Baltic J. Modern Computing, Vol. 5(2017), No.3, 295-305
http://dx.doi.org/10.22364/bjmc.2017.5.3.04

Automated Analysis of the Content of Selected Open Access Internet Sources as a Tool for Government Decision Making

Vladislav V. FOMIN[1,2,3], Tomas KRILAVIČIUS[2], Vytautas MICKEVICIUS[4], Daiva VITKUTĖ-ADŽGAUSKIENĖ[2], Aušra MACKUTĖ-VARONECKIENĖ[2], Rita VALTERYTĖ[2], Tomas TAUGINAS[5], Dominykas VERŠINSKAS[6], Egidija VERŠINSKIENĖ[6], Evaldas BRUŽĖ[6]

[1]Vytautas Magnus University, Lithuania
[2]Turība University, Latvia
[3]University of Latvia, Latvia
[4]Baltic Institute of Advanced Technologies
[5]Mykolas Romeris University, Lithuania
[6]Lithuanian Cybercrime Centre of Excellence for Training

vvfomin@gmail.com, tomas.krilavicius@vdu.lt, vytautas.mickevicius@gmail.com, daiva.vitkute-adzgauskiene@vdu.lt, ausra.mackute-varoneckiene@vdu.lt, rita.valteryte@vdu.lt, tomas.tauginas@mru.lt, dominykas@l3ce.eu, egidija@l3ce.eu, evaldas@l3ce.eu

Abstract: In this paper we report on the development of the prototype of Internet media monitoring tool for Lithuanian government. Two design specificities are emphasized. First, the tool must to a maximum possible extent utilize the open access and open source tools and resources available. Second, this university-lead open mode of the tool development must be conducted in close collaboration with governmental agencies operating under confidentiality seal. Having successfully developed the media monitoring prototype, two key findings are reported: 1) the conceptual model of the Internet media monitoring tool based on open access Internet infrastructure resources; and 2) the system design method for balancing public and confidential requirements towards the system.

Keywords: prototype, Internet media, monitoring, analysis, open-source, government decision-making.

1. Introduction

Today we witness close dependencies between economic and political developments of a particular country on the one hand, and the information dissemination activities of its

strategic allies or counterparts on the other hand. While such media resources as TV and newspapers are conventionally used to promote a political cause or point of view, they are limited to local or national coverage. Internet has become a domain where any geopolitical entity can promote its own causes or views, while at the same time those views can be countered or balanced by information provided by other entities. The world-wide discussions on the role of media in presidential elections in the USA, for example, witness that the role of Internet media will be the focus of scholarly research for time to come (Green, 2016; Lawler et al., 2016).

A small county by any measure, Lithuania has a number of substantial geopolitical dependencies with respect to neighboring domains –European Union countries in the west and the countries of the Commonwealth of Independent States (CIS) – in the east. Success of economic and political partnership, as well as the development of national market, is highly dependent on the assessment of and timely adaptation to economic and political changes in those neighboring domains.

To a large extent, the information needed for that kind of assessment and continuous monitoring of the state of affairs in the partner countries, is available on the Internet from open access sources. Official information on Internet news portals is often supplemented by valuable comments, discussions, and opinion forums. However, the volume of available information, the frequency with which the information is being updated, render monitoring and analytical tasks ineffective or even impossible if performed by human agents manually.

In order to tackle with the information monitoring and assessment problems, automated analysis of information from open access sources can be performed using Information and Communication Technology (ICT) tools. On the one hand, a number of tools for Internet media monitoring has been developed by educational, governmental, and private entities. On the other hand, the government of any particular country sets its own unique goals and methods in monitoring media of partner countries, which often leads to specialized technological solutions. Use or adaptation of extant generic solutions for media monitoring is often problematic due to language specificities, unique national interests, as well as trade and state security requirements imposed by the given country.

In this paper we report on the development of the prototype of Internet media monitoring tool for Lithuanian government. Two specific requirements had to be satisfied in the development of the prototype. First, the tool had to a maximum possible extent utilize the open access and open source tools and resources available[1]. Second, this public university lead tool development process had to be conducted in close collaboration with governmental agencies operating under confidentiality seal.

