Visualizing World Bank Indicators through Google Earth

William Murakami-Brundage, MS; Jennifer Bopp, MA; Megan Finney; Joselito Abueg, MA

Abstract— The goal of the project is to develop a large visual data resource for Google Earth using major education, gender, and health datasets. With global data increasingly being made public by organizations such as the World Bank, global data modeling has been a significant development in information visualization and geographical information systems. While there is a considerable amount of publicly owned and open-source data sets available, there has been minimal development beyond proof-of-concept ideas. The current research project is to model five major domains of the World Bank's global datasets. The global datasets being modeled are education, women and gender issues, health care access, life expectancy and wellness, and access to the Internet and technology. Each dataset is comprised of between 5-15 sub-elements that are being modeled as well. A data translation key has been developed for data migration into Google Earth's KML format, and modeling is expected to be finished by April 2011. After the Google Earth models are complete, the resulting KML files will be made available for public use. It is hoped that a greater global awareness will develop by using the World Bank/Google Earth data. Additionally, data development will be easier once the data key is published.

Index Terms— Data and knowledge visualization, Spatial databases and GIS, Visualization systems and software.

1 Introduction

THE mission of the World Bank is to reduce poverty by supplying resources and knowledge that leads to improvements in infrastructure and sustained development. To achieve this goal, the World Bank engages in partnerships with public and private sector entities and promotes awareness of issues related to global development. The World Bank publishes reports, statistics, and data sets to support program management, accountability, and academic research. These include over 2,000 economic and human development statistics for 209 countries. In April 2010, the World Bank made most all of these data available on the internet under its Open Data initiative. Data that were previously only accessible to paid subscribers became available for download or through a published Application Programming Interface to anyone with internet access.

It is hardly novel at to suggest that we now live in a globally-connected era. The internet has played a significant role the more recent developments of globalization, and information from around the world is more accessible than it ever has been in history. But with so much information—from sources trustworthy and otherwise—it remains important to not only bring attention to meaningful data, but to present it in ways that promotes new ways of deeply understanding that data. Information visualization can illuminate relationships in data that enables new insights and more contextualized and significant apprehension. Our primary goal is to make the World Bank indicators readily available to users in a compelling form. We want to explore a way in which the user could visualize the data with no programming and with tools that were freely available.

After the World Bank made its data publicly available, it launched its *Apps for Development* competition to encourage developers to make use of the data (World Development Indicators, or WDI) in ways that could support the UN's Millennium Development goals. While several of the entries were concerned with promoting awareness through visual presentation of these datasets, none appeared to make use of Virtual Earth (VE) visualizations. Google has made available some of the World Bank's indicators and several methods to create visualizations on its Labs site. But this site is not as prominent or as widely known as Google Earth. Google Earth's compelling interface and the fact that its software and data are freely available for different platforms, make it an important, educational tool to promote global awareness. Making available World Bank indicators in Google Earth's KML format would therefore be highly desirable. We have started with indicators for *Expected Years of Schooling, Male; Expected Years of Schooling, Female;* and *GNI per Capita, Atlas Method*.

One final goal for this project is to establish a streamlined and cost-free way to convert data into a KML format so that the technology is accessible to a wider user population. There are currently options available for converting data into KML format, but these are cost prohibitive for general use.

After a brief overview of the World Bank data and Google Earth, we discuss the process undertaken to convert World Bank indicators into KML data, as well as some of the challenges involved. Next, we will evaluate the resulting visualizations on Google Earth. Finally, we will assess future directions that could further this effort.

1.1 The World Bank

The World Bank is an international organization with 187 member countries, dedicated to combating poverty globally through grants, interest-free loans, and strategic partnerships to address such issues as education, health, infrastructure to name just a few.

