Extraction of Environmental Data from On-Line Environmental Information Sources

Stefanos Vrochidis, Victor Epitropou, Anastasios Bassoukos, Sascha Voth, Kostas Karatzas, Anastasia Moumtzidou, Jürgen Mossgraber, Ioannis Kompatsiaris, Ari Karppinen, Jaakko Kukkonen

Extraction of Environmental Data from on-line Environmental Information Sources

Stefanos Vrochidis[1], Victor Epitropou[2], Anastasios Bassoukos[2], Sascha Voth[3], Kostas Karatzas[2], Anastasia Moumtzidou[1], Jürgen Moßgraber[3], Ioannis Kompatsiaris[1], Ari Karppinen[4] and Jaakko Kukkonen[4]

[1] Centre for Research and Technology Hellas, Informatics and Telematics Institute
[2] Informatics Systems and Applications Group, Aristotle University of Thessaloniki
[3] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
[4] Finnish Meteorological Institute, Helsinki

stefanos@iti.gr, vepitrop@isag.meng.auth.gr, abas@isag.meng.auth.gr, sascha.voth@iosb.fraunhofer.de, moumtzid@iti.gr, kkara@eng.auth.gr, juergen.mossgraber@iosb.fraunhofer.de, ikom@iti.gr, ari.karppinen@fmi.fi, jaakko.kukkonen@fmi.fi

Abstract. Analysis of environmental information is considered of utmost importance for humans, since environmental conditions are strongly related to health issues and to a variety of everyday activities. Despite the fact that there are already many free on-line services providing environmental information, there are several cases, in which the presentation format complicates the extraction and processing of such data. A very characteristic example is the air quality forecasts, which are usually encoded in image maps of heterogeneous formats, while the initial (numerical) pollutant concentrations, calculated and predicted by a relevant model, remain unavailable. This work addresses the task of semi-automatic extraction of such information based on a template configuration tool, on methodologies for data reconstruction from images, as well as on Optical Character Recognition (OCR) techniques. The framework is tested with a number of air quality forecast heatmaps demonstrating satisfactory results.

Keywords. Environmental, air quality, heatmap, image processing, OCR, data reconstruction, template configuration.

1 Introduction

Analysis of environmental information is considered of utmost importance for human population, as this is strongly related to health issues (e.g. cardiovascular diseases), as well as to a variety of important activities (e.g. agriculture). In everyday life, environmental conditions of the atmospheric environment, in terms of air quality,

weather, pollen measurements and forecasts are also of particular interest for outdoor activities (e.g. trip planning) and therefore they strongly affect the quality of life. Nowadays, the main sources of such information for the everyday user are web portals and sites. In order to support people in everyday action planning considering the environmental conditions, we need to provide them with services, which combine complementary environmental information from several resources, with a view to generate more reliable environmental measurements. The first step towards this direction is the extraction of data from environmental resources. In practice only a few of the data providers make available some means of access to their actual (numerical) forecast data. In this context, this paper addresses the semi-automatic extraction of air quality forecasts from heatmap images.

After studying a number of on-line chemical weather forecasts by various providers [1], it can be said that the air quality information is most usually presented in the form of images representing forecast pollutant concentrations over a geographically bounded region, typically in terms of maximum or average air pollution concentration values for the time scale of reference, which is usually the hour or day [2], [3], [4], [5], [6]. These providers present their air quality forecasts almost exclusively in the form of preprocessed images with a color index scale indicating the concentration of pollutants. In addition, these providers arbitrarily choose the resolution of their images, the color scale and color depth employed for visualizing pollution loadings, the covered region, as well as the geographical map projection. The actual mode of presentation varies from simple web images to more elaborated AJAX, Java or Adobe Flash viewers [7]. While this representation is informative for the casual user (e.g. compared to a table with numerical values), it has the drawback that the data are being presented in a wide range of highly heterogeneous forms, which makes it very complicated to extract and compare their results. To make it worse, some of the images are permanently marked with visible watermarks, text, lines etc. that would make the extraction phase even more challenging.

