Global Energy Observatory (GEO): A One-stop Site for Information on Energy Systems, Infrastructure and Emissions

Rajan Gupta
Theoretical Division, Los Alamos National Lab, New Mexico, 87544
rajan@lanl.gov

Harihar Shankar and Aswin T. Y. Venkata
Electrical and Computer Engineering, University of New Mexico, Albuquerque
New Mexico, 87544
harihar@unm.edu and aswintyv@unm.edu

Abstract

This paper describes our attempt to create a one-stop open web tool for collecting, collating, managing and analyzing global energy systems at multiple scales. Participation by the public and experts through volunteered information is a key component of all aspects of this project and for the organic growth and scrutiny of the scientific databases. The geospatially and time referenced data in GEO will be viewable using open GIS tools such Google Earth for mashups and integrated analyses. The framework and databases are intended to serve multiple communities. (i) Policy makers interested in the growth and/or evolution of the systems, resource optimization, emissions and environmental management; (ii) Negotiators of international treaties; (iii) monitors and advocates of environmental treaties; (iv) energy companies and carbon traders; (v) researchers and academics; (vi) educationists interested in creating "real-life" demonstrations of the network of energy systems and projects for high school and college students; (vii) journalists for informing the public and (viii) the public and various government and non-government organizations interested in a quick and ready reference. Our immediate goals are to finish the first version of the framework for the four interlinked databases described in this paper by July 2009 and to scrape as much of publicly available data as possible by end of 2009 to motivate the use of GEO. An illustration of the opportunity created by such a one-stop shopping site is the integration of the large amounts of data on power plants available in the open. Once the existing highly fragmented data collected with different levels of sophistication and validation, and located in hundreds of different places and in many different formats is available in GEO, much more comprehensive real-time analyses can be done. We believe we can capture enough of the power plant data in GEO to provide basic information and geo-location of 75 percent of the global power generation capacity by the end of 2009.

Introduction

The energy-environment-development-climate (EEDC) challenge is perhaps the most urgent and multi-dimensional problem facing humanity. Until technological innovations can provide environmentally sustainable and reasonably priced energy to the global population, it is imperative that we develop tools to help us quantify and understand the techno-socio-economic-political aspects of the challenge, and motivate investment in energy efficiency and lifestyle changes that would reduce emissions of greenhouse gases and environmental impacts of the use of fossil fuels.

Energy and environmental data collection needed to perform a comprehensive and defensible analysis continues to grow in scale, diversity, and complexity. Unfortunately, the collected data is very incomplete and not easily available to consumers, scientists, and decision makers. The abundant yet fragmented, incomplete and heterogeneous data and data sources have created increasing demands on, and opportunities for, information technologies. Even if data exist, today we are asking new questions that were unanticipated even a few years ago. This is requiring repurposing, transforming, and integrating multiple, uncoordinated, and sometimes variously restricted or proprietary data sources. Having a comprehensive framework and database such as GEO will significantly enhance our ability to analyze the EEDC challenge.

Our goal is to build, with public participation, a web-based real-time Global Energy Observatory (GEO) of geo-spatially referenced global energy systems and their emissions, use it to educate a much larger public and motivate them to change their lifestyles in order to promote reduced use of fossil fuels, and encourage the development of alternate energy technologies. We are developing the framework for collecting, collating, managing, visualizing, and understanding diverse digital content in circumstances ranging from individuals through groups, organizations, and societies, and from individual devices to globally-distributed systems. It is anticipated that the availability of such data to anyone with web access, will ultimately lead to the engagement of a much larger and diverse group providing solutions to the EEDC challenge.

Our aims for GEO are to provide a “one-stop shopping” portal for adding, enhancing, correcting and validating energy and emissions data, develop semantic tools for enriching the knowledge associated with this data, and integrate helpful tools for accessing and analyzing these enriched data sets. The framework has been designed to be an open, geo-spatial, and time-referenced database (maintained under the Creative Commons license) housing details of humanity’s energy systems, and tracking emergent phenomena like distributed energy generation (renewable energy systems for homes, businesses and small organizations) as well as changes in public opinions and lifestyles.

