Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Towards an era of epidemiological databases for autoimmune diseases

Vassiliki Gkantouna, Marina Ioannou, Athanassios Tsakalidis, Emmanouil Viennas
Department of Computer Engineering and Informatics
Faculty of Engineering, University of Patras
Patras, Greece

Konstantinos Poulas
Department of Pharmacy
School of Health Sciences, University of Patras
Patras, Greece

John Tsaknakis, Giannis Tzimas
Department of Applied Informatics in Management & Economy
Faculty of Management and Economics, Technological Educational Institute of Messolonghi
Messolonghi, Greece

Abstract—**Nowadays, autoimmune diseases are among the leading causes of death for a remarkable number of patients all around the world. Recent studies have witnessed that the epidemiological indices for a specific disease can vary according to ethnic and geographical parameters. As a result, the genetic epidemiology of autoimmune diseases is a major matter of study for the worldwide scientific community. We have previously reported the development of dAUTObase (www.dAUTObase.org), a database recording solely epidemiological data of autoimmune diseases in various populations around the globe. Here, we present an important upgrade of the dAUTObase system focused on the development of new data visualization tools oriented to further assist the effective data querying and the mining process.**

Keywords-autoimmune diseases; epidemiological databases; visualization tools; querying tools

I. INTRODUCTION

Autoimmune diseases (ADs) are characterized by loss of self-tolerance and an altered immune response which lead to tissue and organ damage caused mainly by autoantibodies that attack the body own healthy tissues. A large group of heterogeneous conditions are classified as ADs, ranging from Sjögren's syndrome to systemic lupus erythematosus (SLE). Depending on the number of organs affected, ADs are divided in two main categories: organ-specific ADs, which target a single organ, or tissue and systemic ADs that correspond to a multisystem involvement. Overall, many millions of individuals worldwide are affected by ADs. According to an epidemiological analysis of 24 preselected ADs, carried out in 1997, it was estimated that 8.5 million individuals in the USA were afflicted with an AD, a number that corresponds to a prevalence rate of approximately one in 31Americans [1]. As a consequence, ADs represent an enormous burden on both the patient and society. The quality of patients' life is impaired, while some diseases can reduce lifespan. Also, in many cases, the early age of onset results in long disease duration, increased health-care costs, or even loss of patients' productive years.

In spite of the indisputable socioeconomic impact, the mechanisms and etiologies of autoimmunity remain elusive. Twin studies reviewed in [2], revealed that high concordance rates among monozygotic twins, observed in several ADs, are suggestive of a strong genetic component. In contrast to mendelian diseases, ADs are considered to have a multifactorial underlying basis. In fact, it is assumed that the pathogenesis of ADs is caused by a complex interaction between multiple genotypes of low penetrance and environmental factors, including pathogen exposure, particularly Epstein-Barr virus [3] [4] [5], sex hormones [6] and lifestyle habits, such as cigarette smoking [7]. Recently, the field of epigenetics has emerged in the study of autoimmunity due to the potential role of epigenetic regulation as a gene-environment link.

Spatial epidemiology or geoepidemiology is the description and analysis of geographically indexed health data with respect to demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors. Recent advances in data availability and data visualization technologies have offered new opportunities for researchers to improve on the traditional reporting of disease at national or regional scale. By collecting and comparing such epidemiological data (e.g. prevalence, incidence etc.) one can obtain important clues to the genetic and triggering environmental mechanisms of autoimmunity. We here present the first attempt, to the best of our knowledge, for the development of a database targeted on the geoepidemiology of ADs, an invaluable tool for the study and comprehension of autoimmunity.

To date, there are several data repositories that fall under the banner of ADs databases. However, these databases, which are mainly literature-derived, focus on the documentation of ADs and provide no possibility to study them by an

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

epidemiological point of view. To this end, we are currently experimenting on various visualization techniques in an effort to set the scene for a new interesting and promising area in the field of ADs, called "geoepidemiological databases". The aim of such databases will be the study of the descriptive epidemiology of ADs and to characterize the risk of their coexistence within individuals around the world.