Having developed the prototype, two findings are reported: 1) the conceptual model of the Internet media monitoring tool based on open access Internet infrastructure resources; and 2) the Information System (IS) design method for balancing public and confidential requirements of the involved stakeholders.

2. Political and technological background

The global discussion sparkled around the 2016 U.S. elections (Lawler et al., 2016), and many other global developments widely covered by international press, demonstrate that

[1]Such as developed under university-lead projects and consortia for public use.

governments should not underestimate the potential impact of the Internet media on political actions of citizens and governments (journalistsresource.org, 2016).

Each government seeks to minimize the possible (negative) impact of external to the state subjects on the internal development processes. To succeed in this task, the governments should be able to identify the "impact groups", the prominent past and present events, reactions to which can have impact on present and future geo-political processes in the country (see Fig.1).Continues monitoring of the media, identification of prominent events and reactions to them with a potential to affect the future developments, can help government generate foresight on the possible outcomes and develop informational or other appropriate counter measures for preserving its long-term strategic plans.

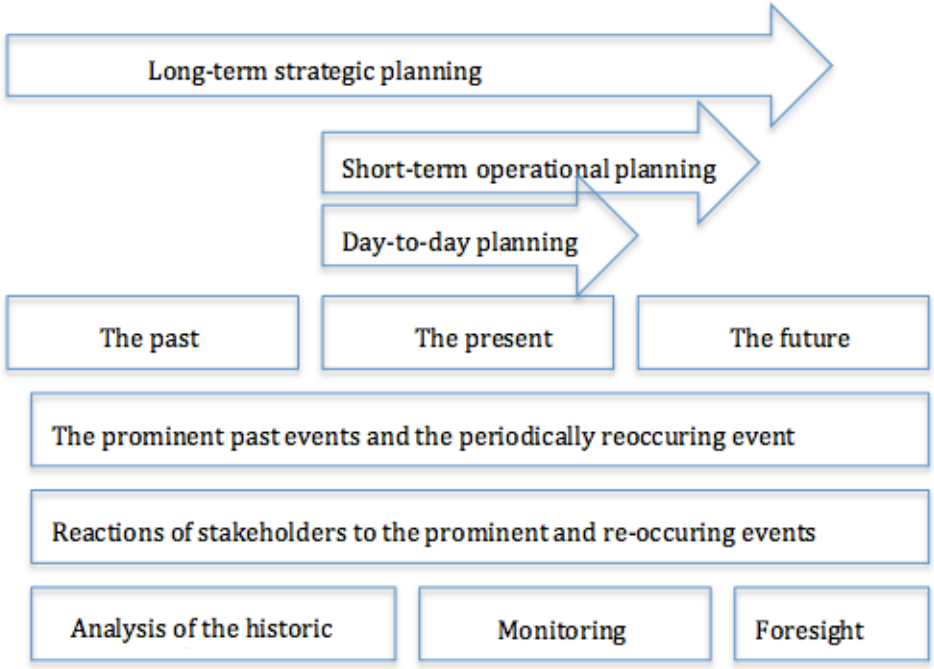

Fig. 1. Governmental decision making as an ongoing process.

Continuous monitoring of media resources, especially the Internet, presents a serious challenge. The volume of information available on the Internet, and the frequency with which it is being updated, the speed with which is being disseminated – all factors combined render manual monitoring and analysis ineffective or even impossible. Human's job of written text analysis must be delegated to computer-based tools.

To date, there are many computerized language analysis tools available, with varying degree of sophistication and specificity. A class of tools, which can be referred to as platforms offers broad spectrum of functionality and can serve as infrastructure to specific user-developed software via different Application Programmable Interfaces (APIs). Among most popular systems are GATE[2] and UIMA[3], where the latter facilitates the analysis of not only text, but also audio and video.

[2]General Architecture for Text Engineering. https://gate.ac.uk

Among popular tools with narrower functionality, are systems for natural language processing such as Stanford CoreNLP[4], LingPipe[5], a platform for building Python programs to work with human language NLTK[6], Apache OpenNLK[7], and other. Most of them are easy to integrate (or are already integrated) with the aforementioned GATE and UIMA platforms.