In 2009, the World Bank announced its API to encourage developers to create applications that used its data. To further encourage these efforts, the World Bank launch its *Apps for Development* competition in October of 2010, offering a total of $45,000 in awards for applications that addressed issues identified in the United Nation's Millennium Development Goals. According to the World Bank (http://data.worldbank.org): "[a]pplications were submitted from 36 countries across every continent; more than half came from Africa, Asia, and Latin America." Though the Apps for Development competition is closed for new entries, it serves as inspiration for the current project.

1.2 Visualizing World Bank Indicator Data

Providing a tool to better understand the roots of economic disparities and leverage the World Bank's data catalog resonates with the principal goal of visualization: namely, generating insights [1]. In the year prior to the public release of its full bank of WDI, the World Bank partnered with Google, providing data from 17 of the datasets to be searchable and displayed in simple graph formats on Google. In March of 2010, Google Labs launched its Public Data Explorer, which now included 54 of the World Bank's WDI, and the ability to create line and bar graphs as well as maps and bubble charts. Figure 1 provides one such example.

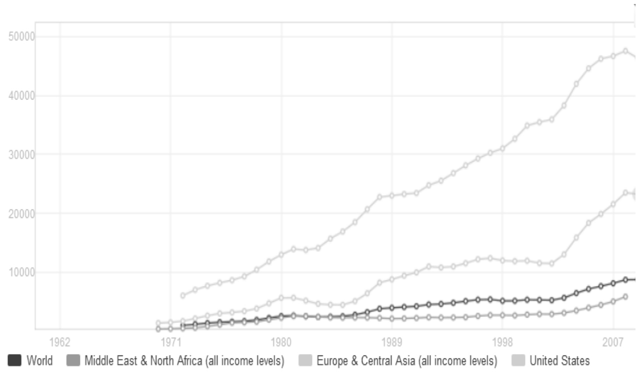


Fig. 1. Gross National Income (GNI) per capita over a 30 year period on an interactive graph found on the World Bank data website (http://data.worldbank.org/indicator/NY.GNP.PCAP.CD/countries?display=graph). A visitor to the website may choose to display any number of individual countries or regions on the graph.

The World Back anticipated the importance of visualizing it data, and provided ways to map and graph the data. However, the methods by which these data are represented are not without issues. For example, in simple line charts as in Figure 1, care must taken so as not to present too much data that individual countries cannot be identified, especially in a static display. These kinds of visualization also rely on the view to recognize any geographical influence on the data.

On the other hand, overlaying data on a map makes it easy to visualize the importance of location/country on the variable of interest, and is an obvious choice of data coded by country. However, flat maps are in fact, a projection of three dimensional object (i.e., the earth). This introduces actual spatial errors (relative sizes and locations of borders are incorrect), whose degree depends on the type of projection made (e.g., Mercator vs. Robinson). Just as actual maps are "warped" representations of reality, so are cognitive ones (e.g., [2],[3],[4]). Fortunately, though people make errors, they seem not to be influenced by poor global-scale geographic information at least when it comes to area estimations by young adults [5]. Young children might not fare so well.

That said, maps are understood to carry with them certain biases that can influence how their users perceive the world [6]. For example, plotting countries on a flat map requires the cartographer to make decisions about what location is center of the two-dimensional world map, and how to orient the countries from left-to-right (see Figure 2). Of course, by convention, north is up, and countries are centered about the European continent. But, at least with respect to the mission of the World Bank, centering a flat map on the African subcontinent or other regions of the developing world might yield greater impact.

