

The View from the Chathams: Geovisualisation of Web Site Hits using Google Earth™

*Nigel Stanger*¹

¹Department of Information Science
University of Otago, Dunedin, New Zealand
Phone: +64 3 479-8179 Fax: +64 3 479-8311
Email: nstanger@infoscience.otago.ac.nz

Presented at SIRC 2006 - The 18th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
November 6th-7th 2006

ABSTRACT

A useful approach to visualising the geographical distribution of web site hits is to geolocate the IP addresses of hits and plot them on a world map. In this paper we examine the efficacy of Google Earth for this purpose.

Keywords and phrases: web traffic, geolocation, geovisualisation, digital repository, Google Earth

1 INTRODUCTION

When administering a web site, it is normal to want information on the nature of traffic to the site. Information on the geographic sources of traffic can be particularly useful in the right context. For example, an e-commerce site might wish to determine the geographical distribution of visitors to the site, so as to decide where best to target marketing resources. One approach to doing so is to plot the distribution on a map. Geographical information systems (GIS) were already being used for these kinds of purposes prior to the advent of the World Wide Web (Beaumont 1991), and it is a natural extension to apply these ideas to online geovisualisation of web site hits.

In November 2005 the author implemented a pilot digital institutional repository¹ for the University of Otago School of Business (Stanger & McGregor 2006), using the GNU EPrints² repository management software. This repository quickly attracted interest from around the world and the number of abstract views and document downloads began to steadily increase. There was great interest within the University in tracking this increase, particularly with respect to where in the world the hits were coming from. The EPrints statistics management software developed at the University of Tasmania (Sale & McGee 2006) proved very useful in this regard, providing detailed per-country download statistics, as illustrated in Figure 1. However, while this display provides an ordered ranking of the number of hits from each country, it does not provide any further detail below the country level, nor does it provide any visual clues as to the spatial distribution of hit sources around the globe.

The author therefore began to explore techniques for plotting the repository's web traffic onto a world map. There have been several prior efforts to geovisualise web activity. Lamm, Reed & Scullin (1996) developed a sophisticated system for real-time visualisation of web traffic on a 3D globe, but this was intended for use within a virtual reality CAVE (Cruz-Neira, Sandin, DeFanti, Kenyon & Hart 1992), thus limiting its general applicability. Papadakakis, Markatos & Papathanasiou (1998) described a similar 2D system called *Palantir*, which was written as a Java applet and was thus able to run within any Java-enabled web browser. Dodge & Kitchin (2001, pp. 100–103) describe these and several other related systems for mapping Web and Internet traffic.

These early systems provided impressive visualisations, but suffered from a distinct limitation in that there was no public infrastructure in place for geolocating IP addresses (that is, translating them into latitude/longitude coordinates). They generally used *whois* lookups or parsed the domain name in an attempt to guess the country of origin, with fairly crude results (Lamm et al. 1996). Locations outside the United States were typically aggregated by country and mapped to the capital city (Lamm et al. 1996, Papadakakis et al. 1998, Jiang & Ormeling 2000).

¹<http://eprints.otago.ac.nz/>

²<http://www.eprints.org/>










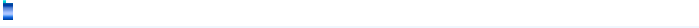

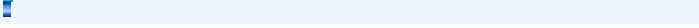

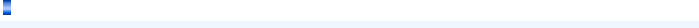

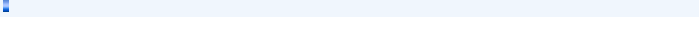
Country	Abstracts	Downloads	
 United States	44186	9308	
 United Kingdom	807	2150	
 New Zealand	1426	1340	
 Switzerland	556	910	
 China	187	593	
 Australia	613	524	
 Otago Intranet	1908	452	
 India	152	388	

Figure 1: A portion of the by-country traffic display for the Otago EPrints repository, generated by the Tasmania statistics software (Sale & McGee 2006)

Reasonably accurate and detailed databases were commercially available at the time (Lamm et al. 1996, p. 1466), but were not generally available to the public at large, thus limiting their utility.