A significant part of our effort is also aimed at designing the framework to facilitate education and behavioral and social science research on how people perceive and respond to the EEDC challenge. We are striving to make GEO accessible to high school and undergraduate students for creating energy and environment related projects, for educating the public, and for providing a ready and comprehensive reference that can be accessed by policymakers. We believe that education on, and awareness of, the EEDC challenge are essential for transformational changes, for example, early adoption of initiatives such as smart green grids. We also hope to advance scientific knowledge regarding how people perceive of the EEDC challenge, and develop and implement tools for promoting and assessing the success of this GEO. The data and analysis tools comprising GEO will also be useful for researchers who rely on detailed geo-spatial and time referenced information. It is anticipated that GEO will serve as a prototype for similar frameworks that could be used to address other techno-economic-social-political global challenges. These include water resources, bio-diversity, habitats, ecosystems and their management, public health and terrorism.

Overview of Current Efforts:

The current version of GEO has been developed using a traditional web-based LAMP (Linux, Apache, MySQL and PHP) infrastructure. The alpha version is available at http://openmodel.newmexicoconsortium.org/. It consists of a robust framework for collecting, adding, editing, storing, visualizing, moderating and querying data. Four linked databases that will house global information on

1. GEOpower (Power plants): Coal, gas, geothermal, hydro, nuclear, oil, solar PV, solar thermal, waste, and wind power plants. We impose no cut-offs on the size of plants included. Different sizes can be resolved easily by appropriately written queries.
2. GEOresources (Fuels and resources): Gas and oil fields, coal and uranium mines, crude oil refineries, solar and wind potential, and biomass (agricultural) resource.
3. GEOtransmission (Transmission of energy): Gas and oil pipelines, coal, LNG and oil ports, rail and road links, shipping lanes and electric transmission grid.
4. Carbon footprint of individual end-users, their understanding of the EEDC challenge and installations of distributed generation (solar, wind, heat pumps) and storage systems for individual homes, commercial buildings, institutions and communities. This survey is called "Reducing Our CO2 Footprint" and the main source of data will be volunteered information by end-users.

A very significant amount of open data on energy infrastructure already exists on the Web. We propose to harvest sufficient amount of these data to populate databases 1-3 in order to engage the public and the experts in the project. The eventual goal is not to build complete databases ourselves but to make the framework easy to use. We, therefore, plan to concentrate on developing tools to maintain very high quantity control of information and to engage a large and diverse global community that is then willing to participate in maintaining it and building it further. Nevertheless, we recognize that the success of any such project depends on the usefulness of these databases, therefore, maintaining very rigorous quality control and providing real-time information is essential. We discuss our strategy for implementing moderation and validation and for automation of this process.

Examples of data already populated are: (i) all the available EIA and EPA data for USA power plants and their emissions. (ii) All nuclear power plants from the IAEA PRIS data base. (iii) Significant amount of data on Australia, Canada, India, Mexico, South Africa and the UK. (iv) We are in the process of harvesting data for Europe (EU25 from the archive EPER), (v) Significant amount of additional data on oil refineries, oil and coal ports and pipelines has been collected and is being sorted for inclusion.

Framework

The information science (LAMP) framework has been designed to facilitate public and expert participation on the following:

1) To view, edit and add information on individual systems through web-based forms. For each unit (for example, a power plant or a refinery or a coal port) there are two linked databases – one structured and the other format-free. In the structured database, the information has to be input (by experts or general public)

in a specified format (integer, float, entries selected from a drop down menu or character strings) so that it can be coupled to analysis.

2) Format-free discussions and data sharing are facilitated by an associated page called *discussion forum* and attached to each infrastructure unit. This is designed to encourage a non-scientifically oriented user to share their ideas and information.
3) The structured database includes an inventory of the plant characteristics, timeline of changes, performance and emissions data, and a list of the associated infrastructure. We have incorporated graphical analysis of performance data and built in visualization of associated infrastructure using Google Maps.
4) Analyze and download information in different formats (currently only KML for visualization using Google Earth has been implemented).
5) Capture the growth in small, distributed generation and storage through a web survey, thus relying on volunteered information. This is part of the "reducing our CO2 Footprint" database and will include installations of renewable energy (solar, wind, geothermal, and fuel cells) and storage (battery) units in homes and commercial buildings. The same survey will also query the user on their energy use and provide estimates of their carbon footprint and ways of reducing it. We propose to build tools that will allow users to harvest this database to study the dynamics of change in both off-grid and grid-based user-level distributed systems.

Populating the last database capturing energy end-use, CO2 footprint, and distributed energy systems installed by individuals requires public participation, as public databases with such information do not exist. For this database to succeed, user interactions with GEO will need to provide adequate motivation for people to volunteer information on their energy systems, lifestyles and actions.