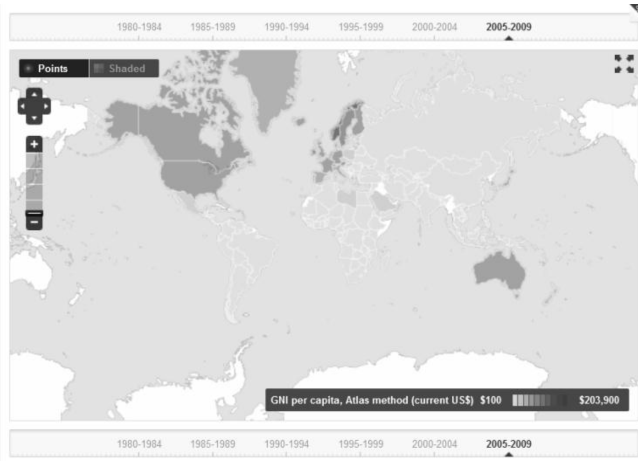


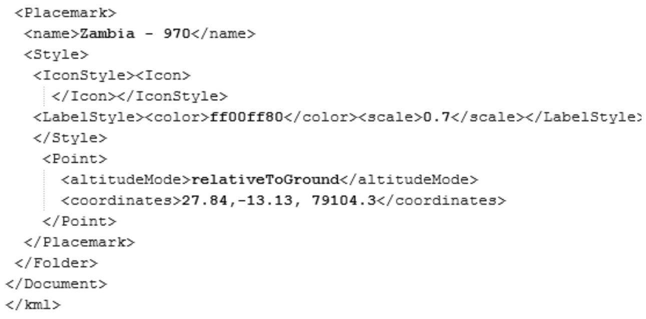
Fig. 2. Gross National Income (GNI) per capita over a 30 year period on an interactive map found on the World Bank data website (http://data.worldbank.org/indicator/NY.GNP.PCAP.CD/countries?display=map). Though an effective way of presenting data that are intrinsically geographical, compromises are made.

In part to bring greater understanding to issues such as these, in 2006, the National Research Council called for K-12 curriculum changes that gave an explicit role for development of spatial thinking (see review in [7]). They also suggested that geographical information systems may support and extend cognition, which appears to implicitly assume that interactive systems which have representational fidelity can make a difference in children's understanding. Physical globes have long been a part of the classroom; their virtual earth counterparts offer a compelling way to visualize geographically-related data in context.

1.4 Google Earth

Google Earth was launched in 2005 as a free downloadable, highly interactive, geographical information application. Users are easily able to navigate and explore the Earth, turning or tilting the globe, zooming rapidly from full globe view to close views of specific locations. Movement is rendered in a smooth, continuous "flying" manner that allows for a stimulating sense of context. A wide variety of information (e.g. shipwrecks, monuments, tours, routes, etc.) can precisely layered onto the land (and ocean) with the use of Keyhole Markup Language (KML), a tailored version of XML (see Table 1). Users are able to create and open KML data files on their desktops, or submit them to be uploaded to be available for other users.

TABLE 1
KEYHOLE MARKUP LANGUAGE (KML) FRAGMENT



```
  <Placemark>
   <name>Zambia - 970</name>
   <Style>
    <IconStyle><Icon>
      </Icon></IconStyle>
    <LabelStyle><color>ff00ff80</color><scale>0.7</scale></LabelStyle>
    </Style>
     <Point>
       <altitudeMode>relativeToGround</altitudeMode>
       <coordinates>27.84,-13.13, 79104.3</coordinates>
     </Point>
   </Placemark>
  </Folder>
 </Document>
 </kml>
```

The text fragment above showed the end of the GNI per capital file. Each World Bank Indicator (e.g., Gross National Income per capita) data set (values by country) is geocoded such that a specific location for each county is determined. The geocoded data are then encoded into a Keyhole Mark Language format document, which is readable by Google Earth.

Since its launch, Google Earth has become an important tool in many contexts, ranging from the classroom to tracking human rights violations remotely (Toney).

2 METHODOLOGY

The project utilized the publicly available World Bank Gender Statistics dataset. The program GE-Graph was selected to transform the data to the Google Earth proprietary KML format. The development process from dataset acquisition to 3D geospatial visualization will be detailed in the following selection, applying Chi's Data State Model as a framework for understanding key elements of the project and its scope.

2.1 Data State Model

Chi [10] introduced the Data State Model as a comprehensive taxonomy of information visualization techniques. According to the model, data progresses through four distinct Data Stages via three distinct Data Transformation steps in order to arrive at the final visualization. The four Data Stages are: Value, Analytical Abstraction, Visualization Abstraction, and View. The three types of Data Transformation are: Data Transformation, Visualization Transformation, and Visual Mapping Transformation.