In order to address this challenge we propose a semi-automatic framework for extracting air quality information from such images and store them into a numerical format. The proposed system is based on an annotation tool, which supports an administrative user to generate a configuration template for each heatmap, and on Optical Character Recognition (OCR) techniques for text information extraction. The basic functionality of the system (i.e. the information extraction from heatmaps), is based on AirMerge [4], [6], [8], [9], a system that allows for the automatic harvesting, annotation, harmonization and reverse engineering of heatmaps, in order to come up with easily deployable numerical values of chemical weather forecasts.

The contribution of this paper is the methodology and the framework for user-assisted air quality information extraction from heatmaps, which extends previous works (i.e. AirMerge) by further adding OCR techniques, as well as allowing user configuration with the aid of a dedicated graphical user interface. More specifically, we propose a framework, which is based on a novel heatmap Annotation Tool (AnT), on the application and optimization of OCR techniques for textual information extraction from heatmap images and on the AirMerge tool [4] for image processing.

This paper is structured as follows: section 2 presents the related work, while section 3 describes the framework architecture. Section 4 presents the Annotation Tool, section 5 the OCR techniques and section 6 the AirMerge system. The results are presented in section 7 and finally, section 8 concludes the paper.

2 Related Work

Existing maps can be grouped into map types based on the placement and presentation of their information. Discriminating factors between map types can be found in their scale, colorization, quality, accuracy, topology and many other aspects. In case of air quality (or chemical weather) maps there are mainly two types of information covered by the map data: a) Geographical information: points and lines describing country frontiers or other well-known points of interests or structures (e.g. sea, land, capitals) in a given coordinate system, b) Color information: measured data of any kind (e.g. average temperature), which are coded via a color scale representing the measured values. Single values are referenced geographically by a color value at the corresponding geographical point. Chemical weather maps often use this type of maps called raster map or heatmaps images to represent measured data. There are several approaches to extract and digitalize this image information automatically.

First, the authors in [10] describe the process of the vectorization of digital image data. Hereby the geographical information, in form of lines, is extracted and converted to digital storable vector data. Only the lines are processed. The work in [11] makes use of the specific knowledge of the known colorization in USGS maps, to have the ability to automatically segment these maps based on their semantic contents (e.g. roads, rivers). In [12] the segmentation quality of text and graphics in color map images is improved, to enhance the results of the following analysis processes (e.g. OCR), by selecting black or dark pixels from color maps, cleaning them up from possible errors or known unwanted structures (e.g. dashed lines), to get cleaner text structures.

Although research work has been conducted towards the automatic extraction of information in maps, very few works address the automatic extraction of information from chemical weather maps. In such works [4], [6], [8], a method to reconstruct environmental data out of chemical weather images is proposed. In a first step the relevant map section is scraped from the chemical weather image. After that disturbances are removed (e.g. country lines) and a color classification is employed to classify every single data point (pixel), to recover the measured data. With the aid of the known geographical boundaries, given by the coordinate axis and the map projection type, the geographical position of the measured data point can be retrieved. In case of missing data points, a special interpolation algorithm is used to fill these gaps.

The proposed work goes one step beyond the aforementioned heatmap extraction methods, since it introduces a configurable user-assisted environment, which facilitates the application of the framework on new heatmaps without requiring programming skills and low level configuration on the user's side.

3 Framework Architecture

The architecture of the proposed framework is illustrated in Figure 1 and includes two main components: the Annotation Tool and the data extraction service.

The first phase, called "Template Configuration" (1→2→3), includes the manual annotation of an image with the AnT, and the generation of a configuration file. This process is controlled by an administrative user with the aid of AnT. The second phase includes the "Data extraction" (1+3→4→5), which uses the configuration file to extract data from the specific heatmap. During this phase, the parts of each image are analyzed using image and text processing techniques. Specifically, the heatmap is processed with the AirMerge system, while the text information located in the image is extracted and processed using OCR techniques and text processing.