Complementing the aforementioned infrastructural and open text- and language-analysis systems, there is a variety of specialized proprietary solutions, such as Palantir[8], OntoText[9], Orunmila[10]. Among tools specifically tailored for social media data analysis can be named Crimson Hexagon (CH)[11] and Meltwater[12].

European Commission has developed its own tool for media monitoring and analysis - Europe Media Monitor (EMM)[13], which, among other functionalities, allows users to see the major news stories in more than 20 languages for any specific day and to compare how the same events have been reported in the media written in different languages.

At the backdrop of a sufficiently broad variety of extant technological solutions for text analysis, it must be emphasized that those tools exist either as "development platforms" or "open projects", or as specialized, fixed functionality proprietary tools. Such dichotomy implies that, in the case of proprietary solutions, the user must be satisfied with the functionality offered by the developer, or else in the case of open systems – the specific user requirements can only be met with a substantial programming and tinkering efforts.

The highly dynamic nature of geo-info-political environment, as we witness today, brings forth a requirement of having possibility for frequent change of criteria and variables for media monitoring and analysis, which becomes a serious (financial) obstacle in opting for tailored proprietary solutions, such as e.g., Citer 360[14]. If opting for one of the extant open-source solutions for media monitoring, the confidentiality seal of strategic governmental decision-making becomes a serious obstacle, unless the government can afford to maintain own pool of programmers (which, again, besides the substantial cost also implies organizational burden).

The aim of this project was to find a compromise solution by developing a design method, which would, on the one hand, give government authorities access to rich open-source language analysis resources, but on the other hand allow for (frequent) customizations to match confidential (and changing) requirements of the strategic decision-making without creating a financial strain.

It may be argued, that extant solutions, such as e.g., European Commission's media monitoring tool EMM, can fulfill the aforementioned requirements. EMM can follow several preselected media outlets in different languages and allow performing news

[3]Unstructured Information Management Architecture. https://uima.apache.org
[4]http://stanfordnlp.github.io/CoreNLP
[5]http://alias-i.com/lingpipe
[6]http://www.nltk.org
[7]http://opennlp.apache.org
[8]https://www.palantir.com
[9]http://ontotext.com
[10]http://www.tokenmill.lt
[11]http://www.crimsonhexagon.com
[12]https://www.meltwater.com/uk/
[13]http://emm.newsexplorer.eu
[14]http://www.tm-group.com/products/nice-security-portfolio/citer-360/

analysis to support strategic decision-making. However, such tools as EMM are based on in-house developed proprietary software. The aim of the reported project, on the contrary, was to work with less popular language(s), to have the solution based on open source software, and to propose a framework, which can be easily extended to any language and information source. In addition, our proposed solution works not only with news outlets, but also with social media, e.g. comments, forums, social networks (with specific scanners).

2.1. The open language Internet infrastructure in Lithuania

"Open language infrastructure" is a contemporary concept referring to the development of language tools and resources – dictionaries, translators, thesaurus, etc. – for open Internet access and use. For the so-called "small language" countries like Lithuania,[15] "open language infrastructure" helps mobilize scarce computational resources and developers' efforts needed to tackle the growing popularity of English as de-facto Internet language and the resulting negative impact of the local language (Fomin et al., 2012). During 2012-2015, under the program "Lithuanian Language for Information Society"[16] researchers of two major Lithuanian universities - Vytautas Magnus university and Kaunas University of Technology – developed the National Language Information Infrastructure (system) abbreviated as "LKSAAIS" (Utka et al., 2016; Vitkutė-Adžgauskienė et al., 2016). LKSAAIS consists of two sub-systems (see

Fig. **2**. National Language Information Infrastructure (LKSAAIS)). The first sub-system forms the basis of the system. It contains different modules (tools) for a number of linguistic analyses. The second sub-system of the developed infrastructure is the user layer. It contains few exemplarily tools (applications) for the public use. The open access free tools for the public use include Lithuanian Internet media analysis, semantic search tool, and annotation tool, among other.

The LKSSAIS services are offered in both user-machine (through the Internet web sites' user interface) and machine-machine (as web services) modes (Vitkutė-Adžgauskienė et al., 2016). Prior to this project, a number of European Commission's or Lithuanian government funded projects were focused on Internet Infrastructure development for Lithuanian and other languages, which resulted in creation of tools for language analysis, and the creation National Language Infrastructure and establishment of B-level Centre of CLARIN Consortium.[17]

One of the distinctive goals of the project was to develop specific (and evolving) tool for media monitoring based on the available open Language Infrastructure tools.

