MINING AND AGGREGATION OF CULTURAL DATA IN SOCIAL NETWORKS

Evi Papaioannou[1]*, Elpida Schiza[2]

[1]University of Patras and CTI "Diophantus", GREECE, papaioan@ceid.upatras.gr
[2] University of Patras, GREECE, elpida_sxiza@hotmail.gr
*Corresponding Author

Abstract

The explosive growth of the Internet, the emergence of social networks and recent technological advances enabled an enormous user population to become actuators in this new emerging cultural environment. Handheld wireless devices, like smartphones and tablets, which can be internet-connected, allow users to join the Internet community from any place at any time. Users are of various and diverse cultural profiles. Social networks form a modern global environment where all these users can actually become cultural actuators in the sense that they socialize, communicate, announce and reproduce information promoting local, national and international activities closely related to their cultural background.

Modern social networks, like Facebook or Twitter, form active and vivid channels of cultural information circulation. Thousands of single users or user groups make frequent announcements about special cultural events, related to music, dance, theater, cinema, gastronomy, performances, exhibitions, gatherings of a special cultural character. In addition, such announcements made in the form of short, inclusive posts bear unique online features so that their audience can immediately exploit them. However, it remains an important challenge to efficiently mine useful data from such populated, diverse and vaguely structured spaces.

Motivated by the case of Santorini Island, Greece and a strong recent observation that local traditional activities or special (multi-)cultural events and activities tend to be absent from touristic guides and plans, we present a WordPress-based website which automatically collects cultural data from Facebook and presents it in a comprehensive way for promoting cultural activity in Santorini.

Lack of information implies lack of knowledge which consequently results in a reduced interest and decision space. Utilizing keywords spanning a variety of cultural activities and events, our system serves as an aggregator for Facebook posts of particular cultural interest. While several, mainly not collaborating, entities – like for instance Facebook users or groups, websites, Twitter users or groups - do release this sort of information, lack of organization and timely viewing makes it extremely inefficient for interested entities to locate, evaluate and exploit this highly distributed and unstructured material.

The experimental use of our system so far – as an application offered from the Department of Cultural Heritage Management and New Technologies of the University of Patras – shows that technology can indeed serve an important role towards efficient cultural management and fruitful intercultural cooperation.

Keywords: culture, data mining, social networks

1 INTRODUCTION

The explosive growth of the Internet, the emergence of social networks and recent technological advances enabled an enormous user population to become actuators in this new emerging cultural environment. Handheld wireless devices, like smartphones and tablets, which can be internet-connected, allow users to join the Internet community from any place at any time. Users are of various and diverse cultural profiles. Social networks form a modern global environment where all these users can actually become cultural actuators in the sense that they socialize, communicate, announce and reproduce information promoting local, national and international activities closely related to their cultural background (Kidd, 2008, Furedi, 2014, Sawyer, 2011).

Santorini, located in the southern Aegean Sea, is the southern member of the Cyclades group of islands. Apart from its natural beauty, Santorini has a rich and longstanding cultural environment. However, the existing rich pool of standard touristic information in the Internet regarding Santorini fails to capture and highlight several aspects of its cultural environment. For example, a simple web search would return thousands of posts or advertisements for items that mainly promote economic activity on the island like fancy coffee shops, restaurants, clubs, hotels, etc. Nevertheless, there is an important lack of filtered and organized information sets about events and activities that reflect the actual cultural environment of Santorini as it has been maintained and evolved through the centuries.

We mine the Internet and modern social networks for information about various cultural events all around Santorini which include thematic activities in archaeological sites and museums, exhibitions, cultural events and activities that maintain and promote the local cultural character of Santorini like fairs, gastronomy, wine-making, construction workshops, meetings of cultural groups and events that highlight the cultural history and tradition of the island as it is shaped and reflected within a modern multi-cultural setting.

Towards this aim, we utilize modern social networks which have emerged during recent years over the Internet and tend to form a modern, vivid cultural canvas (Easley, Kleinberg, 2010). Exploiting a mixture of web technologies accommodated in the WordPress platform, like html, php, sql, and Facebook, we automatically mine cultural data for Santorini and present it in a comprehensive way. We present Santorini culture miner, a WordPress-based web aggregator which currently runs at the webserver of the Department of Cultural Heritage Management and New Technologies of the University of Patras, Greece. Our work and evaluation activities so far imply that technology can indeed serve an important role towards efficient cultural management and fruitful intercultural cooperation.

The rest of the paper is structured as follows: in Section 2 we discuss the potential of modern, internet-based social networks to form modern cultural environments. In Section 3, we provide a high-level description of our system together with its technical characteristics and implementation details. In Section 4, we present current evaluation results. We conclude in Section 5 where we also outline our vision and future plans.