Although the initial launch of dAUTObase included some basic data visualization functionality [8], the recently upgraded version does allow further data querying to be coupled to advanced data visualization. In this work, we present the development of new interactive web-based data visualization and querying tools for dAUTObase. These tools allow users to compare the epidemiological indices of the documented in dAUTObase ADs among different populations and thus identify hidden relationships between individual pieces of information. More specifically, we have built an elegant web-based multimedia front-end based on a software tool launched by Microsoft, namely the PivotViewer [9], in order to provide a high level visualization of the data collection and the mining process. Additionally to the PivotViewer, there are also two alternative visualization querying interfaces based on the Flare visualization toolkit [10] and the JQVmap [11], providing two extra visualization types of the underlying data collection, namely the Disease Treemap and the Disease World Map allowing users to query disease distributions and correlations among populations.

The remainder of this paper is organized as follows: Section 2 presents the system architecture, while Section 3 describes the technologies used in order to develop the new visualization tools. Section 4 illustrates the results produced in this work. Finally, Sections 5 and 6 discuss future steps and provide concluding remarks.

II. System Architecture

dAUTObase (http://www.dautobase.org) is a web application recording solely epidemiological data of ADs in various populations around the globe. It is public and there are no registration requirements for data querying.

dAUTObase is based on a relational database developed with Microsoft SQL Server [12]. Database records include the population, the geographic region, the relative publication, and a number of epidemiological indices like incidence, prevalence, mortality, morbidity, age at onset etc. The overall system architecture is based on a three-tier model which is a robust and flexible enough to aggregate multiple information sources and integrate modular development.

III. Technologies Used

dAUTObase provides a web-based multimedia visualization environment for population-based ADs epidemiological data collection and retrieval. Driven by the need to provide an effective and user-friendly representation of the underlying data collection, we have utilized state-of-the-art visualization technologies and implemented three alternative data querying and visualization interfaces. This way, the system can provide various views of the underlying data collection and enable users to discover hidden patterns between apparently unrelated items. In what follows, a description of the technologies used is presented.

A. PivotViewer

The dAUTObase main querying interface is based on Microsoft Silverlight PivotViewer control [13], a free web browser plug-in that enables interactive media experiences. Microsoft Silverlight is an application framework for developing rich Internet applications, with features and purposes similar to those of Adobe Flash. By leveraging Deep Zoom [14], PivotViewer displays full, high-resolution content without long loading times, while the animations and natural transitions provide context and prevent users from feeling overwhelmed by large quantities of information. It supports dynamic data visualization, sorting, organization and categorization.

The dAUTObase PivotViewer application enables users to interact with thousands of objects at once, and categorize and browse data in multiple ways so that they can reveal new trends in them.

B. Flare Visualization Toolkit

dAUTObase queries can also be performed utilizing the Flare Visualization Toolkit. Flare is an open-source library written in ActionScript 3 [15], an object-oriented programming language, for creating data visualizations that run in the Adobe Flash Player. Including a wide variety of features, ranging from basic charts to complex interactive graphs, it supports data management, visual encoding, animation, and interaction techniques. Flare has already been used in many well-known web-based visualization applications. Among the most representative examples is the BBC SuperPower website [16] that uses Flare for mapping the top 100 sites on the Internet. Also, the IBM Visual Communication Lab uses Flare to build visualizations for Many-Eyes [17]. For example, Many-Eyes has been used for visualization of ADs in the US [18].

We have utilized the Flare to implement an alternative representation of the underlying data collection, namely the Disease Treemap. A treemap is an easy way of analyzing large amounts of data in a small space. Introduced by Ben Shneiderman in 1991, treemaps are a space-filling approach of showing hierarchies in which the rectangular screen space is divided into regions, and then each region is divided again for each level in the hierarchy.

C. JQVMap

Finally, dAUTObase queries output can also be visualized using an alternative interface based on the JQVMap, a jQuery plugin that renders Vector Maps by using resizable Scalable Vector Graphics (SVG) for modern browsers and VML for the rest browsers. It provides a variety of parameters to the developers in order to change the look and feel of the maps by adding colors, events, and interactivity. We have utilized the JQVMap technology to implement the Disease World Map.

IV. Results

Since dAUTObase is the first attempt towards the establishment of geoepidemiological databases for ADs, our primary goal was to provide scientists a powerful tool for the study of the autoimmunity epidemiology.