Over the next 3-4 months our focus will be on finishing the development of this framework for the 4 databases. Major future effort will concentrate on adding more enhanced search and cross-correlations capabilities and a more detailed tracking of the time history of individual fields and their provenance. We are investigating semantic web technologies in this context. Once the framework and integration is complete, we plan to make the schema and details of the databases public (and create mirror sites) so that many more people can contribute to the analyses and develop tools.

Emissions and Impact Inventory

We have implemented tracking three kinds of emissions from fossil fuel fired power plants and oil refineries:

- Green House Gases: CO_2, N_2O and CH_4
- Pollutants: SOx, NOx, mercury, volatile organics, and particulates.
- Water use and impact: In particular water drawn, used and contaminated by power plants, refineries, mines and oil and gas fields.

The environmental impacts and the plant's footprint will be displayed at different levels. First, a user can draw the boundary of visible direct impact (disturbed earth) of the energy infrastructure using an API to Google Maps. Second, the user can plot the time history of emissions and performance characteristics plant by plant and correlate them with deployment of emission control devices, technology insertion and policy changes. In the future, we propose to develop/integrate a plume transport and pollutant dispersal model and tool that would allow a user to estimate and plot contours showing deposition levels on the ground. These tools can then be used by experts to carry out a comprehensive cost-benefit analysis of energy systems that includes their impact on health and the environment.

Another feature we are in the process of developing is a geospatial visualization of associated infrastructure. As an example consider the infrastructures associated with a coal-fired power plant. These are the coalmines supplying the coal, coal beneficiation plants, rail link for coal transportation, coal-fired power plant itself, and waste treatment plants. Providing such an integrated view would allow lifecycle analysis and emissions inventory and management.

User Classes and Roles

The moderation process we have implemented is analogous to the peer review system followed by scientific journals and professional societies for publication of scientific results. The users of GEO are divided into five categories:

Casual Users: This class consists of users not willing to register. They will only be able to view and download data and will not be able to edit or contribute new information.

Contributors: These are registered users who can access all the functions in GEO and contribute new information.

Editors: These are subject area experts that are assigned jurisdiction over contributions from a given region, for example, natural gas power plants in Italy. They will be pinged to periodically review all new contributions to a given plant within their jurisdiction, consolidate all the changes they can validate and create a new version for review by the moderator. To facilitate this review process editors would have overlapping jurisdictions so that more than one editor can act on changes to a given plant. Information that editors cannot validate but consider plausible will be placed in the discussion forum attached to the plant for further comments by the user community. Finally, for each contribution editors would assign a score of accuracy/trust to the contributor. These scores will, over time, be used to automate the process, for example, contributions from users with high scores would be accepted automatically in most cases until random checks and user feedback indicates otherwise. In performing these tasks an editor serves the role of a referee for scientific journals.

Moderators: These are professionals with experience with the moderation process. The moderator’s job is to (i) review submissions by the editors and incorporate them into the database, (ii) score the editors for accuracy in performing their checks, (iii) recommend high performing editors for consideration as additional moderators, and (iv) recruit new editors from high value contributors. The moderator is the analogue of an editor of a scientific journal.

System Administrators: They have the overall responsibility for preserving the scientific quality of the database and for its development, and for selecting and maintaining the pool of potential editors and moderators.

Maintaining Data Integrity and the Moderation Process

Our goal is to maintain GEO as an open scientific database that is continually built and maintained with public participation. This requires simultaneously maintaining its integrity at all times and providing users instant gratification and recognition to motivate them to contribute and develop useful tools to query and visualize the data. Our anticipation is that over time an increasing fraction of the data in all four databases will come as volunteered contributions from both the public and hopefully from the energy companies themselves as part of their public outreach and transparency. To preserve the scientific integrity of the databases, it is therefore essential to develop defensible processes for moderation, verification and validation.

Our current strategy for the moderation process is as follows. The data any user views at any given time is the last moderated version. Changes to a given record (e.g., information on a power plant) are placed in a stack called "history of edits" until moderated. Users can examine all entries in this stack and the differences from any previous version of that record in a view only mode. The process of moderation evaluates all the pending contributions to a given record and accepts and incorporates into the main database all changes that can be validated. All contributions that are plausible but require further verification and validation are entered into the discussion forum for further comments and input. Thus, users can check on the status of their contributions instantly (by viewing the stack) and follow what actions have been taken and why through the discussion forum. This process is illustrated in Figure 1.