2 MODERN SOCIAL NETWORKS AS CULTURAL ENVIRONMENTS

The notion of a network usually implies a structure which includes an interconnected set of items and involves some sort of exchange or interconnection among these items. Typically, a network can be defined as a set of nodes and links where there is at least one link between two nodes as long as these nodes are related according to some predefined manner. For example, a transportation network is a structure emerging from the interconnection among different places of interest which involves the exchange of people or goods among them. A communication network is a structure emerging from the interconnection among communication devices which involves the exchange of messages i.e., data, among them. A social network is a structure emerging from the interconnection among social actors - individuals or organizations - and involves some sort of actual social interaction among them. In the Internet age, the term “social network” directly points to software platforms like Facebook, Twitter, MySpace, etc, which act as virtual societies: they do emerge from interconnections among actors but, now, both the actors and their interconnections are virtual and powered by an underlying, global communication network, i.e., the Internet. In either case, “the social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures” (Wasserman, Faust, 1994).

Social networks have played a very influential and critical role in almost every aspect of human socio-economic and political activity and, therefore, in human culture (Easley, Kleinberg, 2010, Furedi 2014). In particular, modern social networks, like Facebook or Twitter, form active and vivid channels of cultural information circulation. Thousands of single users or user groups, simply using electronic devices like smartphones, tablets, notebooks and an active Internet connection, can make frequent announcements about special cultural events, related to music, dance, theater, cinema, gastronomy, performances,

exhibitions, gatherings of a special cultural character. In addition, such announcements made in the form of short, inclusive posts bear unique online features so that their audience can immediately exploit them. However, it remains an important challenge to efficiently mine useful data from such populated, diverse and vaguely structured spaces.

We decided to focus on Facebook, since it is a very popular social network, with millions of subscribers all over the world offering a wide range of applications and communication services to its members. Facebook, which started as a Harvard social-networking website, has emerged to a global internet phenomenon (Phillips, 2007). Facebook was launched in 2004 with approximately 1.500 Harvard students as its initial subscribers and soon enough extended beyond educational institutions to anyone with a valid email address. We indicatively mention that according to the online statistics portal "statista" (https://www.statista.com/) "in the third quarter of 2012, the number of active Facebook users had surpassed 1 billion". Facebook can be accessed over the Internet and mobile networks by a wide range of computing devices like smartphones, tablets, laptops and desktop computers.

Through their Facebook profile, singe users or user groups can instantly, at almost no cost, inform their followers for activities or events in every place of the globe. Focusing on cultural information circulation, it is easily seen that Facebook, or any other similar internet-based social network, forms a modern, multicultural virtual environment where, ideally, all participants have equal opportunities to promote their own culture and receive fruitful influence from other cultures. Without the existence of the Internet and the Web and in the absence of modern social network platforms such information spreading would be infeasible or would suffer from severe locality and financial constraints and limitations. Indicative examples of cultural information exchange in Facebook include multimedia material regarding virtual tours in distant cities and monuments therein, food recipes, local festivals, traditional activities and customs, or even information regarding the everyday life of geographically or mentally distant populations. Such interaction provides knowledge and therefore promotes freedom, respect and solidarity in a multicultural environment.

3 OUR WORK

Motivated by the case of Santorini Island, Greece and a strong recent observation that local traditional activities or special (multi-)cultural events and activities tend to be absent from touristic guides and plans, we present a WordPress-based cultural miner for Santorini which automatically collects cultural data from Facebook and presents it in a comprehensive way for promoting cultural activity in Santorini.

Lack of information implies lack of knowledge which consequently results in a reduced interest and decision space. Utilizing keywords spanning a variety of cultural activities and events, our system serves as an aggregator for Facebook posts of particular cultural interest. While several, mainly not collaborating, entities – like for instance Facebook users or groups, websites, Twitter users or groups - do release this sort of information, lack of organization and timely viewing makes it extremely inefficient for interested entities to locate, evaluate and exploit this highly distributed and unstructured material.

3.1 High-level description of Santorini cultural miner environment

We have used WordPress platform and a set of assisting tools, technologies and methods for building a dynamic online environment which works as an aggregator for cultural information for Santorini. Our environment is currently available at the webserver of the Department of Cultural Heritage Management and New Technologies of the University of Patras, Greece at the following URL: http://culturalminer.culture.upatras.gr

Our basic component, namely the "Santorini cultural miner", automatically collects and presents in a unified way public Facebook posts of single users and groups for a wide range of cultural activities and events in Santorini. Data mining is performed according to a small initial set of keywords which guide the data collection process. We have exploited online services offered by Google, like Google maps and Google automatic translation, for increasing the friendliness and usability of our environment to visitors. In particular, instead of providing long, detailed, geographical verbal descriptions for points of interest, we directly locate them on a map thus making information globally usable avoiding linguistic constraints. In addition, we provide an automatic online translation for posts and information originally edited in some particular language. Automatically translated text suffers from inefficiencies of automatic translation, i.e., violated syntactic rules, inaccurate terms, etc. However, it certainly provides fresh and comprehensive cultural information which could be ignored otherwise due to linguistic constraints and limitations.