We have critically collected and screened a large number of epidemiological studies on ADs, analyzing their frequencies in various populations around the globe. We have given special emphasis on the identification of incidence and prevalence rates, while we also registered additional information, if presented. The next step was the effective visual representation of the collected information in a way that would give researchers the opportunity to explore, analyze and understand the screened data. To this end, we have developed a data visualization environment for the dAUTObase data collection, designed to provide users a much more intuitive way to epitomize the large amounts of information without losing their orientation.

On this basis, we have built a visualization tool providing three basic querying interfaces that rely on the following visualization approaches: (a) the main interface based on the PivotViewer, (b) an interface based on the Flare Visualization Toolkit and (c) an alternative interface based on the JQVMap. They all support sophisticated data exploration and allow users to apply advanced filtering criteria upon a set of multiple views of the same underlying data collection, giving them an active role in the mining process. Moreover, they offer users the possibility to zoom-in from the extensive dAUTObase datasets to particular disease-specific and/or population-specific data. This way, users can detect the links between distant data items and handle them in a way that intuitively reflects their semantic proximity. It is noteworthy here that while user experiments with different querying scenarios, he/she may sometimes result in making incidental discoveries of potentially high medical importance. Some such indicative examples are the discovery of demographic patterns and the extraction of population-based patient characteristics.

A. *PivotViewer*

The form of a card (Fig. 1A) is used for the representation of each AD, in order to provide a more human-centric visualization approach. This card is accompanied by a sidebar information panel (Fig. 1B) providing in-depth information concerning the particular AD and population, appearing when users zoom-in on the card.

The entire dAUTObase data collection, as produced by PivotViewer (http://www.biodata.gr/dautobase/pivot/), is shown in Fig. 2. A data filtering panel (displayed on the left side; Fig. 2) is available providing a variety of filtering criteria to be applied on the underlying data collection. This way, users can sort, organize and categorize data dynamically according to common characteristics that can be selected from the data query menu and then zoom-in for a closer look, by either filtering the collection to get a subset of information or clicking on a particular card.

By exploiting the above functionalities, users have the opportunity to experiment with different disease and/or population specific scenarios which may guide them to identify and extract the valuable knowledge that lies in them. Such an example would be the query of the occurrence of the Myasthenia gravis in different populations around the world, sorted by the population name (Fig. 3).

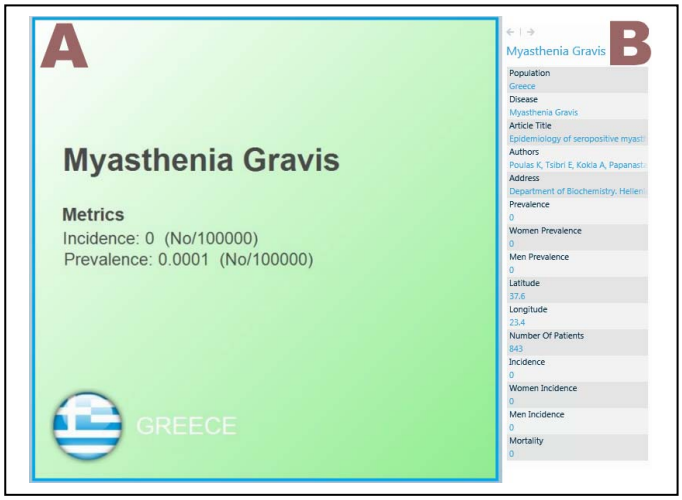

Figure 1. dAUTObase card item: (A) A card represents a single AD in the underlying collection occurring in a specific population. The card consists of the AD name, information about the incidence and prevalence rate and a flag corresponding to the particular population. (B) A sidebar information panel is appearing when users zoom-in on a card providing in-depth information about the card currently selected.

Figure 2. dAUTObase data collection: The entire dAUTObase AD data collection, as produced by PivotViewer. A filtering panel is available to the users containing several categories that can be used to include or exclude specific cards and to help users apply a particular querying scenario.

B. *Disease Treemap*

The dAUTObase data collection can also be presented in an alternative representation namely the Disease Treemap (http://www.biodata.gr/dautobase/DiseaseTreeMap/) (Fig. 4). It is a treemap presenting the incidence rates of ADs over populations, with each rectangle representing the specific incidence rate of an AD for a specific population. The color of each rectangle corresponds to a different population. In case that there are more than one rectangles included within a country's rectangle, the sub-rectangles refer to different ADs occurring within the same population or within regions of it. The area of each rectangle encodes the incidence rate of the AD in the specified population. By clicking on a rectangle, dAUTObase Disease Treemap displays all the relative populations having less than 10% average difference of

incidence. Additional information is also provided such as the population, the disease, the region name, and the corresponding value of the incidence of the specified AD, when users hover the mouse over a rectangle.