[15] Lithuanian language is native to less than 3 million inhabitants of Lithuania. Together with Latvian language native to less than 2 million in Latvia, it belongs to Baltic language group of Indo-European languages.

[16] Project Nr. VP2-3.1-IVPK-12-K-01-007 „Syntax-Semanthic Analysis System for Lithuanian Language Texts, Internet, and for Public Sector Use."

[17] http://clarin-lt.lt. See examples of services at www.semantika.lt, www.rastija.lt, and www.epaslaugos.lt.

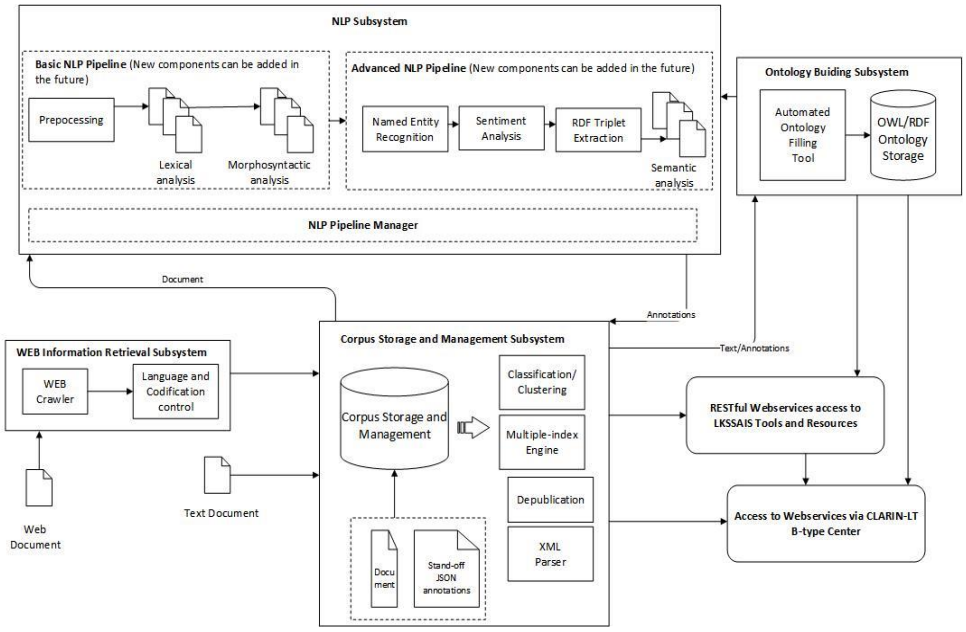

Fig. 2. National Language Information Infrastructure (LKSAAIS)

3. Design of the tool

3.1. Design method

Internet Media Monitoring Tool (IMMT) prototype was developed under the project funded by Lithuanian Science Council responding to request from Lithuanian government.

Specific requirement of the project was to create a conceptual model of the Tool and its prototype based on already available open infrastructure tools and resources for language analysis, such as the aforementioned LKSSAIS.

The general concept of the functionality of the Tool was to provide Internet media monitoring results to Lithuanian government officers in charge of economic and political decisions. Having witnessed the growing importance of Internet media in international development context, especially the effects of propaganda and trolling conducted over regular news services (and through readers' commenting on news), it was assumed that intensity of news and reader discussions on specific topics can inform government officers on the vector of political and economic developments in the country, as well as in its neighboring countries, which can affect economic and political developments in Lithuania.

The stakeholders of the Tool development process were a number of Lithuanian government agencies. The project group included researchers and specialists from two Lithuanian universities and a non-governmental organization Lithuanian Cybercrime Center of Excellence for Training, Research and Education "L3CE" focused on cybercrime studies prevention (education, development). The latter also took the role of intermediary between the university researchers and the government agencies.

Within the project duration of 1 year, the development plan anticipated two stages – requirements solicitation and the proof-of-concept development. Requirements solicitation was done in continuous discussion between the stakeholders (through regular meetings, visits to stakeholder premises, analysis of extant tools, etc.). Altogether, during the development stage from September 2015 until August 2016, over 30 meetings were held, including the research group visit to Joint Research Centre overseeing the development of EMM, a number of presentations at different government offices, university meetings, etc.