3.1.1 Structure and design

Our environment is structured as a four-part panel which includes the header section, the footer section, the

main body section and the side bar. The main body section is actually the interface for accessing our home page and a five-item menu. As depicted in Fig. 1, the five-item menu together with our logo is placed into the header section. Each menu item is then linked to content about the development team (About), a contact functionality (Contact us), the actual Cultural miner and additional static content (A cultural view of Santorini) which is hosted in the main body section.

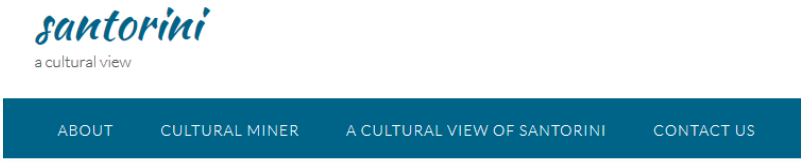

Fig. 1: the header section.

The footer section includes functionalities for the retrieval of recent posts, archiving and traffic statistics (see Fig. 2).

Fig. 2: the footer section.

The side bar (see Fig. 3) contains a calendar, functionalities for the representation of cultural information on the map of Santorini, search and automatic translation, as well as an rss-based frame where selected online cultural information is presented.

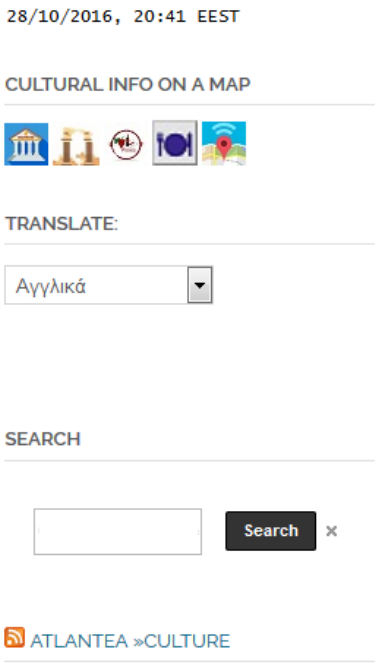

Fig. 3: the side bar.

Using a small initial set of seeds, i.e., keywords for cultural events and activities and the Custom Facebook Feed plugin, Santorini cultural miner automatically collects information from public Facebook posts and announcements of single users and groups and presents it in the form of a list. This list is constantly updated with fresh material placed at its top. Old posts and announcements are archived on a monthly basis and are accessible through an archive menu included in the footer section. In Fig. 4, an indicative snapshot of the cultural miner content is depicted.

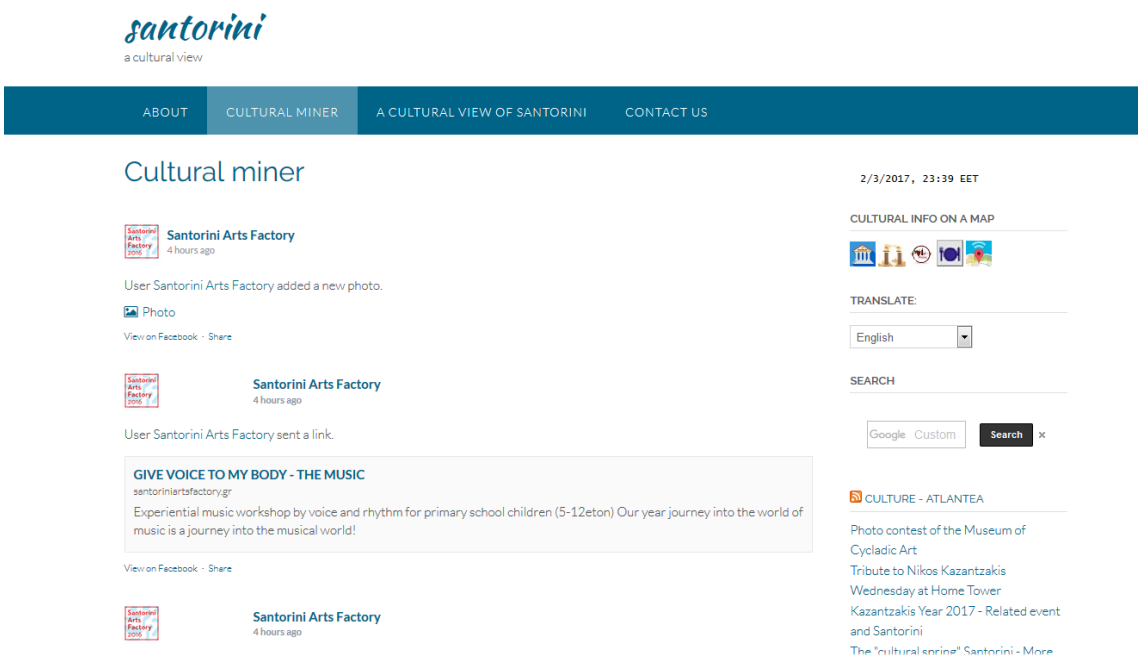

Fig. 4: a snapshot from Santorini cultural miner.