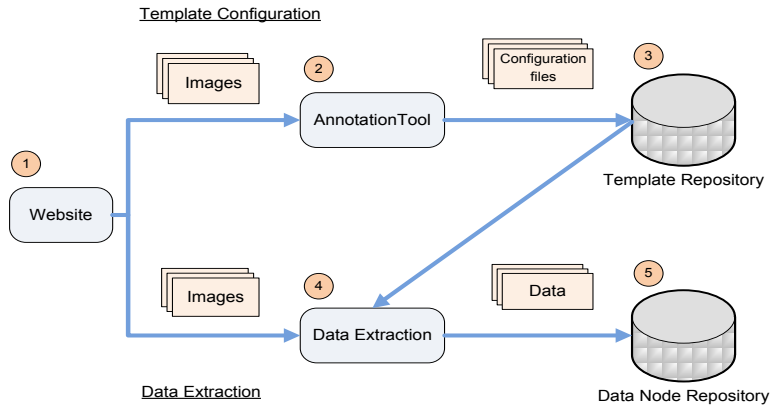

Fig. 1. Air quality data extraction framework.

4 Annotation Tool

The Annotation Tool (AnT) is used to interactively annotate heat maps and it was developed in order to make the annotation process easier for the user.

To make the tool platform independent, the QT Framework[1] was used. The implementation is designed via the MVC (Model/View/Controller) pattern, to ensure its expandability. To allow for different interaction possibilities, two views were implemented. First a simple Tree View, which represents the XML structure and its entries as traversable tree, and a window, which represents the selected tree data graphically. Regions of Interest (ROIs) and Points of Interest (POIs) are drawn onto this window.

Figure 2 depicts the AnT tool after a heatmap from GEM's Project[2] site is loaded. The air quality heatmaps contained in the site are typical examples of images used for representing chemical weather forecasts. The left part of the tool contains the heatmap as well as the ROIs, which are the following: a) the map itself, b) the x and y axis

[1] http://qt.nokia.com/products/

[2] http://gems.ecmwf.int/d/products/raq/

related to the heatmap, c) the color scale, d) the numbers corresponding to the color scale and e) the title of the heatmap. The ROIs are depicted as red bounding boxes and are defined by the user. Finally, their values are recorded to the right part of the AnT inside the XML template under the corresponding nodes.