3.2. Functionality of the tool

With regard to media sources to be processed and analyzed, the Internet Media Monitoring Tool (IMMT) was agreed to process open news resources in Lithuanian and Russian languages, as their associated reader forums.[18] Facebook[19], Twitter[20] and VKontakte[21] were analyzed using external tools and integrating results.

The overall design of the Tool is presented in Fig. 3 below. The data analysis performed by the Tool consisted of 4 distinctive steps:

1. Data collection – targeted scanning of selected news portals/sites with or without specification of keywords.
2. Preprocessing – boilerplate removal, part-of-speech tagging (POS), etc.
3. Semantic analysis of the data:
    a. Named Entity Recognition (NER);
    b. Sentiments.
4. Clustering and classification (Naïve Bayes, SVM) for event identification:
    a. Retrospective analysis (past events according to given keywords);
    b. New event identification (based on data content).

In developing the Tool, two requirements had to be satisfied - utilization of the available open access and open source tools and resources to a maximum possible extent, and balancing the public (university-lead) nature of the development with the confidentiality seal imposed by the government on specific functionality of the Tool.

Regarding the public infrastructure reuse requirement, the functionality of the Tool was implemented through the use of LKSSAIS REST web services of Lithuanian National Language Information Infrastructure (see

Fig. **3**). The same web services can also be accessed via CLARIN-LT repositories. The following components of the Tool are linked with the LKSSAIS infrastructure:

1. Component of information collection (scanning):

[18] The following news sites were included in the prototype: lrytas.lt, delfi.lt, gazeta.ru.
[19] www.facebook.com
[20] www.twitter.com
[21] vk.com

1.1. LKSSAIS resources were used for Lithuanian language;
1.2. in-house scanner was developed for Russian language.

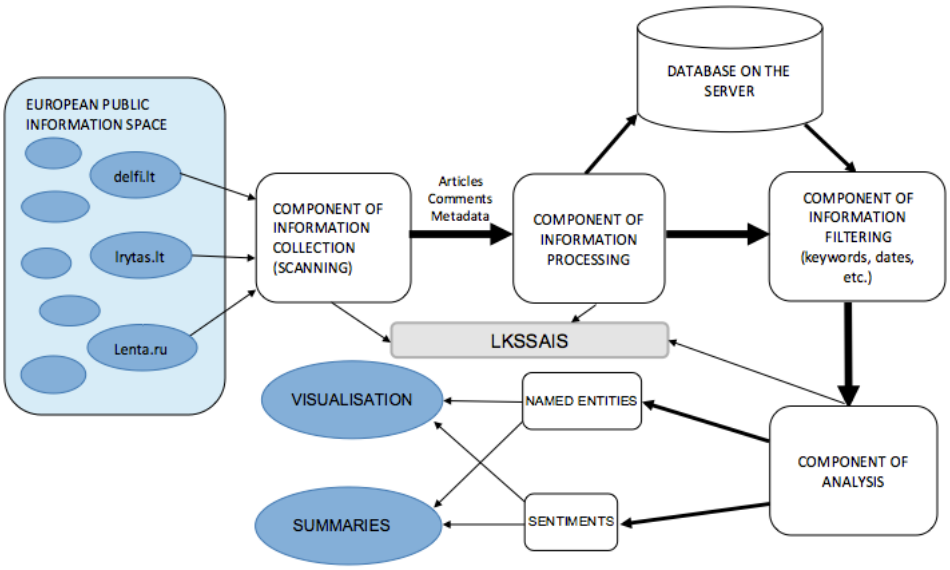

Fig. 3. Internet Media Monitoring Tool (IMMT)

2. Component of information processing:
   2.1. LKSSAIS resources were used for Lithuanian text preprocessing;
   2.2. open resources were used for Russian language preprocessing.
3. Component of information filtering was developed in-house using R libraries.
4. Component of analysis:
   4.1. LKSSAIS resources were used for Lithuanian text analysis;
   4.2. in-house developed routines and open resources were used for Russian text analysis.