The situation has improved considerably in the last five years, however, with the advent of freely available and reasonably accurate geolocation services³ with worldwide coverage and city-level resolution. For example, Maxmind's *GeoLite City* database is freely available and claims to provide "60% accuracy on a city level for the US within a 25 mile radius" (Maxmind 2006). Their commercial *GeoIP City* database claims 80% accuracy for the same parameters. This means that it is now feasible to generate reasonably precise displays of the geographic origins of web site hits.

Another exciting development in recent years has been the advent of powerful, freely available map visualisation software from Google, Inc. Google Maps and Google Earth both enable dynamic and sophisticated interaction with detailed maps and satellite imagery of the Earth. Google Maps is 2D and works in any web browser, while Google Earth is 3D and runs as a separate application. Both run on commonly available hardware. The arrival of this software means that almost anyone can create sophisticated map visualisations with relatively little effort. Google Earth is of particular interest because of its powerful 3D visualisation capabilities.

The remainder of the paper explores the efficacy of various approaches to geovisualising web site hits, focusing primarily on Google Earth. Section 2 briefly discusses four representative approaches to 2D geovisualisation of web site hits, and the limitations inherent with these approaches. Section 3 then discusses the author's experience of building 3D web hit geovisualisations using Google Earth. Issues that were encountered are discussed in Section 4.

2 TWO-DIMENSIONAL APPROACHES

The author initially focused on 2D map visualisations that could be easily displayed within a web browser, such as that shown in Figure 2. Four representative techniques, corresponding to different distribution styles (Wood, Brodlie & Wright 1996, MacEachren 1998), were identified for implementing these visualisations:

1. *server-side image generation*, where points were plotted onto a base map image at the server, and the single combined image returned to the browser;
2. *server-side image overlay*, where one or more transparent overlay images were generated at the server and returned to the browser along with a base map image, then composited at the browser;
3. *server-side HTML overlay*, where one or more overlays comprising absolutely positioned HTML elements (e.g., DIV elements) were generated at the server and returned to the browser along with a base map image, then composited at the browser; and
4. *Google Maps*, where the browser asynchronously requested map images and overlay data from the server, then generated the final map at the browser.

The scalability of these four techniques was tested in a series of experiments⁴, to see whether they could handle the volumes of web hit data that might be generated by a typical web site.

³Such as <http://www.maxmind.com/> or <http://www.ip2location.com/>.

⁴A paper discussing the results of these experiments is currently under review.