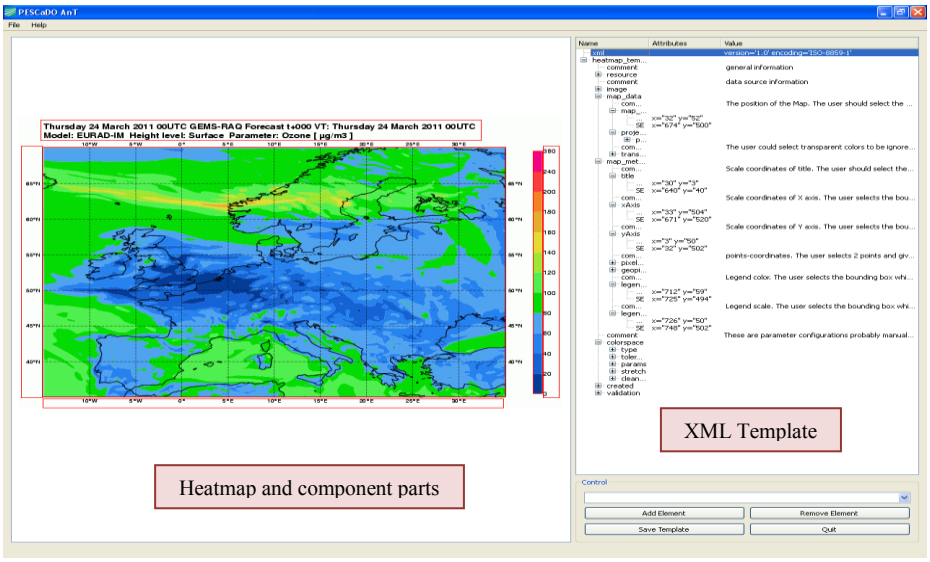

Fig. 2. Screenshot of Image Annotation tool

5 OCR techniques

The OCR module uses the information of the configuration file to extract textual data from images and improve the results using text processing based on heuristic rules.

The first processing step of this module includes the application of OCR on specific parts of the input image as the title, the color scale, the map x and y axes parts. The OCR software that was used is Abbyy Fine Reader[3].

In the second step, we apply text processing based on heuristic rules in order to correct, extract and understand the semantic information encoded in the aforementioned locations. Each of these locations was treated in a different way.

The title, if it exists, usually contains the name of the aspect, the measurement units and the date/time. The measurement units are usually standard depending on the measured aspect so we do not need to extract them. The date/time is considered as the element that is the most difficult to extract, given the fact that many different formats exist. In order to correct possible mistakes in the textual format of the month, day or aspect we exploited the Levenshtein distance. More specifically, three English ground truth sets were created for the three aforementioned elements and were compared to

[3] http://www.abbyy.com/

the corresponding OCR result. Then, we have selected the element from the ground truth that had the smaller Levenshtein distance from the text generated by the OCR.

The color scale contains the values that each color of the map corresponds to. The processing and extraction of information from the color scale element can be divided into two main parts. In the first, we attempt to check and correct OCR results for the scale, while in the second we correlate values to colors. In order to correct the OCR results, we find the most common range among the scale values and adapt accordingly the mistaken values. The correlation of values to colors is achieved by pointing at the middle of each color by using the coordinates of the values.

The last two elements that are analyzed with OCR are the x and y axes. In the case of heatmaps the two axes contain similar information and thus we will apply similar processing techniques to them. The information that can be obtained by each axis is the geographical coordinates of the points of the map. In order to realize this, we have to identify successfully at least two points (x, y coordinates) of the map axes and define their position in relation to the map. In order to identify these points, we first correct most of the errors produced by OCR and then use the coordinates of the elements, as defined by OCR to specify the position of those points.

6 AirMerge

AirMerge is a web-based system that supports harvesting Chemical Weather forecast images and converting them to numerical data. A derivative of its image processing engine is used in the Data Extraction phase of the proposed toolchain, and it is already used for creating harmonized, numerical Chemical Weather data[4].

The AirMerge system combines elements of screen scraping and innovative image processing algorithms [4], [6], [8] in order to produce uniform, indexed data. These data are then stored in a back-end database and may be recalled for further processing such as numerical applications, model ensembles, visualization, transformation etc.

The main task of AirMerge is the extraction of data from heatmaps. This is achieved by using a processing chain that consists of a "screen scraping" phase, where raw RGB pixel data are extracted from heatmaps, a mapping phase, where RGB values are classified to a color scale and mapped to ranges of numerical values and a linear deprojection phase, where the images' raster is interpreted as a geographical grid in a specified geographical projection, centered on key points.

Screen scraping procedure: This step handles the cropping of the original image to a region of interest and parsing of it into a 2D data array directly mapped to the original images' pixels. Also, it associates the color to minimum/maximum value ranges of the air pollutant concentration levels, which is often implied by the color scale associated with the original images. It should be noted that the information about where to crop, where each color on the legend is, to which index it should correspond, etc. are provided by the configuration template of the AnT in the proposed system. In this phase, the mapping of the images' raster to a specific geographical grid is per-

[4] http://projects.isag.meng.auth.gr/airmerge/

formed, since the images themselves represent geographical region. The configuration system allows choosing between the most commonly encountered geographical projections (equirectangular, conical, polar stereographic etc.) and choosing keypoint in the image to allow for precise pixel-coordinate mapping.