In Fig. 5, a snapshot of the content of “A cultural view of Santorini” is depicted. This item is an additional stand-alone Facebook aggregator collecting cultural information for Santorini from a list of predefined sources.

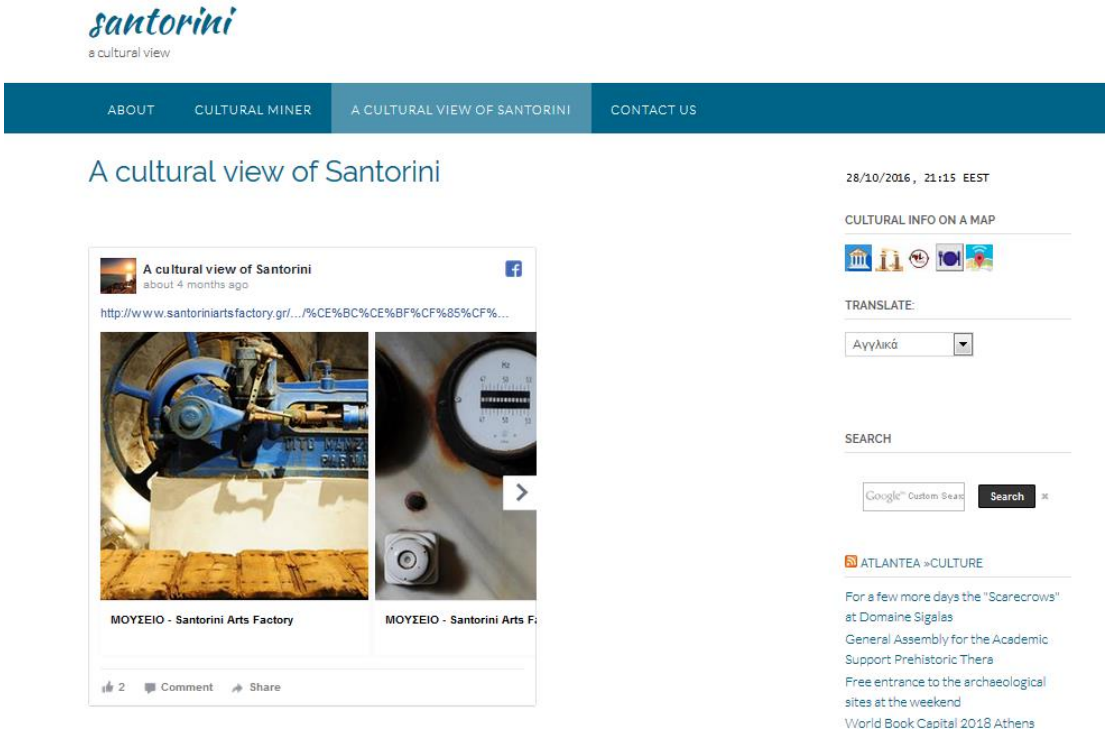

Fig. 5: a snapshot from “A cultural view of Santorini”.

3.1.2 Mobile Web compatibility

All parts and details of our environment are also accessible via mobile web (see Fig. 6). As part of our initial design, our implementation follows standards for mobile web applications and meets corresponding constraints. Mobile web compatibility receives additional significance given than the vast majority of our targeted audience will normally have web access through mainly handheld devices like tablets or smartphones rather than laptops or desktop computers.