Figure 3 displays the transformation process of the current project as interpreted within the Data State Model. It is evident in the figure that the Gender Statistics data set technically constitutes the second Data Stage, Analytical Abstraction, and so the discussion of project methods will begin at this stage.

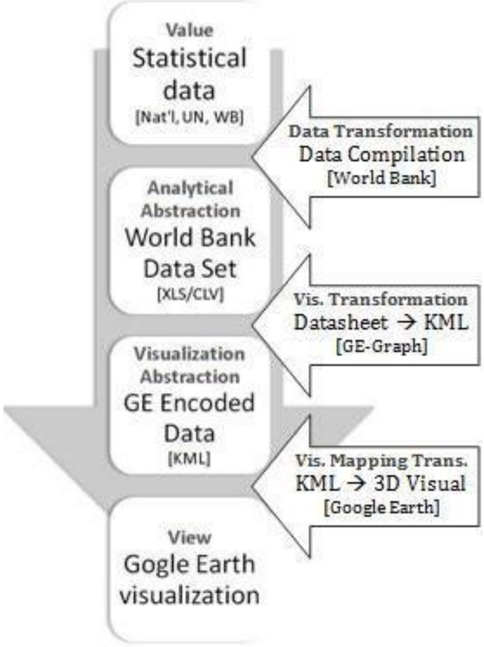


Fig. 3. The Data State Model applied to the World Bank visualization project. Work by project members begins at the Analytical Abstraction stage of the process.

2.1 Gender Statistics Datasheet

The Gender Statistics database (*GenderStats*) is a comprehensive resource of vital gender statistics gathered from national agencies, the United Nations, and World Bank-conducted or funded surveys (http://data.worldbank.org/data-catalog). The database is available for download in the .XLS (Microsoft Excel) and .CSV (simple datasheet) formats.

The project begins in earnest at the Analytical Abstraction stage (cf. Data State Model), after Value stage data, or raw data, has been compiled in a form amenable to analysis. The *GenderStats* datasheet was retrieved from the World Bank website and downloaded to a local drive on the project developer's personal computer. The project developer undertook all data manipulation further identified in this discussion on the same personal computer. The computer used was an ASUS CM5571, Pentium Dual-Core 2.7 Ghz with 6 GB of RAM. It was concurrently being used for typical daily use. Thus, the speed of data processing would be vastly higher on a higher-capacity or dedicated processing system.

2.2 Master Index

During the course of data transformation, the developer discovered an incompatibility of ISO country codes. World Bank data identifies countries based upon the ISO 3166-3 standard of three letters (e.g., USA). Google Earth labels countries according to the ISO 3166-2 standard of two letters (e.g., US). While seemingly a minor difference of one letter, this proved problematic to the transformation process. Wikipedia's entry on ISO-2 allowed the developer to cross-reference ISO-2 codes effectively (http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2).

As a solution, a master index was created in Microsoft Access 2007. This may be considered an extension of the initial Analytical Abstraction stage. It allowed for ease of perusal and efficiency in re-coding the data for processing in the Visualization Transformation stage. Using Google Maps allowed the developer to further develop the index by locating the latitude and longitude for each country's ISO-2 code. This was entered into the master index. Lastly, some entries had to be eliminated from the World Bank datasets, specifically those focused on regional areas and general economic status. After the refinement, the developer was left with a functional, ISO-2 coded, cross-referenced Microsoft Access database that could be sorted and viewed according to variable, country, and time span.

2.3 GE-Graph

One of the major goals was also a major challenge of the project: discovering publicly available free software to perform data processing and transformation. The intention is to inspire open-source systems. However, it required several weeks of searching and encountering issues with software that was either free but incompletely supported and/or poorly documented, or else well-crafted but financially unfeasible.