A typical query example is shown in Fig. 4 presenting the incidence rate of Type 1 Diabetes in different populations and regions around the world.

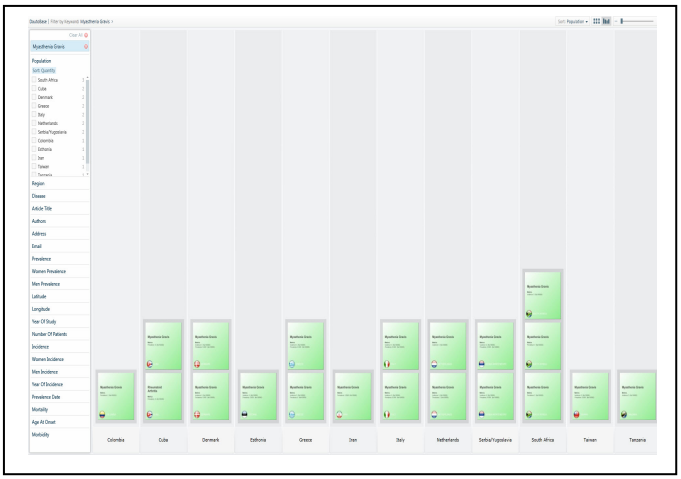

Figure 3. PivotViewer query example: Query output screen for the occurrence of the Myasthenia gravis in different populations around the world sorted by the population name. The query returns 12 populations, which can be further separated according to other filtering criteria such as the incidence rate.

C. *Disease World Map*

The Disease World Map (http://www.biodata.gr/dautobase/DiseaseWorldMap/) (Fig. 5) is a visual depiction of the worldwide map enriched with a variety of epidemiological indices per population. Through two drop-down lists, the user has the ability to make a selection of the specific AD and the type of epidemiological indices to be displayed. When these selections are made, the Disease World Map dynamically updates the data appearing on the map. The color scale, corresponding to each country, depends on the screened data and reflects the value of the selected index for the specified disease and population. By hovering or clicking on a specific country, an additional information panel appears on the left, presenting details like the name, the flag and the value of the selected index for the specified population (the selected country is always denoted by the green color).

A typical query example is shown in Fig. 5 presenting the incidence rate (among men and women) for Type 1 Diabetes around the world.

V. FUTURE WORK

In the near future, we are planning to experiment on more querying and alternative visualization types and techniques able to manipulate multidimensional data, in order to create a global visualization framework allowing users to conduct multifaceted comparative studies of the querying results.

The administration environment of dAUTObase is another aspect which we consider to be of great importance. The system will support multiple user profiles, providing scaled access to the information with users divided into three main groups according to their privileges, namely the administrators, the curators and the simple users. Data entry and modification in dAUTObase will be possible only for registered users. This way, researchers worldwide could concurrently contribute the epidemiological data of their country of origin and thus dAUTObase can become the first worldwide up-to-date and comprehensive geoepidemiological data repository.

We also intend to incorporate data mining techniques to further enhance and automate the discovery of valuable information and we are going to expose dAUTObase dataset through web services based on the oData protocol [19] giving the ability to the research community to freely exploit our data.