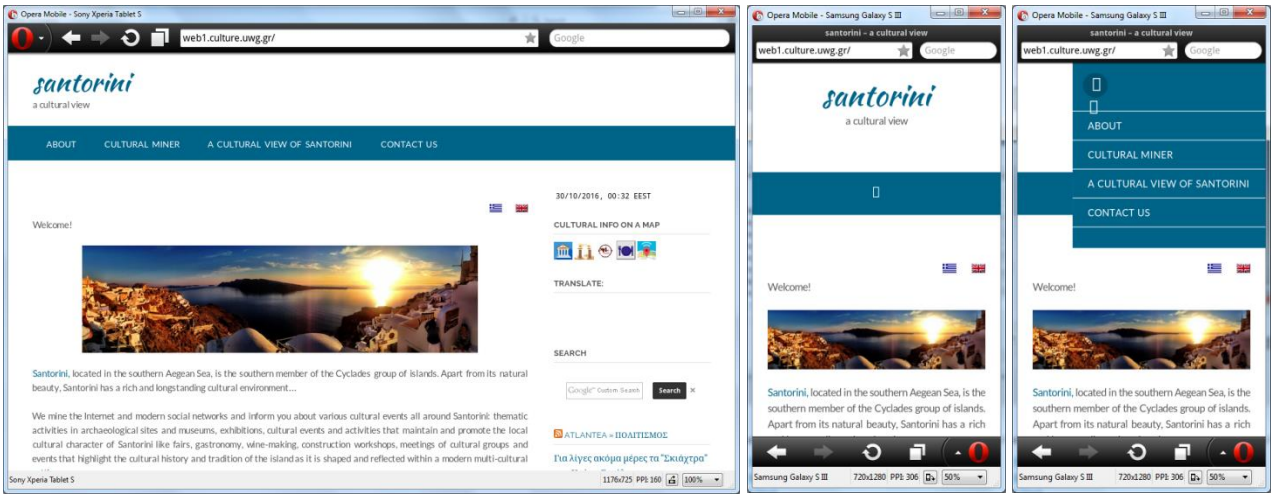

Fig. 6: Sony Xperia Tablet S and Samsung Galaxy S III indicative previews via Opera Mobile Classic Emulator.

3.2 Technical description of Santorini cultural miner environment

In this section, we provide technical details regarding the design and implementation of our environment.

3.2.1 WordPress

We have used WordPress as our main development platform (Corbin 2010, Starr, Coyier, 2009). WordPress, first released in 2003, is a free open source web software for developing websites, blogs and web applications built and updated by a wide, global community of volunteers. Based on PHP and MySql, WordPress soon emerged to a full content management system (CMS in short) supported by a rich list of additional plugins, widgets and themes (https://wordpress.org/) extensively used by the web community. We preferred WordPress instead of some other relevant CMS, like for example Joomla!, Drupal, etc., since it tends to be the most popular – both from a user and a developer point of view – and easy-to-use CMS for structured content management in the web. According to a recent web technology survey by W^{3}Techs on the usage of content management systems for websites released in October 2016, "WordPress is used by more than 26.7% of the top 10 million websites". Furthermore, BuiltWith (https://trends.builtwith.com/cms) reports WordPress as the most popular blogging system in use on the Web supporting more than 300 million websites out of a total of approximately 17 billion.

For developing our environment, we used the current 4.6.1 WordPress version which, as official recommendations suggest, works well with PHP 5.2.4 or greater and MySQL 5.0 greater. Currently, our environment is installed and running on the webserver of the Department of Cultural Heritage Management and New Technologies, University of Patras which is an Apache webserver running PHP 5.2.6 and a MySQL client version 5.0.51a.

We further enriched our WordPress basic installation by exploiting a number of additional plugins (Corbin 2010, Starr, Coyier, 2009). Apart from plugins related to appearance and formatting, the catalogue of our most important plugins includes:

- *Anti-spam*: with more than 100.000 active installs, it provides protection against spam comments and, thus, protects our environment and the webserver hosting it from malicious traffic.
- *Automatic Plugin Updates*: with more than 7.000 active installs, it enables automatic background updates of plugins excluding selected plugins from being automatically updated.
- *Custom Facebook Feed*: with more than 200.000 active installs, it can be used for displaying a customizable, responsive and search engine crawlable version of Facebook feeds from multiple different Facebook pages and groups. Making the appropriate modifications, both on our client and server sides, we use this plugin as a basic tool for our cultural miner.
- *Google Language Translator*: with more than 90.000 active installs, it offers fast and relatively accurate automatic translation also providing configuration options for layout and style as well as language selection. Our thorough investigation and testing on a long list of available plugins for automatic translation suggests that for our purposes Google Language Translator achieves a very

good performance ratio as far as complexity of installation and use and quality of obtained translation are concerned.

In addition, we used JavaScript technology for adding interactivity to our environment. We indicatively refer to our script implementing the advanced search functionality included in the side bar (see Figure 5).

3.2.2 Google maps

Instead of providing long, detailed, geographical verbal descriptions for points of interest, we directly locate them on a map thus making information globally usable avoiding linguistic constraints.

We extensively use Google maps (https://maps.google.com/) in order to present points of cultural interest, like archaeological sites, museums, traditional wineries, restaurants (see Fig. 7). Furthermore, in order to facilitate internet connection we provide information regarding Wi-Fi hot spots. In this way, mined cultural and networking information can be efficiently exploited by visitors of different linguistic background.