Validation and Verification Process

To maintain the scientific integrity of such public databases that house information from multiple sources – official databases, academics, experts, environmental organizations, and the public – constant verification and validation is required. In addition to constant scrutiny by the users, some of the processes and tools we have incorporated and will develop further include:

- Validation of geospatial information on large infrastructure (power plants, mines, ports, etc.) can be done by inspection by any internet savvy user in regions for which Google Earth has recent high-resolution imagery or if they are knowledgeable of the geolocation. Over the long term we envisage integrating automated pattern recognition tools, however, at present the human eye remains the best option. This seemingly enormous task becomes possible with successful public participation – any user with knowledge of the location of the plant and willing to use our interface to Google Maps (or view the infrastructure use Google Earth or any other GIS tool) could verify and validate the geo-spatial information.
- A framework to help the editors and moderators integrate this heterogeneous input and develop a process for validation. Over time, based on this experience, we propose to develop increasing levels of automation.

- Performing consistency checks on the data. Routines to exercise these checks will be built into the database and run in the background to flag anomalies. For example, knowing the composition of the coal being burned in a coal-fired power plant and the details of the generating units and the scrubbers provides reasonable estimates of the emissions. Using these routines we will flag anomalies that will be entered in the discussion forum for input. A second check will be to compare performance figures of plants within similar operating environments and similar components and quantify the range of variations. Another integrity check will be whether changes in time histories of performance or emissions records are correlated with by recorded changes in infrastructure and/or infusion of new technology.

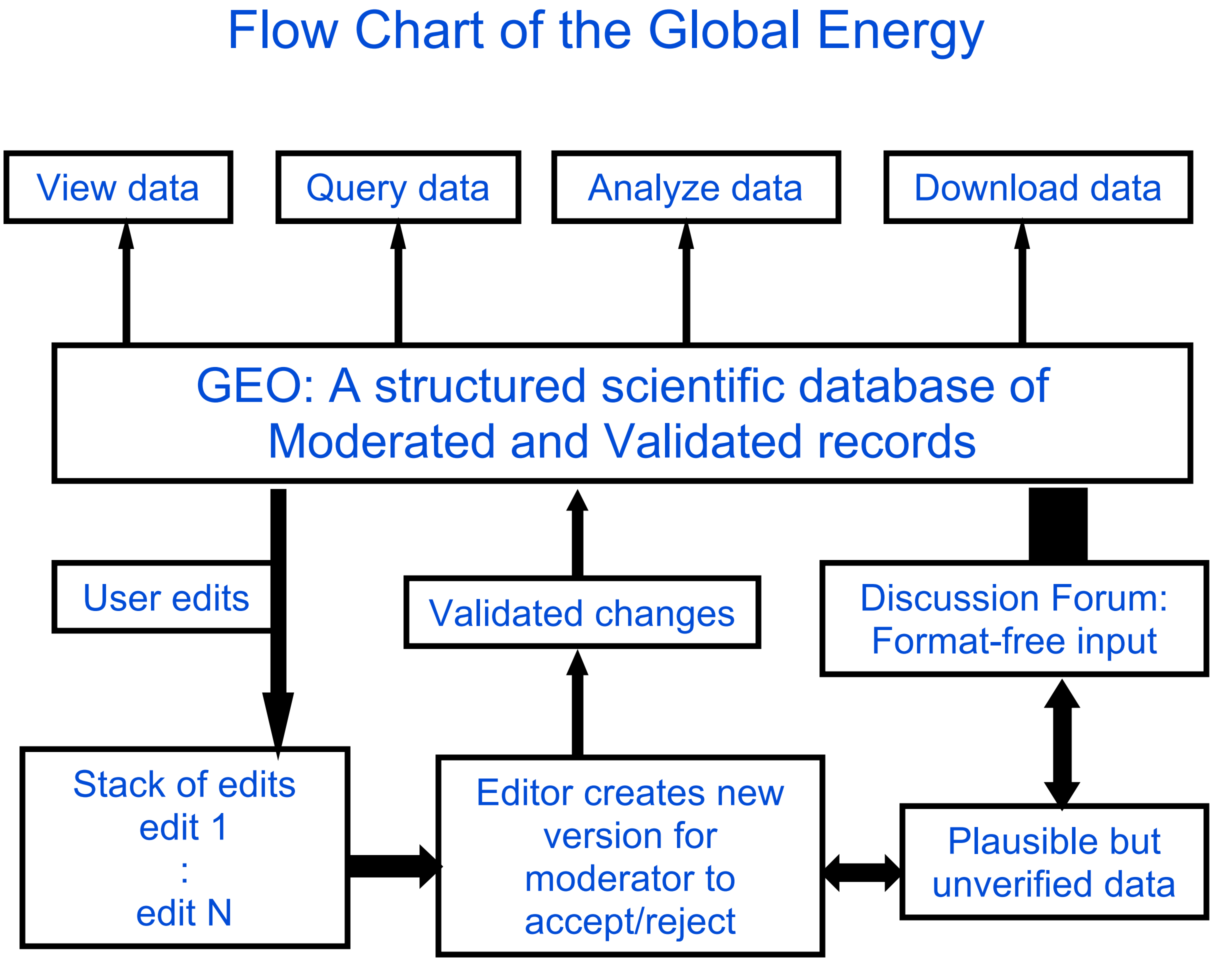
Flow Chart of the Global Energy
View data
Query data
Analyze data
Download data
GEO: A structured scientific database of Moderated and Validated records
User edits
Validated changes
Discussion Forum: Format-free input
Stack of edits
edit 1
:
edit N
Editor creates new version for moderator to accept/reject
Plausible but unverified data