The solution came care of Brazil. GE-Graph is a publicly available "freeware" (free of charge) program developed by Ricardo Sgrillo of the Escola Superior de Agricultura Luiz de Queiroz (ESALQ) of the University of Sao Paolo (http://www.sgrillo.net/googleearth/gegraph.htm). GE-Graph is intended to aid in the import and export of KML data, the file format used by the Google Earth program. For the purposes of this project, GE-Graph's import functions were utilized to transform the *GenderStats* datasheet file format to a KML format. There were many other program features that unfortunately could not be investigated due to time constraints, but remain of interest to future developments.

The project developer imported the re-indexed *GenderStats* datsheet file to GE-Graph manually, one year at a time. Due to GE graph limitations, every variable needed to be imported and processed on an annual basis. The program processed the datasheets and output the data in KML format, taking roughly 10 minutes per annual data set. Due to a system limitation, the KML files generated are non-timespan supported; data had to be processed by individual years. This required a long-term intensive approach to processing the dataset.

2.4 Google Earth Visualization

Once processed, GE Graph output the *GenderStats* data in KML format. As previously discussed, KML (Keyhole Markup Format) is a geographically enhanced extension of the XML (Extensible Markup Language) standard. In Google Earth, KML files incorporate geographic elements such as placemarks, ground overlays, paths, and polygons. All of these elements are customizable and may be tailored to suit the developer's stylistic choices.

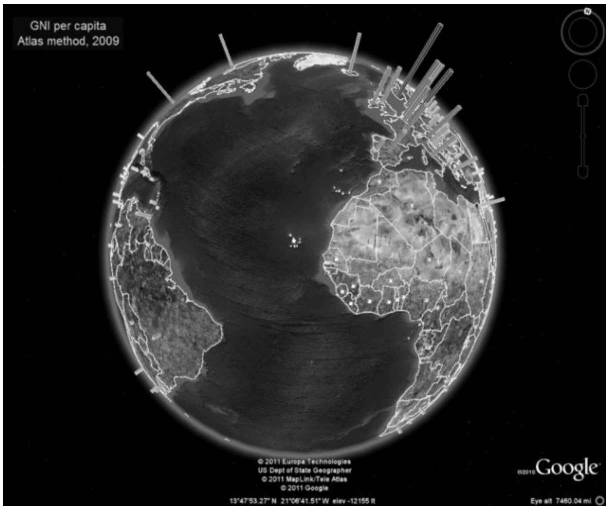

GNI per capita
Atlas method, 2009
© 2011 Europa Technologies
US Dept of State Geographer
© 2011 MapLink/Tele Atlas
© 2011 Google
13°47'53.27" N 21°06'41.51" W elev -12155 ft
©2010 Google
Eye alt 7460.04 mi

Fig. 4. Gross National Income (GNI) per capita in 2009. Note the high incomes in the US and Europe related to the African continent. Also, note the orientation of the earth required to see this.

3 Results

Three dimensional rods were chosen from a variety of options for this project. Due to time constraints, a simple approach was preferable. The height of each rod correlated to the value of the variable processed by the GE-Graph program (e.g., income, education level). Figure 4 provides an example of the rods generated in Google Earth.

While the code itself was initially generated by GE Graph, it may be altered simply by working within the Google Earth application or by opening the KML file in a text editor (cf. Table 1) and manually making changes to the markup. It is more time-consuming and coding-intensive to perform these edits manually. Thus, within the context of this project, manual coding was not a viable option unless absolutely necessary.

We used blue, round rods to represent the Expected Years of Education (for males) and yellow-green, rectangular bars that encased the rods to represent GNI. Due to development tools, graphics were restricted to simple polygonal shapes, but height, width, and color could be altered when data processing occurred. Another data development issue was that the core KML files could not be easily altered once the graphic data was modeled; in order to change even a small feature of the Google Earth graphic, the variable or year in question would need to be re-generated.

