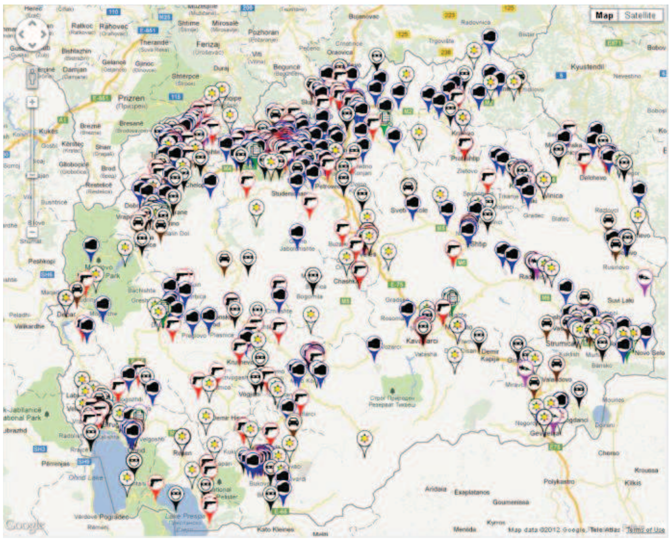
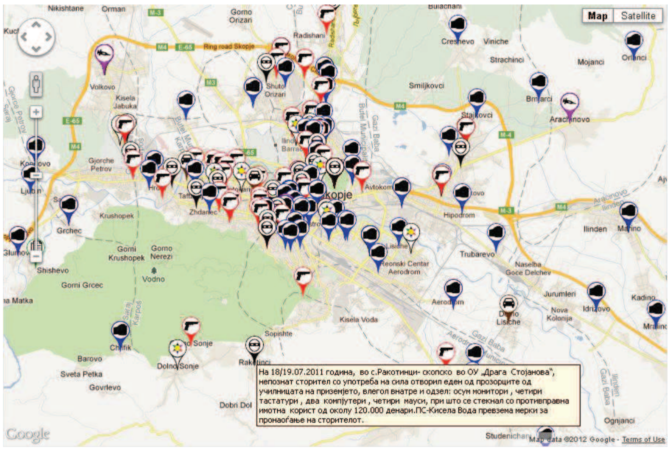


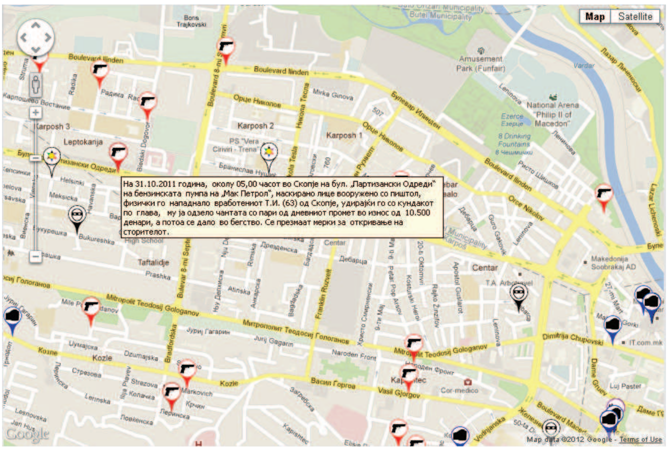
Figure 5: The crime map for the entire country.

Our project is a web application, for which we used PHP as a server-side programming language, with MySQL as a database engine and HTML and JavaScript for the presentation layer. The web application automatically updates the database with new information from MOI's website.



На 18/19.07.2011 година, во с.Ракотинци- скопско во ОУ „Драга Стојанова", непознат сторител со употреба на сила отворил еден од прозорците од училницата на приземјето, влегол внатре и одзел: осум монитори , четири тастатури , два компјутери , четири мауси, при што се стекнал со противправна имотна корист од околу 120.000 денари.ПС-Кисела Вода превзема мерки за пронаоѓање на сторителот.

Figure 6: The crime map for Skopje, the capital of Macedonia.