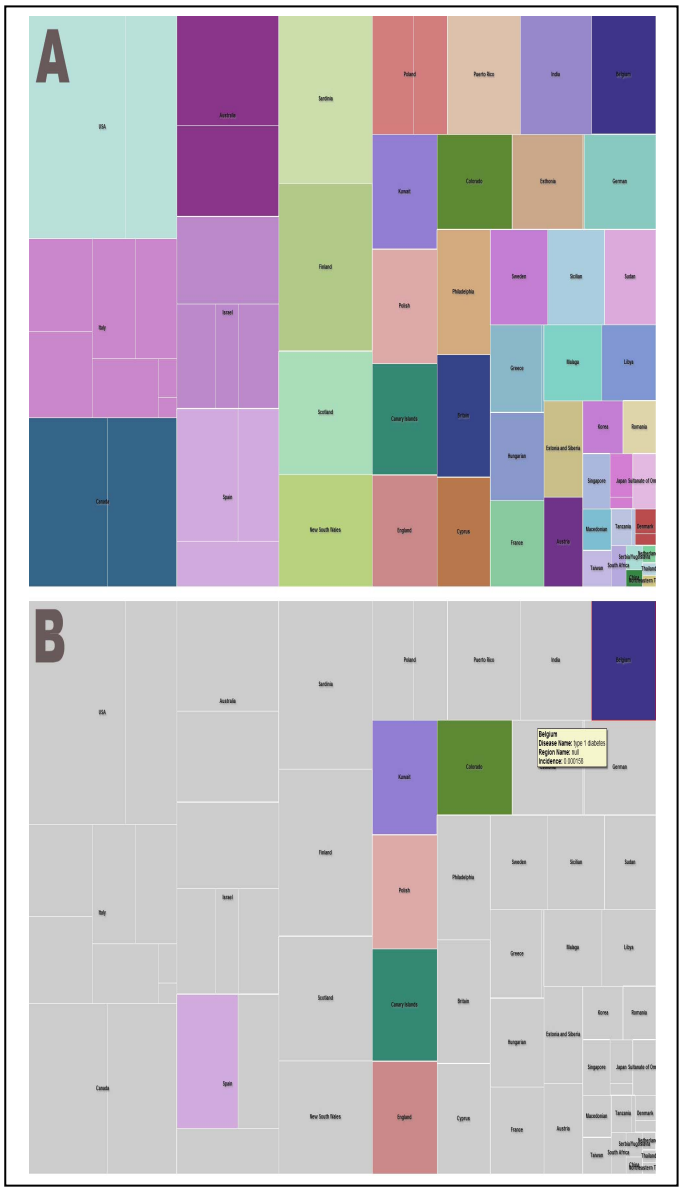

Figure 4. Disease Treemap: (A) the entire dAUTObase AD data collection, as produced by Flare visualization toolkit. (B) Display of the incidence rate of Type 1 Diabetes in different populations and regions. Only the areas with less than 10% difference of incidence are colored.

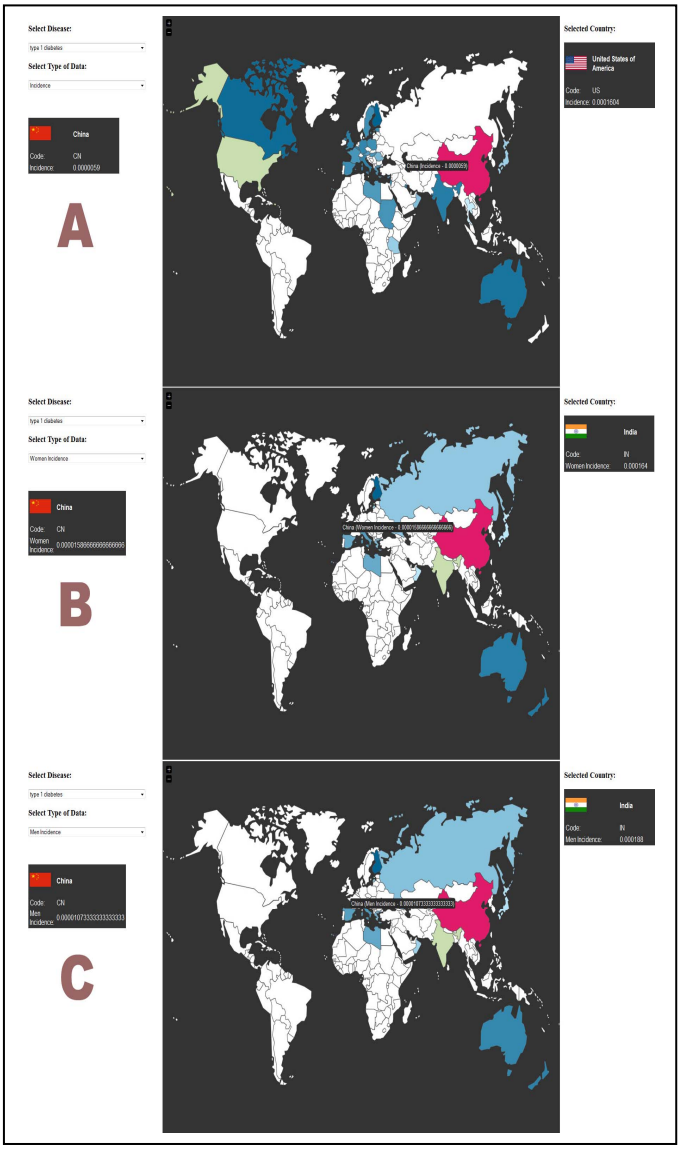

Figure 5. Disease World Map: Display of (A) the overall, (B) the women and (C) the men incidence rate of Type 1 Diabetes in different populations around the world. By hovering or clicking on a specific country, an additional information panel appears on the left.