4 Discussion

Using a virtual model of the world such as Google Earth allows the World Bank data to be displayed in geographic context and should provide a greater understanding of that data through locating it. Yet the evidence is equivocal. For example, Tavanti and Lind [9] found that that three dimensional graphics more strongly supported spatial memory. However, Cochburn [10] ran a similar experiment adjusted a few of the previously uncontrolled parameters, and found that dimensionality did not greatly impact spatial memory. Both studies, however, worked only with static 3D graphics rather than navigable virtual world graphics (but see Koua et al. [11] and Keehnerm et al. [12]). While it is beyond the scope of this paper to evaluate the benefits of virtual spaces (3D representations), we suspect that the ways in which a virtual world matches our mental model would make a stronger sense of space and context than would a two dimensional, static model.

Using three dimensional shapes such as rods, the heights of which correspond to a desired value, should allow for a broader overview and prevent the occlusion that might result with two-dimensional objects with varying sizes. Nevertheless, the resulting visualizations in these first stages of design pose some difficulties in terms of usability and communicating data.

One significant problem is the difficulty in achieving the necessary position to attain data. Labels that identify the country and the variable are only visible at a specific distance to the rod. If the user zooms out too far, the labels disappear. Zooming in too close also results in the loss of labels, as well as the solidity and sense of height of the rod. When in position to see the data, the user is hovered over the rods, resulting in a severely foreshortened shape that does not visually communicate its value. The circular ends of the rods are given a color, the intensity of shade of which corresponds indicates the range in which the value falls (with a key included at the bottom of the screen). To compare countries in close proximity (e.g. European or African countries), the user must be both at a distance and an oblique angle to see the rods in a sort of three-quarter view. Of course, this means that labels and even a sense of which country a rod represents are lost. One possible solution might be to find a way to render cast shadows from the rods. Both the Education and the GNI variables displayed together is only effective (i.e. a visual ratio is apparent) if the height of the GNI bar does not surpass the height of the education rod.

Another significant difficulty is displaying time series data. The user can go into the data folder and check or uncheck as many years of data as desired. But if multiple years are displayed, the labels positioned along the rod appear somewhat chaotic. In the case of closely positioned countries, the density of labels makes them nearly illegible. Making sense of the information conveyed by the labels is further complicated by the fact that the corresponding year is not included with each. Programming the dataset with Google Earth's timespan feature is the next step for the World Bank variables. The first step to modeling data over time is to develop all the necessary years in Google Earth. After this, the data will be joined together to form a cohesive, timespan-capable file.

Shneiderman’s [13] visual information seeking mantra, “Overview first, zoom and filter, then details-on-demand” might be instructive for considering what would benefit these visualizations. The overview may be the most successful aspect of the visualization; as the user turns the globe, the variable heights of the rods are useful for giving a general impression of different regions of the globe. Additionally, we can zoom in on areas of interest. But as discussed earlier, zooming in must be at a fairly specific degree and creates visual distortion. Labels may or may not be informative, depending on how many years are selected. And this brings up filtering and details-on-demand. While the user is able to filter by checking or un-checking years or countries in the data folder, this operation is cumbersome and the former view cannot be held in the memory for comparison.

Difficulties aside, some striking comparisons can be made, such as the discrepancy in income between Eastern and Western Europe, when the education levels are roughly equivalent. Education in Eastern Europe does not seem to result in a higher income.

While Eastern European countries have roughly the same levels of education as Western European Countries, their GNI figures significantly lower. But one must discern this through the density of labels for multiple years that obscure the representative shapes and do not organize in an informative way.

5 Future Work

Google Earth has addressed 3D primarily in terms of allowing users to create buildings and cities. This interest in surface features might provide a clue to some of the difficulties in displaying data in a satisfying way. While Google Earth has an enormous range of zooming, most layers refer to physical features on the surface. The current programming features may just not have developed to enable effective ways of visualizing abstract data.

Further goals, therefore, would be to find ways in which data could be visualized with a wider array (and perhaps easily changeable) shapes, as well as an effective use of color. While we would like to have had time to examine the Time Slider feature of Google Earth, another effective way of differentiating years might be to use different colors to create strata along the rods. Improved labeling that would indicate the year along with the numeric data would clearly be desirable. Labels might additionally match the color of the strata to which they refer.