"Reconstruction of missing values and data gaps" procedure: This step is introduced to deal with unwanted elements such as legends, text, geomarkings and watermarks, as well as regions that are not part of the forecast area, which might be present after the screen scraping phase. The image pixels are classified into three main categories: valid data (with colors that satisfy the color scale's classification), invalid data (with colors not present in the color scale), and regions containing colors that are explicitly marked for exclusion, and which are considered void during processing. Such marked regions are not considered as part of the forecast, and thus do not undergo data correction. However, regions containing unmarked invalid data are considered as regions with correctable errors or "data gaps" which can be filled-in. This distinction is due to their different appearance patterns: void regions are usually extended and continuous (e.g. sea regions not covered by the forecast, but present on the map), while invalid data regions are usually smaller but more noticeable (e.g. lines, text, watermarks etc.) and with more noise-like patterns, and thus it is more compelling to remove them by using gap-filling techniques. These techniques include traditional grid and pattern-based interpolation techniques using neural networks.

It should be noted that the AirMerge system functionality is also provided via an API, which is available as a REST service [9]. Therefore, AirMerge can serve any request related to the heatmaps of many chemical weather models (e.g. every-day processing), thus making it suitable for environmental service-oriented applications.

7 Results

In this section, we present the results of the framework when applied in three air quality heatmaps from different providers. Since an evaluation of AirMerge is already provided in [8], we evaluate the results of the OCR and the total system output. Regarding the OCR, we focus on the recognition of the x and y axes, since this is the most important information in order to correctly map the air quality index onto the right coordinates. The following providers are considered for the evaluation: GEMS[5], Laboratory of Atmospheric Physics of the Aristotle University of Thessaloniki[6] and the Atmospheric and Oceanic Physics Group[7].

7.1 GEMs website

Figures 3 and 4 depict the original and reconstructed image by the Airmerge system, which are almost identical, and any noise (e.g. black lines) was removed [4], [8].

[5] http://gems.ecmwf.int/d/products/raq/
[6] http://lap.physics.auth.gr/forecasting/fore_images/.
[7] http://www.fisica.unige.it/atmosfera/bolchem/MAPS/

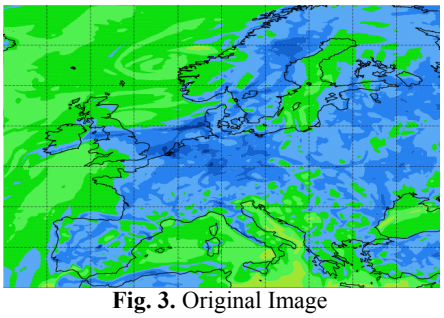

Fig. 3. Original Image

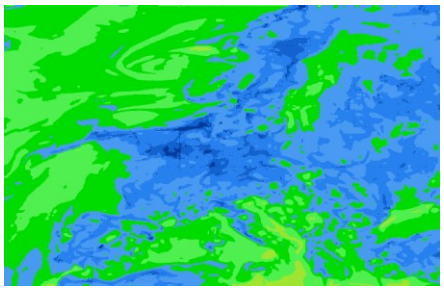

Fig. 4. Image Reconstructed from AirMerge

Table 1. Results for the GEMs website

	Resolution Steps	Correct Value	Estimated Value	Absolute Difference	Error
Longitude step	0.0791	5	5.0634	0.0634	1.25%
Latitude step	0.0777	5	4.9776	0.0224	0.45%

During this process an error is usually introduced mostly due to the inability of OCR to perfectly identify the position of each coordinate on the map axes. In table 1 we report the longitude and latitude steps (i.e. the coordinate step for each pixel), the correct step value between two subsequent coordinate marks (e.g. when the marks are 0^0, 5^0, 10^0, etc. the step value is 5), the estimated value, the absolute difference and finally the introduced error. In both cases the error is very low and acceptable (in general we assume that an error is acceptable, when it is less than 3%).