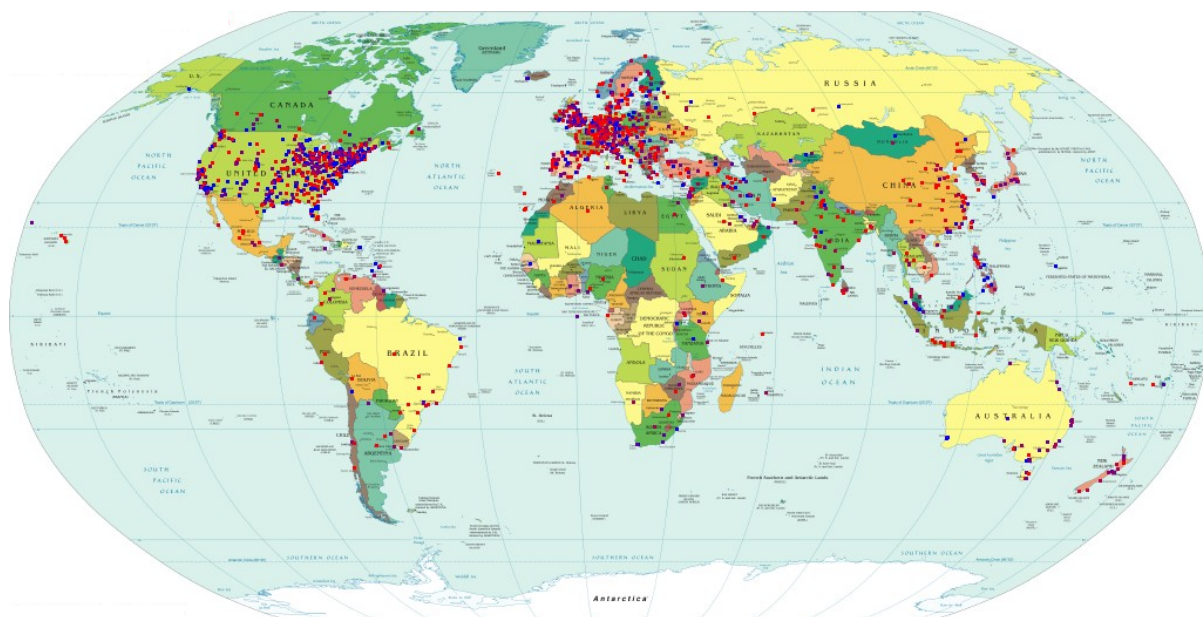


Figure 2: A typical 2D map displaying web hits from the Otago EPrints repository

The server-side image generation and image overlay techniques were found to be the best performers, as shown in Figure 3. Maps were relatively easy to generate and the two techniques scaled well to extremely large numbers of data points, taking on the order of twenty seconds to generate a map with about one million (2^{20}) points. This performance advantage was offset, however, by the maps' static nature and lack of interactivity.

It was expected that Google Maps would provide a more dynamic experience, but as can be seen in Figure 3, it was the worst performer of the group, taking over seven minutes to load a page with only 4,096 points! This was most likely due to memory overload within the browser's JavaScript engine, due to the amount of data being manipulated. Google Maps was thus eliminated as a serious contender.

3 GOOGLE EARTH

Google Earth is a free "Earth explorer" that combines satellite imagery, maps and the Google search engine into a powerful tool for geovisualisation. Google Earth runs under Windows, Mac OS X and Linux, on any reasonably modern PC with a 3D video card. Users can easily create and apply their own overlays, thus gaining much of the visualisation power of a GIS without the associated complexity. Google Earth has been used for many different applications, including topographical analysis (USGS 2006), geology (Thompson, Keith, Swan & Hamblin 2006, de Paor 2006), palaeogeography and archaeology. It is also an obvious "mass market" successor to the earlier resource-intensive 3D geovisualisation systems developed by the likes of Lamm et al. (1996).

Google Earth overlays are specified using an XML dialect called Keyhole Markup Language (KML), and can be displayed or hidden at will within the Google Earth application. KML files are relatively simple to generate, and there are many available for download from the Internet. In addition to static KML files, web sites can provide *network links*, which are dynamic data sources generated by server-side scripts and accessed remotely by the Google Earth application. This enables automatic refreshing of the data on a regular basis.

The author's first KML effort was a simple port of the 2D visualisation shown in Figure 2. Traffic data were extracted from the Otago EPrints repository statistics database (which were in turn derived from the Apache web server logs). The IP addresses were geolocated using MaxMind's free GeoLite City database, and the number of abstract views and document downloads for each distinct geographical location was accumulated. A KML icon was then generated for each distinct location, coloured according to the proportion of abstract views (blue) versus document downloads (red) at that location. This produced an interesting display, but provided no additional benefit beyond "looking cool". In particular, it took no advantage of the third dimension that was now on offer.

The next step was to use this third dimension to display the traffic volume information inherent in the data set. For each distinct location, two KML polygons were generated, a blue one representing abstract views and a red one representing document downloads. The polygons were extruded vertically by 1,000 metres per hit, producing a 3D bar. This approach produced visually stunning images such as those shown in Figure 4. The display is reminiscent of the 3D web traffic visualisations produced by Lamm et al. (1996), but is accessible to a much

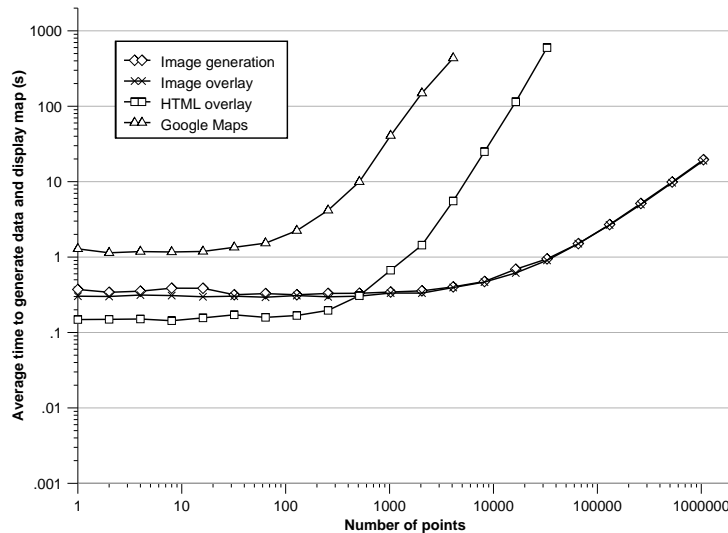


Figure 3: Comparison of page load time for four online map generation techniques (log-log scale)