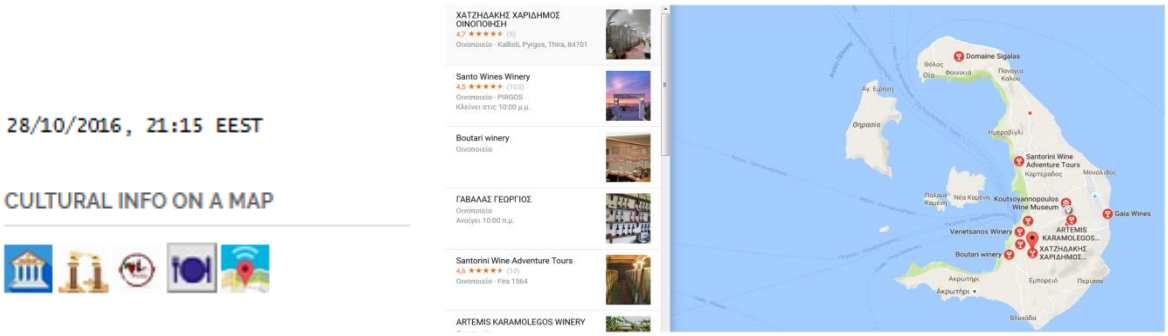

Fig. 7: points of cultural interest marked on the map of Santorini together with further related information.

3.2.3 Google online automatic translation

An important constraint that visitors have to face in a foreign-language environment is that while plenty of cultural information is available, yet, this information is not actually exploitable. A usual approach involves the offline creation of alternative versions of relevant material, e.g., websites, leaflets, signs, etc, in different languages. However, such an approach can work well for rather static information with low update rate. For instance, it is definitely functional to provide a printed touristic guide created offline in several languages, but this process becomes completely inefficient when it comes to small pieces of information produced online, with a usually short exploitation lifetime, like an announcement or a post for an event.

For limiting the negative effect of linguistic constraints, we use Google Language Translator (Shankland, 2013) in order to provide an automatic online translation for posts and information originally edited in some particular language. This particular plugin is based on the online use of Google Translate (https://translate.google.com/), a service launched by Google initially in 2006 for rule-based machine translation and subsequently updated in 2009 for multilingual statistical machine translation. We decided to use Google Language Translation plugin instead of several other plugins for language translation, like for example WPML, qTranslate, Polylang, Tansposh, Weglot, etc. Three main factors supported our decision. First, the - mainly - online and dynamic nature of our material suggested that we focus our investigation on translation plugins which could work well for online translation of several, simply-structured, short statements like posts or short announcements. So, we decided to leave out plugins which either require an additional site version per language or provide satisfactory translation for offline material. Second, the lack of available budget for purchasing a translation plugin suggested that we avoid plugins requiring any sort of paid subscription. Third, the complexity of installing and configuring a translation plugin should definitely not exceed the corresponding complexity of our overall implementation.

An important observation (Aiken, Ghosh, Wee, Vanjani, 2009) is that automatically translated text suffers from known inefficiencies of automatic translation, i.e., violated syntactic rules, inaccurate terms, etc. However, it certainly provides fresh and comprehensive cultural information which could be ignored otherwise due to linguistic constraints and limitations (Aiken, Balan, 2011).

3.2.4 JotForm

For contact purposes we use an online contact form (see Fig. 8). We created this form online and published it using the form builder software provided for free by JotForm (https://www.jotform.com/). Visitors can

communicate with us by including their message in the contact form together with personal information, like name, surname and email. The captcha field ("Completely Automated Public Turing test to tell Computers and Humans Apart") at the bottom of the contact form determines whether or not the sender is human (von Ahn, Blum, Hopper, Langford, 2003), thus avoiding messages from automated malicious processes and agents. Using this particular form allows us to receive email notifications for responses and also facilitates data collection and processing.

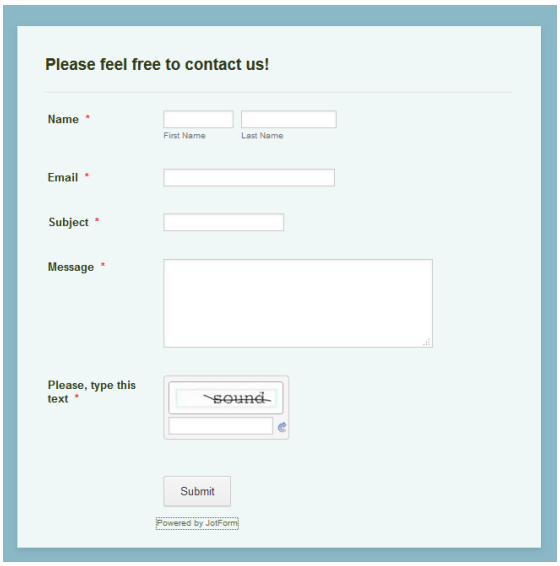

Fig. 8: Out contact form maintained at JotForm platform.