7.2 Laboratory of Atmospheric Physics of the AUTH site

In a similar way we present the initial and the reconstructed image of this website in Figures 5 and 6. The results are reported in table 2 and the error is again very small.

Table 2. Results for the Atmospheric and Oceanic Physics Group website

	Resolution Steps	Correct Value	Estimated Value	Absolute Difference	Error
Longitude step	0.0311	2	1.9924	0.0076	0.4%
Latitude step	0.0276	1	1.0236	0.0236	2.3%

7.3 Atmospheric and Oceanic Physics Group site

Finally, in table 3 we present results for the last provider reporting an average error of 0.35%. The initial and the reconstructed map are illustrated in figures 7 and 8. It should be noted that the white region in figure 7 is treated as “void space” in figure 8, and considered as a distinct case than national border lines, which are instead treated as unwanted noise and filled-in.

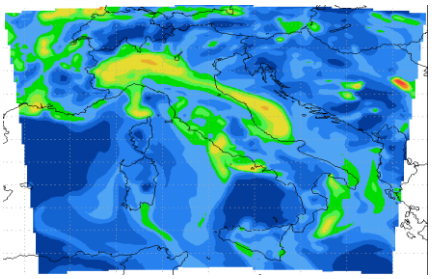

Fig. 7. Original Image

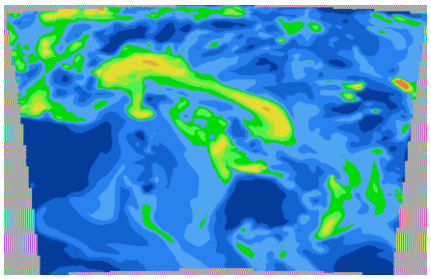

Fig. 8. Image Reconstructed from AirMerge

Table 3. Results for the Atmospheric and Oceanic Physics Group website

	Resolution Steps	Correct Value	Estimated Value	Absolute Difference	Error
Longitude step	0.0289	2	1.9937	0.0063	0.3%
Latitude step	0.0249	1	0.9958	0.0042	0.4%

8 Conclusions

Despite the fact that the current presentation format of air quality forecasts might be ideal for casual users, it is not easily accessible by automatic services which would expect a structured and numerical format of the forecast data. In this context, we have proposed a framework for air quality information extraction from heatmaps, combining existing (AirMerge), as well as new (AnT and OCR) components. This framework could serve as a basis for supporting environmental systems that provide either air quality information from several providers for comparison or orchestration purposes or high level suggestions on everyday issues (e.g. travel planning) based on advanced decision support [13], which could facilitate the quality of life. The proposed work overcomes the limitation of not having access to the raw data, since it only considers information being publicly available on the Internet, thus respecting also data access policies. Although the system has been tested with forecast air quality heatmaps it could also deal with observed pollutant and pollen concentrations represented in the same way. Future work includes extensive evaluation with more images in different projections (e.g. conical) and addressing of pollen heatmaps.

Acknowledgments. This work was supported by PESCaDO project (FP7-248594).

References

1. Balk, T., Kukkonen, J., Karatzas, K., Bassoukos, A. and Epitropou, V.: A European open access chemical weather forecasting portal, Atmospheric Environment 45, pp. 6917-6922, doi:10.1016/j.atmosenv.2010.09.058 (2011).
2. Karatzas, K.: Internet-based management of Environmental simulation tasks. In Farago I., Georgiev K. and Havasi A. (eds) Advances in Air Pollution Modelling for Environmental