Figure 1: An organizational overview of the tools and moderation process implemented in the current version of the Global Energy Observatory (GEO).

Measures of Accuracy of the Data and Trust Management

In many parts of the developing world, in particular China and India, energy systems are being installed at a very rapid rate. In the industrialized world they are being replaced and beginning to evolve to more environmentally benign solutions and are incorporating more renewable options. Developing an integrated understanding of the dynamics of change is still elusive as information is fragmented, incomplete, of uneven quality and accepted standards for reporting and managing data are lacking. Analyzing the fast dynamics of change and increasing complexity of options and players requires such scientifically motivated databases to include a measure of accuracy of the data, trust of its contributors, and an assessment of its completeness and relevance.

To address these important issues of trust and transparency we propose to implement measures of accuracy for both the data and its contributors. To track accuracy, all visitors to a page (a record of, for example, a coal-fired power plant) will be asked to rate the accuracy of existing information, and their score and the time of submission will be recorded. In addition, users will also be able to view graphs of the time history of these scores and their distribution. Correlations between these scores and improvements and/or enhancements in data with each successive moderated version will be made manifest on the time history graph by superimposing the dates on which new moderated versions were added to the database. To prevent large-scale manipulation and bias through multiple ratings by the same user, any given user will be allowed to rate a given moderated version of the record only once. To rate the plant again, they will have to wait for the next moderated version of that plant.

Completeness of data is subjective and depends on the questions being addressed. To address the issue that people are asking new questions in response to new opportunities provided by data and computational tools, we will ask users to identify missing fields and the analysis these fields will facilitate. These fields will then be added to the database. Second, most official databases lag by 2-4 years in making public the data they collect. We have found that a large part of these data are often available earlier on the web as part of news articles or reports, so we propose to develop [semi]-automated tools to first identify the sources and then harvest the desired nuggets using well-formulated pattern searches. In this way the databases in GEO will be continuously and organically enhanced by contributions from experts and the public and by harvesting other open databases and the WWW in general.

To manage ratings of trust of the contributors, the editors, as part of the moderation process, will assign, for each contributor and contribution, a score capturing the accuracy of data provided and the quality of associated documentation needed for verification. The database will maintain this score, the username of the contributor and the editor, time of submission and the plant ID of the energy system edited. From this information we will build a trust (reliability) index and profile of frequent contributors segmented by trust classes defined by geographical and subject areas. Such an assignment of *reputation-based direct trust*, i.e., belief in the capabilities of a contributor with respect to a given trust class, and models of deriving trust relationships have been discussed in [1,2,3].

Once there exists a statistically significant distribution of scores for a particular contributor, we propose to use these trust ratings to help automate the moderation process. The expectation is that if the mean score reaches a certain threshold, further data contributed by that user for the energy systems judged to be within their domain of expertise (trust class) would automatically go to the moderator for acceptance or even directly into the database. Furthermore, individuals with high trust ratings will be added to the pool of potential editors and moderators.

Analysis

GEO currently provides simple analysis and visualization features. These include (i) the data can be downloaded as KML files for mashup and visualization on Google Earth, and (ii) country profiles of power generation (time line of installed capacity, power generation, and emissions by year) and visualization of geo-spatial location of the power plants. Once the framework is complete, we will focus our attention on developing additional analysis tools along with continuous refinement of the overall framework.