The ability to display multiple variables simultaneously in a useful way would also be significant improvement. Our use of rods and shapes that enclosed the rods was somewhat successful, but had serious limitations. Perhaps enabling two varying shapes to be side by side within the country's parameters without being in the exact would be helpful, although the small size, or density of layout of some countries would not make this an ideal solution. Moving toward a solution might lie in improved navigation where one could come close to shapes, circle around them, and scale their heights. Incorporating animation in some way would also be valuable.

KML files are non-proprietary, but must function with the host application's framework. There are several two-dimensional geographical information system programs, including some notable open-source contenders. Given enough resources, cross-capability and higher levels of functionality can be coded within the KML files. Working within the open-source framework is somewhat resource limited, but pushing the open-source envelope can grant global access to everyone.

Lastly, the farthest reach of visual systems is modeling in a true 3-dimensional space, with a true X, Y, and Z axis denoted for every point of interest. Current technology may be reaching this tipping point of data display, or it may be some years away. When the day arrives, effective geographic information systems will need vastly more processing power and high-quality modeling algorithms.

All of these wished-for features to produce an effective visualization would require easier ways for the user to interact with the data in order to allow for simple filtering, details-on-demand, and the ability to combine, compare and change the parameters of the data producing the visualization.

References

[1] C. Chen. "Information visualization, " *Wiley Interdisciplinary Review: Computational Statistics*, vol. 2, no 4, pp. 387-403, July/August 2010.

[2] A. Friedman and N. R. Brown, "Reasoning About Geography," *J. Experimental Psychology: General,* vol. 129, no. 2, pp. 193-219, 2000.

[3] B. Tversky, "Distortions in Memory for Maps," *Cognitive Psychology,* vol. 13, pp. 407-433, 2000.

[4] M.J. Egenhofer and D.M. Mark, "Naïve Geography," Technical Report 95-8, National Center for Geographic Information and Analysis, June 1995

[5] S.E. Battersby and D.R. Montello, "Area Estimation of World Regions and the Projection of the Global-Scale Cognitive Map," *Annals of the Association of American Geographers, 99*(2), pp. 273-291, 2009.

[6] J.B. Hartley, "Deconstructing the Map," *Cartographica: The International Journal for Geographic Information and Visualization,* vol. 26, no. 2, pp. 1-20, Summer 1989.

[7] D.R. Montello, "Review of `Learning to Think Spatially' by the Committee on Support for Thinking Spatially: The Incorporation of Geographical Information Science Across the K-12 Curriculum, the National Research Council," *J. Environmental Psychology, 28*, 104-106.

[8] E.H. Chi, "A Taxonomy of Visualization Techniques using the Data State Reference Model," *Infovis '00*, pp. 69-75, 2000.

[9] M. Tavanti and M. Lind, "2D vs 3D, Implications on Spatial Memory," In *INFOVIS '01 Proceedings of the IEEE Symposium on Information Visualization 2001* (Washington, DC: IEEE Computer Society), 2001.

[10] A. Cochburn, "Revisiting 2D vs 3D Implications on Spatial Memory," *User Inferfaces*, vol. 28., 2004.

[11] E. Koua, A. MacEachren, and M-J. Krak,."Evaluating the usuability of visualization methods in an exploratory geovisualization environment," *International Journal of Geographical Information Science,* vol. 20, no. 4, pp. 425-428, 2006.

[12] M. Keehnerm, M. Hegarty, C. Cohen, P. Khooshabeh, and D. R. Montello, "Spatial Reasoning With External Visualizations: What Matters Is What You See, Not Whether You Interact," *Cognitive Science,* vol 32, pp. 1099-1132, 2008.

[13] B. Shneiderman. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," In *Proc. of IEEE Symposium on Visual Languages,* Los Alamos, pp. 336-343, 1996.