broader audience. The bar visualisation is highly interactive, and also highly informative when combined with the icon visualisation described in the previous paragraph. For example, while it was apparent from the icon view that the Otago repository had been frequently accessed from Europe, there was no clear indication of the relative levels of traffic across Europe. The bar view, on the other hand, instantly revealed very high traffic levels from Southampton, London, Manchester and Stockholm, in particular.

4 ISSUES ENCOUNTERED

4.1 Geolocation

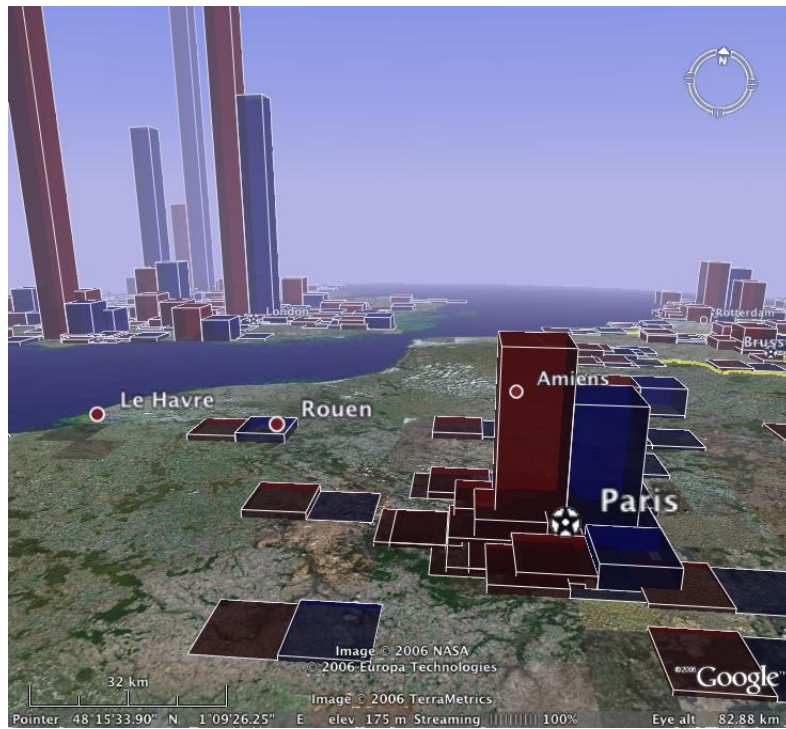
Several issues arose from the geolocation tools used. The geolocation process is independent of the visualisation method, so these issues will affect any attempt at geovisualising data derived from IP addresses.

First, IP address ranges are continually changing and being reassigned across organisations. This means that the physical location of an IP address may change over time. These changes will eventually be reflected in the various geolocation databases, so the same source data may produce different results at different points in time. However, this is only a significant issue if precise measurements are required. In the case of the Otago EPrints repository, for example, such precision is not as important as the ability to easily identify overall access patterns.