The requirement to balance opens (public) and closed (confidential) interests in the development process presented a substantial challenge to both the developers and the government and contributed to a high degree of complexity in this project.

Any software development work in contemporary organizations requires inter-disciplinary expertise to achieve a complex synthesis of specialized knowledge domains of involved stakeholders. In what Boland and Tenkasi (Boland et al., 1995) referred to as iterative loops of "perspective making" and "perspective taking", specific to the reported project, development work required iterative exchange of views, ideas and requirements between the two pools of stakeholders. This additional challenge to the of process of "distributed cognition" (Boland and Tenkasi, 1995) was successfully tackled by developing a specific working pattern.

The specific working pattern of the project stakeholders was such that the university representatives, using their expert knowledge of LKSSAIS, would set up a tool according to general requirements of the group working under confidentiality seal. The latter would configure the tool to meet their specific requirements, test the specific

functionality, and provide feedback to the university group – either requiring amendments to the tool or signaling acceptance of the tool (see Fig. 4).

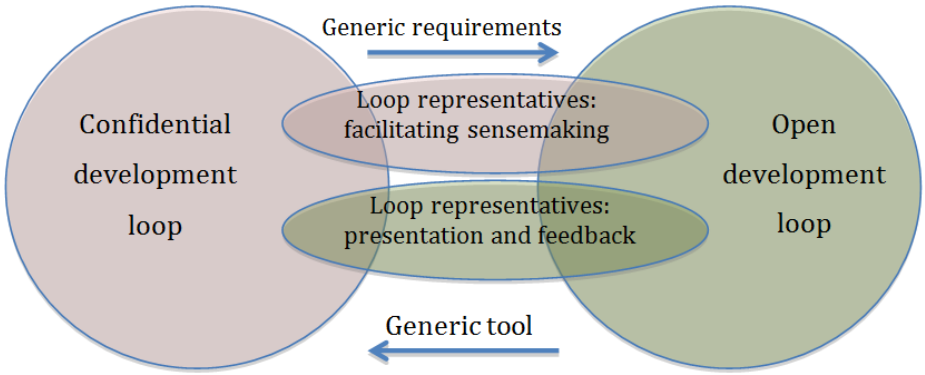

Fig. 4. Interaction of open and closed loops during prototype development process

Collaboration between the two groups was facilitated by intermediaries, who had government granted access to confidential information. These intermediaries would help establish meaningful exchange of ideas and requirements in the context, when certain requirements cannot be disclosed.

4. Conclusions and further developments

Having developed a prototype for open access Internet media monitoring, the project group had confronted a number of issues. Some of them were successfully resolved – e.g., the use of open access National Language Internet Infrastructure tools, the resolution of other requires regulatory action – e.g., the regulation on media archiving.

The two requirements brought forward to the developers were satisfied –the Tool utilizes open access resources and can be maintained and further developed by public entity (university), while specific functionality of the Tool remains under the government‘s confidentiality seal.

One of the important findings of the project was that the imperfect nature of Internet media imposes new challenges to the task of media monitoring the governments did not face in the age of printed media, as explained below.

Monitoring of Internet media presents considerable technical and organizational challenges to the monitoring agency. At least three problems must be addressed when developing a media monitoring tool: accessibility of historic media, assessment of historical validity of information, and the frequency and volume of data being processed. Of the three challenges, only the latter can be tackled by technological genius of the developer. The two former ones present more serious challenges requiring amendments to the national or international legislation.

The problem of accessibility of historic media is related to the need of media providers to optimally use own resources while delivering high speed access to the media. This dilemma is usually solved by archiving or removing older information.

Public access to the archived information becomes restricted, often resulting in fee-based access or removal of information. This means that access to historical information – while important for governmental decision-making process (as shown in

Fig. **1**) – can become limited or unavailable for media monitoring purposes.

Additional obstacle for historical data access is the fact that search engines organize information search and retrieval based on proprietary indexing methods. One particular function of the proprietary indexing methods is data filtering, which means that search engine user is obtaining historical data not in its totality, and cannot determine whether or not (and which) pieces of information are missing.