На 31.10.2011 година, околу 05,00 часот во Скопје на бул. „Партизански Одреди" на бензинската пумпа на „Мак Петрол", маскирано лице вооружено со пиштол, физички го нападнало вработениот Т.И. (63) од Скопје, удирајќи го со кундакот по глава, му ја одзело чантата со пари од дневниот промет во износ од 10.500 денари, а потоа се дало во бегство. Се презмаат мерки за откривање на сторителот.

Figure 7: The crime map for the neighborhood near our Faculty.

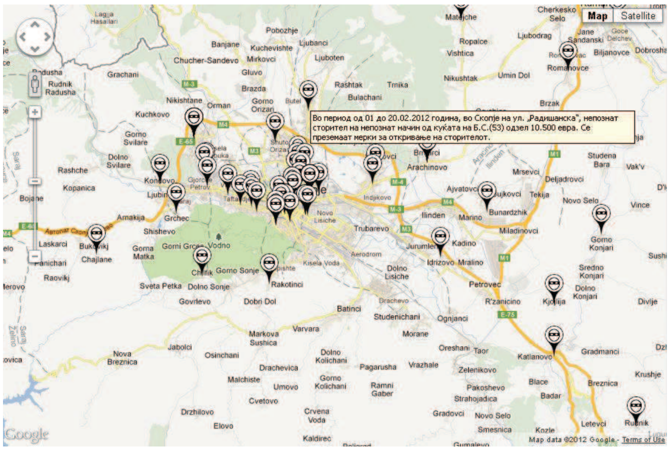
We developed an interactive user interface allowing the user to filter the events by the category of crime they belong to, or by location, as shown on Fig.8. This is done by sending an AJAX request to a page which returns a JSON formatted response with the latitude and longitude values, a description and the category of each event that satisfies the particular filter condition.

A key issue when working with data in Macedonian language is the character encoding. A good practice is to use the international Unicode encoding system, which encodes not only Latin and Cyrillic, but all the possible alphabets and symbols. We used UTF-8 as it proves to be the best and most popular encoding [20].

C. *Encoding Challenges*

When dealing with strings and substrings, all programming languages work with indexes or character positions. We pay a lot of attention to this, because PHP's byte level string division and comparison allows us to see the full effect of UTF-8.

Many people misbelieve that UTF-8 encodes symbols in 8 bit (1 byte) arrays. This is not the case. In fact, that answer is closer to the number of bits the American ASCII uses - which is 7. On the other hand, UTF-8 works with octets. The most commonly used characters, such as the ones in ASCII are encoded with one octet and are compatible with the ASCII codes. The less used characters, such as most letters in the Macedonian alphabet, are encoded with 2 octets (2 bytes) and therefore take up 2 index-position spaces in PHP's substring function. The Macedonian quotation marks „ and ", being even rarer, are encoded with 3 bytes [21].



Во период од 01 до 20.02.2012 година, во Скопје на ул. „Радишанска", непознат сторител на непознат начин од куќата на Б.С.(53) одзел 10.500 евра. Се преземаат мерки за откривање на сторителот.

Figure 8: A filtered view of the crime map, showing only theft and burglary crimes.

D. *Geocoding*

Google Maps is a web mapping service application and technology provided by Google. It has become popular due to its precision and correctness, as well as the fact that it is free for public use. In 2005 the people at Google launched a free API to allow developers to integrate Google Maps into their websites. We have used the newest version (v3) of the JavaScript API (there are other ways to use Google Maps not necessarily with JavaScript) [22].

Google allow users a free account key with certain limitations. For instance, the key can be used to access a map only 25,000 times a day, after which Google will block your access or allow you more for a fee if you prefer.

Google Maps, as well as all other mapping services, process map locations with their longitude and latitude values. This is why we keep this information within out database. There is a way to get these values (usually shortened as *latlng*) from an address. This process is called geocoding [23].