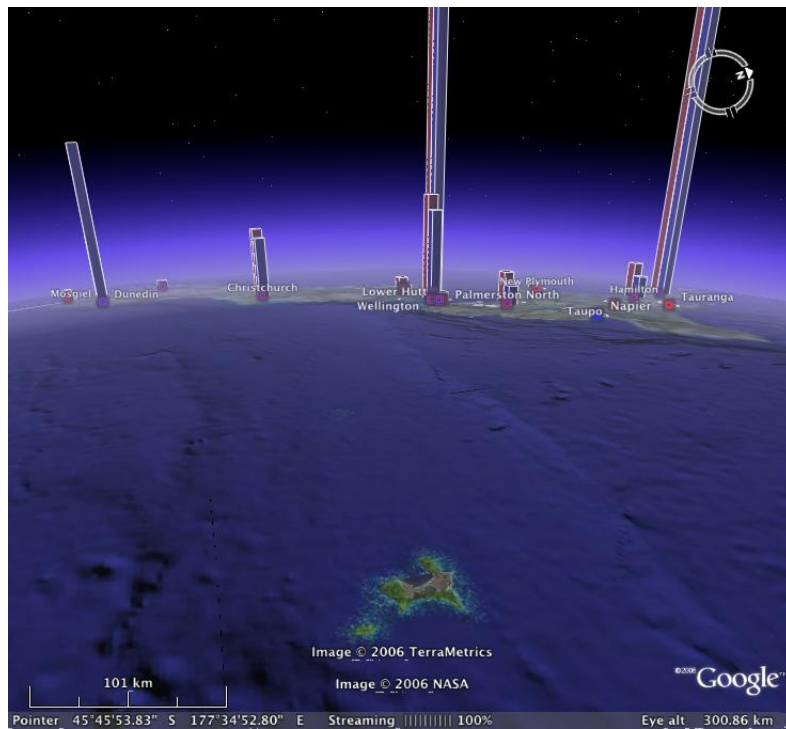
A second issue is that of homonyms and synonyms. *Homonyms* occur when the same name is associated with two or more distinct locations. If the homonyms occur in distinct countries (e.g., London, UK versus London, Canada), then the country name can be used to disambiguate the two. If the homonyms occur within the same country (e.g., Stafford, Georgia versus Stafford, Texas), then more care must be applied. Disambiguation based on the geographic coordinates is an obvious solution, but this is complicated somewhat when the geolocation database returns slightly different coordinates for IP addresses in the same city. *Synonyms* occur when there are multiple names for the same geographic location. Examples noted by the author include Beijing/Peking, Mumbai/Bombay, and somewhat disconcertingly, Louisville/LOUISVILLE. The latter case could be remedied by case normalisation, but there is no simple solution to the other cases.

Third, many IP addresses cannot be successfully geolocated below the country level. In these cases, the geolocation database appears to return a default location in the approximate geographical centre of the country in question. As a consequence, very large apparent traffic rates pop up in some extremely odd places, such as a wheat field in Kansas, the middle of the Sahara Desert or the Marlborough Sounds, or a yurt Outer Mongolia. There appears to be no satisfactory solution to this issue other than simply omitting these data, or perhaps placing them into a separate KML “folder”, which can then be displayed separately.

A final issue arises from the way that traffic data for the Otago EPrints repository were gathered. Two custom “countries” were created representing traffic from the Otago intranet and repository administrative staff, respectively. The former was done in order to track interest from within the University, while the latter was done to prevent routine administrative access from artificially inflating the Otago intranet figures. Since these are stored as distinct “countries” in the statistics database, visualisation generators must special-case them to avoid plotting



(a) View from Paris across the English Channel; the tall bars over England are (from left to right) Southampton, Wolverhampton, Manchester and London



(b) View of New Zealand from the Chatham Islands; note the large spike of non-geolocated traffic in the Marlborough Sounds

Figure 4: Google Earth geovisualisations of web traffic to the Otago EPrints digital repository

them as distinct entries at the same geographic location. In this particular case, the two extra “countries” were re-mapped to Dunedin.

4.2 Google Earth

Only one significant issue relating to Google Earth itself has been noted so far: KML supports only latitude/longitude coordinates. This is not an issue when simply plotting an icon at a given location, but it does cause a problem when drawing a polygon whose vertices are offset a constant distance relative to that location. For example, in the original version of the bar visualisation, the north and south edges of each bar were offset ± 0.075 degrees, or approximately 8 km, from the actual location. A similar offset scheme was used for the east and west edges of the bar. However, the size of a degree of longitude shrinks as one moves from the equator towards the poles, so the bars were square at the equator (approximately 16 km on a side), but became progressively more oblong at higher latitudes.

Fortunately it is relatively simple to compensate for this effect. Using simple trigonometry, a lookup table can be derived that provides a latitude adjustment factor at whatever level of resolution is desired (e.g., one degree intervals). This can then be used as a longitude multiplier when calculating polygon coordinates relative to the base location.

5 CONCLUSIONS

Google Earth has proven to be an effective tool for geovisualising web traffic, generating visualisations that are highly interactive and visually spectacular. While it has been applied to only a single web site (the Otago EPrints digital repository), the results should be readily generalisable to any web site that keeps adequate traffic logs, as all web sites use essentially the same underlying mechanisms.