The problem of validity of information is manifested in media archiving practices. Periodical mass media is being archived according to relevant national regulation, which imposes rules for media archiving in national archives. Archived printed media cannot be altered, i.e. the archived historical information is intact and "valid". In the case of digital media, information providers can modify historical information at their discretion. Information can be altered or deleted altogether. In such circumstances, access to (the original) historic information is also distorted, which, if done on purpose, corrupts the very concept of mass media as a reliable mirror of past events and related societal discussions.

For digital media, especially for information provided by third parties, national regulation for archiving does not exist. This means that there is no possibility to assess validity of claims with regard to historical media information. In the context of increasing influence of fake news and propaganda (Jackson, 2017), this situation should urge the national governments issue regulatory acts on media archiving and ensure that the regulation is being followed, or even take the role of national archive for a selected number of media resources (public and private).

Finally, the volume of (daily) information being published on the Internet is extremely high. The volume is even higher if one considers not only the published information, but also associated reader comments, forums, blogs, etc. The volume of the information to be processed renders the manual processing impossible, unless data is filtered to dramatically reduce the volume, which would inevitably lead to the subjectivity of analysis results, up to the misinformation. Objective analysis of information, accordingly, requires computerized automated solutions capable of acquiring and processing high volume of complete data sets from the selected sources.

To summarize, while the legislative issues pertaining to media archiving were outside of the competence pool of the project team, technical and organizational requirements were addressed successfully.

A final note should be given on the use of "small languages". Language specific resources on the Internet are mostly available for such languages as English. To achieve required quality of language analysis for such languages as Lithuanian or Russian, additional resources are needed. In the case of Lithuanian language, most of them can be based on the results of LKSAAIS project and other open access/open source developments (e.g., TokenMill language pack).[22] The existence of open National Language Internet Infrastructure helps leverage creation of new language analysis tools and services, which, in turn, help grow the National Infrastructure and hence tackle the "small language" problem.

[22] https://github.com/tokenmill

5. Acknowledgements

This work was supported by Lithuanian Science Council, grant Nr. REP-09/2015.

We are thankful to Lithuanian Ministry of Defense, Lithuanian Ministry of Interior, Gerhard Wagner of the EC Joint Research Centre's IPSC GlobeSec TP267 for support and guidance in developing the tool.

References

Boland, R. J. Jr., and Tenkasi, R.V. (1995). Perspective Making and Perspective Taking in Communities of Knowing. *Organization Science* 6: 350–72.

Fomin, V., Laužikas, R., Vitkutė-Adzgauskiene, D., Vaitkevičius, V. (2012). Building Knowledge Society in Lithuania – towards „heritage-Aware" National Information Infrastructure. *Transformations in Business and Economics* 11 (2(26)): 180–201.

Green, R. K. (2016). The Game Changer: Social Media and the 2016 Presidential Election, November 16. http://www.huffingtonpost.com/r-kay-green/the-game-changer-social-m_b_8568432.html.

Jackson, D. (2017). Trump: Russia Collusion in Election Story Is 'Fake News.' *USA Today*, March 20. http://www.usatoday.com/story/news/politics/2017/03/20/donald-trump-barack-obama-wiretapping-james-comey-russia/99401374/.

journalistsresource.org. (2016). The Arab Spring and the Internet: Research Roundup, January. https://journalistsresource.org/studies/international/global-tech/research-arab-spring-internet-key-studies.

Lawler, D., Henderson B., Allen, N., Sherlock, R.. (2016). US Election: Donald Trump Claims Victory despite Media Consensus That Hillary Clinton Won First Presidential Debate, September 28. http://www.telegraph.co.uk/news/2016/09/26/donald-trump-and-hillary-clinton-to-face-off-in-first-us-preside/.

Utka, A., Amilevičius, D., Krilavičius, T., Vitkutė-Adžgauskienė, D. (2016). Overview of the Development of Language Resources and Technologies in Lithuania. In *Human Language Technologies–The Baltic Perspective*. Vol. 289. Riga, Latvia: IOS Press. http://hlt2016.tilde.eu.

Vitkutė-Adžgauskienė, D., Utka, A., Amilevičius, D., Krilavičius, T. (2016). NLP Infrastructure for the Lithuanian Language. In *LREC 2016 Proceedings* , 2539–42. Portorož, Slovenia. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1070_Paper.pdf.

Received June 12, 2017, revised September 12, 2017, accepted September 18, 2017