We decided to use JotForm online platform instead of simply implementing our own PHP contact form from scratch mainly due to security constraints imposed on our webserver side. Firewalls and other security policies (like for example virtual private networks) adopted and implemented on our webserver side would make it completely inefficient to handle email communication internally, despite the use of appropriate anti-spam software within our environment.

4 EVALUATION

Current evaluation of our environment is mainly based on the use of web analytics. We trace our internet presence via the online TraceMyIP.org free service (https://www.tracemyip.org/). We collect and study visitor analytics exploiting information of all inbound connections to our website (see Fig. 9).

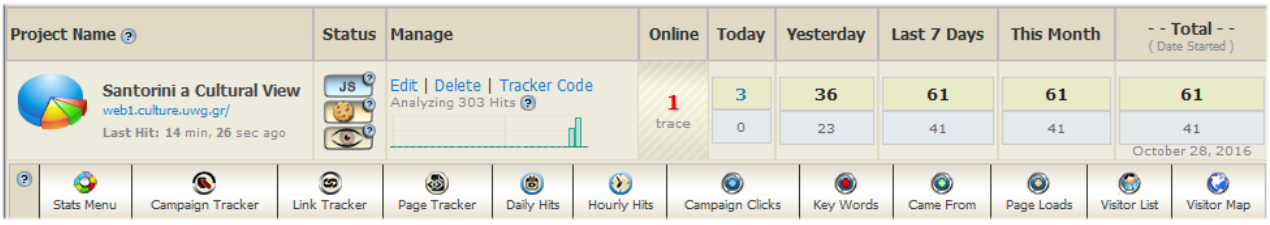

Fig. 9: Information of all inbound connections provides a wide range of visitor analytics.

We track information like for example individual IP activity, traffic sources, page tracking and popularity, visitor platform (see Fig. 10) and geographical information (see Fig. 11).

Santorini a Cultural View [edit]

Web Browsers
Real-Time Traffic Statistics

Percentage	Visitors	Browser Name
51.61 %	16	Chrome
22.58 %	7	Safari
16.13 %	5	Mozilla Firefox
3.23 %	1	Internet Explorer
3.23 %	1	Opera Mobi
3.23 %	1	Opera

Results 1 to 6 (returned 6) from 303 log records

Santorini a Cultural View [edit]

Operating Systems
Real-Time Traffic Statistics

Results 1 to 10 (returned 19) from 303 log records

Percentage	Visitors	Operating Systems
15.15 %	5	Windows 7
12.12 %	4	Windows 10.0
9.09 %	3	iOS 10.1
9.09 %	3	Android 4.4.2
6.06 %	2	Android 6.0
6.06 %	2	iOS 10.0.2
6.06 %	2	Android 4.4.4
3.03 %	1	Windows Vista
3.03 %	1	Windows XP
3.03 %	1	Android 5.0.2

Fig. 10: Real-Time Traffic Statistics: visitor web browsers and operating systems