Certainly the author has only scratched the surface of what may be possible with this tool. The current visualisations are generated as static KML files, which could raise scalability issues for very large data sets. One solution might be to implement the visualisations as network links that return only data that are visible from the current viewpoint. The current visualisations also display all the available data; it would be useful to limit this to traffic for specific eprints, countries or time periods. Finally, version 2.1 of KML adds support for temporal animation of overlays. This adds a fourth dimension to the three already available and thus opens the door to even more compelling visualisations.

ACKNOWLEDGEMENTS

“Google”, “Google Maps” and “Google Earth” are all trademarks of Google, Inc. Neither the author nor this research are affiliated with Google, Inc. in any way. The map shown in Figure 2 has been placed into the public domain by the CIA (2006).

References

- Beaumont, J. R. (1991). “GIS and market analysis” In D. J. Maguire, M. F. Goodkind & D. W. Rhind (eds), *Geographical Information Systems, Volume 2: Applications*. Longman. Harlow, UK pp. 139–151.
- CIA (2006). “The World Factbook” Central Intelligence Agency, Washington, DC, USA.
*<https://www.cia.gov/cia/publications/factbook/>
- Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V. & Hart, J. C. (1992). “The CAVE: Audio visual experience automatic virtual environment” *Communications of the ACM*. **35**(6): 64–72.
- de Paor, D. G. (2006). “Towards a single, planet-wide, scale-independent, multidimensional geological map” *2006 GSA Annual Meeting and Exposition*. Geological Society of America, Philadelphia, Pennsylvania. (In press).
- Dodge, M. & Kitchin, R. (2001). *Mapping Cyberspace*. Routledge. London, UK.
- Jiang, B. & Ormeling, F. (2000). “Mapping cyberspace: Visualizing, analysing and exploring virtual worlds” *The Cartographic Journal*. **37**(2): 117–122.
- Lamm, S. E., Reed, D. A. & Scullin, W. H. (1996). “Real-time geographic visualization of World Wide Web traffic” *Computer Networks and ISDN Systems*. **28**(7–11): 1457–1468.
- MacEachren, A. M. (1998). “Cartography, GIS and the World Wide Web” *Progress in Human Geography*. **22**(4): 575–585.

- Maxmind (2006). "GeoLite City: Free IP address to city database" Maxmind LLC. Accessed on 22 September 2006.
*<http://www.maxmind.com/app/geolitecity>
- Papadakis, N., Markatos, E. P. & Papathanasiou, A. E. (1998). "Palantir: A visualization tool for the World Wide Web" *Proceedings of the INET'98 Conference*. Geneva, Switzerland.
*http://www.isoc.org/inet98/proceedings/1e/1e_1.htm
- Sale, A. & McGee, C. (2006). "Tasmania Statistics Software" University of Tasmania, Hobart, Australia. Accessed on 23 September 2006.
*<http://eprints.comp.utas.edu.au:81/archive/00000262/>
- Stanger, N. & McGregor, G. (2006). "Hitting the ground running: Building New Zealand's first publicly available institutional repository" *Discussion Paper 2006/07*. Department of Information Science, University of Otago, Dunedin, New Zealand.
*<http://eprints.otago.ac.nz/274/>
- Thompson, K., Keith, J., Swan, R. H. & Hamblin, W. K. (2006). "Linking geoscience visualization tools: Google Earth, oblique aerial panoramas, and illustration and mapping software" *2006 GSA Annual Meeting and Exposition*. Geological Society of America, Philadelphia, Pennsylvania. (In press).
- USGS (2006). "Google Earth applications" Integrated Remote Sensing and Modeling Group, United States Geological Survey Center for Coastal & Watershed Studies, Tampa, Florida, USA. Accessed on 22 September 2006.
*<http://coastal.er.usgs.gov/remote-sensing/advancedmethods/googleearth.html>
- Wood, J., Brodrie, K. & Wright, H. (1996). "Visualization over the World Wide Web and its application to environmental data" In R. Yagel & G. M. Nielson (eds), *Proceedings of IEEE Visualization '96*. IEEE Computer Society and ACM, San Francisco, California pp. 81–86.