VI. CONCLUSION

The thorough and systematic study of the epidemiology of ADs can open the way to new perspectives in terms of medical care utilization and patient quality of life. In this work, we present the integration of novel visualization tools within the dAUTObase 2.0 version in an attempt to provide a state-of-the-art data and querying visualization environment for the epidemiological study of ADs. Our primary goal is to offer a robust and stable tool oriented to support and broaden multidisciplinary knowledge among the scientists worldwide.

REFERENCES

[1] D.L. Jacobson, S.J. Ganqe, N.R. Rose, and N.M. Graham, "Epidemiology and estimated population burden of selected autoimmune diseases in the United States," Clin. Immunol. Immunopathol., vol. 84, pp. 223–243, 1997.

[2] D.P. Bogdanos et al,"Twin studies in autoimmune disease: Genetics, gender and environment," Journal of Autoimmunity, vol. 38, pp. J156–J169, 2012.

[3] A. Ascherio, K.L. Munger, E.T. Lennette, D. Spiegelman, M.A. Hernan, and M.J. Olek, "Epstein-Barr virus antibodies and risk of multiple sclerosis: a prospective study," JAMA., vol. 286, pp. 3083–8, 2001.

[4] N. Balandraud, J.B. Meynard, I. Auger, H. Sovran, B. Mugnier, and D. Reviron, "Epstein-Barr virus load in the peripheral blood of patients with rheumatoid arthritis: accurate quantification using real-time polymerase chain reaction," Arthritis Rheum., vol. 48, pp. 1223–8, 2003.

[5] J.A. James, B.R. Neas, K.L. Moser, T. Hall, G.R. Bruner, and A.L. Sestak, "Systemic lupus erythematosus in adults is associated with previous Eipstein-Barr virus exposure," Arhritis Rheum., vol. 44, pp. 1122–6, 2001.

[6] K.H. Costenbader, D. Feskanich, M.J. Stampfer, and E.W. Karlson, "Reproductive and menopausal factors and risk of systemic lupus erythematosus in women," Arthritis Rheum., vol. 56, pp. 1251–62, 2007.

[7] E.W. Karlson, S.C. Chang, J. Cui, L.B. Chibnik, P.A. Fraser, and I. Devivo, "Gene-environment interaction between HLA-DRB1 shared epitope and heavy cigarette smoking in predicting incident RA," Ann. Rheum. Dis., vol. 69, pp. 54–60, 2010.

[8] D.I Antoniou et al, "dAUTObase: Mining Gems on Autoimmune Diseases Utilizing Web Visualization Technologies", in the Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Nov. 2010, Corfu, Greece, pp. 1–5.

[9] Microsoft Live Labs – Pivot, Available: http://www.microsoft.com/silverlight/pivotviewer/

[10] Flare – Data Visualization For The Web, Available: http://flare.prefuse.org/

[11] JQVmap , Available: http://jqvmap.com/

[12] Microsoft SQL Server, Available: http://www.microsoft.com/sqlserver/en/us/default.aspx

[13] Microsoft Live Labs – Pivot - Silverlight Control, Available: http://www.silverlight.net/learn/data-networking/pivot-viewer/pivotviewer-control

[14] Deep Zoom, Available: http://www.microsoft.com/silverlight/deep-zoom/

[15] ActionScript 3, Available: http://www.adobe.com/devnet/actionscript.html

[16] BBC SuperPower: Visualizing the Internet, Available: http://news.bbc.co.uk/2/hi/technology/8562801.stm

[17] Many-Eyes, Available: http://www-958.ibm.com/software/data/cognos/manyeyes/

[18] Many-Eyes, Available: http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/autoimmne

[19] Open Data Protocol, Available: http://www.odata.org/