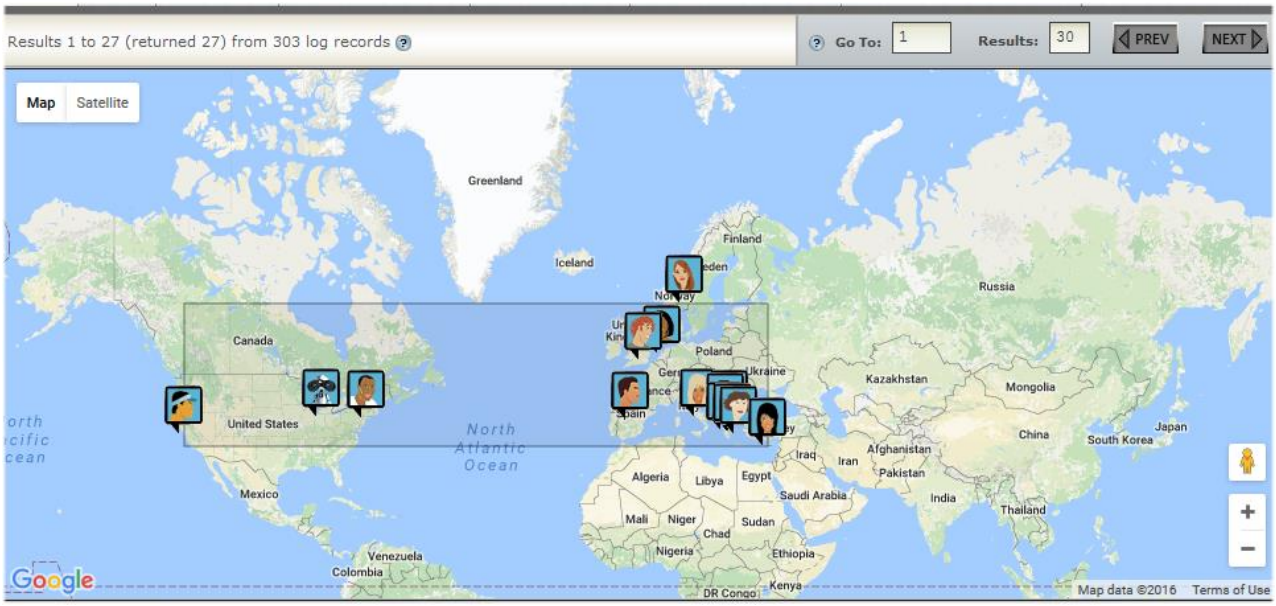

Fig. 11: A snapshot produced using the Real-Time Website Geographical Visitor Tracker

However, current evaluation is conducted mostly in vitro. We believe that transferring our environment to a more powerful webserver which can efficiently serve higher request rates and traffic will enable its exploitation as a real pilotic system for Santorini and, thus, will provide sufficient evidence for its evaluation under real circumstances of operation and use.

5 CONCLUSION AND FUTURE PLANS

Motivated by the case of Santorini Island, Greece and a strong recent observation that local traditional activities or special (multi-)cultural events and activities tend to be disregarded or even absent from touristic guides and plans, we designed and developed a WordPress-based environment which automatically collects cultural data from Facebook and presents it in a comprehensive way for promoting cultural activity in Santorini.

While several, mainly not collaborating, entities – like for instance Facebook users or groups, websites, Twitter users or groups - do release this sort of information, lack of organization and timely viewing makes it extremely inefficient for interested entities to locate, evaluate and exploit this highly distributed and unstructured material. Utilizing keywords spanning a variety of cultural activities and events, we presented a simple web aggregator for Facebook posts of particular cultural interest. To the best of our knowledge, no other similar environment has appeared so far either, in general, for cultural purposes or, in particular, for the case of Santorini.

Our future plans include a refined mechanism for mining cultural data so that information resulting from the interconnection between profiles and posts is also efficiently retrieved. We plan to address in detail the automatic translation part of our work in order to provide a more involved, though similarly efficient translation process for the collected content of our miner. Our challenging major objective is the evolution of our currently centralized system into a collaborative environment where single users can contribute to the creation of a web-based multicultural data mining environment for further promoting efficient cultural management and fruitful intercultural cooperation.

ACKNOWLEDGMENTS

We would like to thank Christina Dimopoulou, Giorgos Fragkogiannis, Giorgos Moisiadis and Dora Monioudi-Gavala for their valuable help and support.

REFERENCE LIST

Aiken M., Balan S. (2011). An Analysis of Google Translate Accuracy. *Translation Journal*, vol. 16, Issue 2.

Aiken, M., Ghosh, K., Wee, J., Vanjani, M. (2009). An Evaluation of the Accuracy of Online Translation Systems. *Communications of the IIMA*, vol. 9, Issue 4, pp. 67-79.

BuiltWith (2016). CMS Usage Statistics: Statistics for websites using CMS technologies. https://trends.builtwith.com/cms

Carlson, N. (2010). At Last—The Full Story of How Facebook Was Founded. *Business Insider*.

Corbin B. (2010). WordPress Top Plugins. *Packt Publishing*. ISBN-13: 978-1849511407

Easley D., Kleinberg J. (2010). Networks, Crowds, and Markets: Reasoning About a Highly Connected World. *Cambridge University Press*. ISBN-10: 0521195330.

Furedi, F. (2014). How The Internet and Social Media Are Changing Culture. Aspen Review, no. 4. *Aspen Institute Prague*.

Kidd, T. T. (2008). Social Information Technology: Connecting Society and Cultural Issues. *Information Science Reference, IGI Global*. ISBN 978-1-59904-774-4.

Phillips, S. (2007). A brief history of Facebook. *The Guardian*.

Sawyer, R. (2011). The Impact of New Social Media on Intercultural Adaptation. DigitalCommons@URI.

Shankland S. (2013). Google Translate now serves 200 million people daily. *CNET*.

Starr, J., Coyier, C. (2009). Digging Into WordPress. *Self Published*. https://digwp.com/book/

von Ahn, L., Blum, M., Hopper, N. J., Langford, J. (2003). CAPTCHA: Using Hard AI Problems for Security. In Proceedings of the 2003 International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2003), pp. 294-311.

W^{3}Techs Web Technology Surveys (2016). Usage Statistics and Market Share of Content Management Systems for Websites. https://w3techs.com/techfeed/survey/content_management

Wasserman, S., Faust, K. (1994). Social Network Analysis in the Social and Behavioral Sciences. *Social Network Analysis: Methods and Applications*, pp. 1–27, Cambridge University Press. ISBN 9780521387071.