Security, NATO Reference EST.ARW980503, 406 p., Hardcover, Springer, ISBN: 1-4020-3349-4, pp. 253-262 (2005).
3. San José, R., Baklanov, A. Sokhi, R.S., Karatzas, K. and Pérez, J.L.: Computational Air Quality Modelling. In Developments in Integrated Environmental Assessment, Vol. 3, Environmental Modelling, Software and Decision Support Edited by: A.J. Jakeman, A.A. Voinov, A.E. Rizzoli and S.H. Chen ISBN: 9780080568867 (2008).
4. Epitropou, V., Karatzas, K. and Bassoukos, A.: A method for the inverse reconstruction of environmental data applicable at the Chemical Weather portal. In Geospatial Crossroads @GI_Forum'10, Proceedings of the GeoInformatics Forum Salzburg, pp. 58-68, Wichmann Verlag, Berlin, ISBN 978-3-87907-496-9 (2010).
5. Karatzas, K., Kukkonen, J., Bassoukos A., Epitropou V. and Balk T.: A European Chemical Weather forecasting Portal. 31st ITM - NATO/SPS International Technical Meeting on Air Pollution Modelling and its Application, Torino, 28 Sept. 2010. Published in Air Pollution Modeling and its Applications XXI, Springer, NATO Science for Peace and Security Series C: Environmental Security, Steyn, Douw G.; Trini Castelli, Silvia (Eds.), 1st Edition, Hardcover, ISBN 978-94-007-1358-1, pp.239-243 (2011).
6. Epitropou, V., Karatzas, K.D., Bassoukos, A., Kukkonen, J. and Balk, T.: A new environmental image processing method for chemical weather forecasts in Europe. In Information Technologies in Environmental Engineering, Proceedings of the 5th International Symposium on Information Technologies in Environmental Engineering, Poznan, 6-8 July 2011 (2011).
7. Kukkonen, J., Klein, T., Karatzas, K., Torseth, K., Fahre Vik, A., San José, R., Balk, T. & Sofiev, M.: COST ES0602: Towards a European network on chemical weather forecasting and information systems, Advances in Science and Research Journal, 1, pp. 1–7 (2009).
8. Epitropou, V., Karatzas, K., Kukkonen, J. and Vira, J.: Evaluation of the accuracy of an inverse image-based reconstruction method for chemical weather data, International Journal of Artificial Intelligence, in press (2012).
9. Epitropou, V., Johansson, L., Karatzas, K., Bassoukos, A., Karppinen, A., Kukkonen, J. and Haakana, M.: Fusion Of Environmental Information For The Delivery Of Orchestrated Services For The Atmospheric Environment In The PESCaDO Project, International Environmental Modelling and Software Society (iEMSs), 2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet, Leipzig, Germany, (R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp, eds.), in press (2012).
10. Musavi, M.T., Shirvaikar, M.V., Ramanathan, E. and Nekovei, A.R.: Map processing methods: an automated alternative. In Proceedings of the Twentieth Southeastern Symposium on, IEEE Computer Society, System Theory, pp. 300 – 303 (1988).
11. [6]Henderson, T.C. and Linton, T.: Raster Map Image Analysis. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition (ICDAR '09). IEEE Computer Society, Washington, DC, USA, 376-380 (2009).
12. Cao, R. and Tan, C.: Text/graphics separation in maps. In: D. Blostein, Y.-B. Kwon (Eds.), Fourth IAPR Workshop on Graphics Recognition, Lecture Notes in Computer Science, vol. 2390, Springer, Berlin, pp. 167–177 (2002).
13. Wanner, L., Vrochidis, S., Tonelli, S., Moßgraber, J., Bosch, H., Karppinen, A., Myllynen, M., Rospocher, M., Bouayad-Agha, N., Brugel, U., Casamayor, G., Ertl T., Kompatsiaris, I., Koskentalo, T., Mille, S., Moumtzidou, A., Pianta, E., Saggion, H., Serafini, L. and Tarvainen, V.: Building an Environmental Information System for Personalized Content Delivery. International Symposium on Environmental Software Systems (ISESS 2011), Brno, Czech Republic (2011).