Education and Outreach

A major goal of this project is education, both formal and of the public at large. With respect to formal education we aim to provide GEO initially as a resource for class projects. Over the next three years we propose to initiate a collaboration with authors of text books leading to the integration of GEO into a undergraduate curriculum on energy and environmental issues. For the public at large our goal is to provide awareness of both the challenges and opportunities that brings the public (consumers) and corporations (providers) together. Energy systems are big business and corporations that figure out how to align with the demands of the consumer, with respect to electric power generation, transportation fuels, technologies for efficient energy use and environmental protection, stand to reap rich rewards. The goal of the GEO databases and analysis tools is to create an environment of cooperation through understanding rather than an adversarial relationship based on mistrust and perceptions.

To evaluate the quality of the web interfaces and data structures, and the efficacy of the GEO framework to facilitate student projects we collaborated with undergraduate classes at three universities, UC Berkeley, Carnegie Mellon and York University, Canada during the fall semester of 2008.

The UCB students were asked to navigate through the database to determine if it was well organized and documented and whether it contained data of interest. The feedback from the students and instructor was very valuable for improving the framework. The reports of students of York University show very high variability – ranging from some who did not understand the framework and how to gather specific information to others who followed multiple leads including calling the utilities and collected significant amount of available data. Twelve CMU students undertook a 30-hour project to gather missing information on energy generation in seven different regions of the world (UK, Taiwan, Hong Kong, India, Mexico, Connecticut, and Texas). They worked in teams of 1-3 students. The goal was to understand current systems and how these regions plan to meet future energy needs and reduce environmental impacts. This exercise of data collection and evaluation exposed students to understanding how fragmented the data actually are when one wants to answer a specific question and how hard it is to get response from power companies to resolve differences between two “official” sources.

The overall lessons we have drawn from the feedback from student participants and their instructors has been that this hands-on experience of the real-world situation is an eye opener in terms of the system's complexity and how hard it is for the public to form educated opinions based on information that is most readily available. From our perspective, the experience gained from working with these three institutions has been very encouraging and rewarding. Our goal is to expand this educational outreach to other campuses and work with faculty teaching courses on energy and environmental systems to integrate GEO into their curriculum both as a learning tool and for helping build the databases.

There are two other web-based efforts we are in touch with that are mapping energy systems and their environmental impacts – the CARMA [4] and Vulcan projects [5]. We plan to share information and tools with these efforts. Overall, the three projects bring very different motivations and strengths but there is a commonality of purpose – to reduce harmful emissions and motivate the development and deployment of clean energy systems.

Conclusions

The motivation behind the Global Energy Observatory is to build a framework through which a large community of people, both experts and the public, can understand and analyze global energy systems. It is designed as a one-stop site for comprehensive information in a scientific format that would help policy makers, energy specialists and providers, academics and educationists. To maintain transparency, build trust and have real time updates we seek public participation through volunteered contributions. We believe such a framework is also required to understand a number of other global challenges that are intrinsically socio-technical-economic-political and require an integrated approach for sustainable solutions. These include water resources, bio-diversity, habitats, ecosystems and their management, public health and terrorism. We have focused on the EEDC challenge because of the global urgency, and because it offers a rich combination of technology, infrastructure and social phenomena occurring at multiple length and time scales. We hope to translate the revolution in information science and computing into novel tools for education on the EEDC challenge, both of the public and as part of classroom curriculum, and for developing novel solutions.

References

[1] D. Artz and Y. Gil. *A Survey of Trust in Computer Science and the Semantic Web.* Web Semantics: Science, Services and Agents on the World Wide Web, 5(2): 58-71, Elsevier Science, 2007.

[2] T. Beth, M. Borcherding and B. Klein. *Valuation of Trust in Open Networks*, in Lecture Notes in Computer Science: Computer Security — ESORICS 94, pp.1-18, Springer, Berlin, 1994.

[3] M. Blaze, J. Feigenbaum and J. Lacy. *Decentralized Trust Management*, in SP '96: Proceedings of the 1996 IEEE Symposium on Security and Privacy, pp. 164-174, IEEE Computer Society, Washington, DC, 1996.

[4] CARMA project, http://carma.org/

[5] Vulcan project, http://www.purdue.edu/eas/carbon/vulcan/index.php