Google Map's geocoding can be applied on client-side, or on server-side. The former does not stress the server as much, but is limited to about 10 a second, so in order to get the latlng values, we used the latter. It implies sending the address in an URL to a Google server and waiting for a response formatted in XML, or in JSON as in our case. The limitations for this service are 2500 per IP address per day, which at this time seems enough for the application.

It is only fair to mention that there are many other web mapping services, most notably Microsoft's Bing Maps, OpenStreetMap and Yahoo Maps. They are all useful and function in a similar manner as Google Maps, but we decided that Google Maps offers the best service for our project at this point.

IV. Conclusion and Future Work

In conclusion, we can say that the Open Data initiative is starting to develop in Macedonia, but there's plenty more room for improvement and at least the computer science community wants to see it happen. In addition, crime mapping is a very useful tool in crime analysis and in fighting crime, but these maps can also be connected and overlaid with open data from education, income salaries, etc., in order to pinpoint the major cause for high percentage of crime in some regions.

Another benefit is that the citizens can take into consideration the crime percentage of the region where they plan to buy their future home, where their children will go to school, or where their work is situated.

However, there are some privacy issues that should be taken into consideration, such as reporting full personal names or exact locations. In the case of crime mapping, the events can be summarized on a point of public interest, like a cinema or a park, for example.

We hope that this project will encourage the local government to be more open towards the public and the computer science community and we hope that this project will be the dawn for the Macedonian Open Data initiative and openness in general.

References

[1] The website of The International Council for Science's World Data System, offering quality-assessed data and data services to the international science community: http://www.icsu-wds.org/

[2] *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007) http://www.oecd.org/document/55/0,3343,en_2649_201185_38500791_1_1_1_1,00.html

[3] *Towards a Science Commons* Creative Commons http://sciencecommons.org/about/towards, http://creativecommons.org/science

[4] *Free Our Data* The Guardian's Technology http://www.freeourdata.org.uk/index.php

[5] The official local crime and policing website for England and Wales: www.police.uk

[6] The website of the US The Omega Group's crime map: www.crimemapping.com

[7] The website of the GIS developed by the US nonprofit community – The Reinvestment Fund: www.policymap.com

[8] The American open-data project Data.gov: www.data.gov

[9] The British open-data project Data.gov.uk: www.data.gov.uk

[10] The Norwegian open-data project Data.norge.no: www.data.norge.no

[11] The Russian open-data project Opengovdata.ru: www.opengovdata.ru

[12] List of open government data catalogues from around the world http://opengovernmentdata.org/data/catalogues/

[13] The website of the Open Government Partnership: http://www.opengovpartnership.org/, http://www.opengovpartnership.org/countries/macedonia

[14] Boba, Rachel (2005). *Crime Analysis and Crime Mapping*. Sage Publications

[15] Maltz, Michael D.; Gordon, Andrew C.; Friedman, Warren (2000) [1990]. *Mapping Crime in Its Community Setting: Event Geography Analysis* New York: Springer-Verlag. ISBN 0-387-97381-8.

[16] Jackson, Peter; Moulinier, Isabelle (2002). *Natural language processing for online applications : text retrieval, extraction, and categorization pp.2-3* John Benjamins B.V. ISBN 1567-8202; v . 5

[17] Witten, Ian H; Frank, Eibe (2005). *Data Mining : practical machine learning tools and techniques* Morgan Kaufmann. ISBN 0-12-088407-0

[18] Manning, Christopher D.; Schutze Hinrich (1999). *Foundations of statistical natural language processing* MIT Press.ISBN 0-262-13360-1

[19] The official website of the Macedonian Public Prosecutor's Office (the DA's office): http://www.jorm.org.mk/zakon-krivicen.shtml

[20] A website showing UTF-8 usage: http://trends.builtwith.com/encoding/UTF-8

[21] The official RFC for UTF-8: RFC 3629 / STD 63 (2003)

[22] The Google Developers website: https://developers.google.com/

[23] Mazumdar S, Rushton G, Smith B *et al.*. *Geocoding accuracy and the recovery of relationships between environmental exposures and health.* International Journal of Health Geographics (2008)