Visualising spatially-coded data at the Australian Data Archive

Rhys Hawkins[1], Ben Evans[1], Deborah Mitchell[2], Steve McEachern[2]
[1]ANU Supercomputer Facility, Australian National University, Canberra, Ben.Evans@anu.edu.au
[2]Australian Data Archive, Australian National University, Canberra, Deborah.Mitchell@anu.edu.au

ABSTRACT

There has been an increasing need for spatial data information to be made available through web-based tools; which link seemlessly to data repositories. The Australian Data Archive, ADA, (formally the Australian Social Science Data Archive - ASSDA) is one such example of a critical research data repository with a potential for such tools. In this paper we will present our work on the ADA spatial data framework and describe our new online tools for exploring spatial social science data. This new capability has had implications for the entire data workflow for archiving of survey data. From the design of surveys to incorporate the accurate recording of geospatial identifiers, maintaining confidentiality of geo-located respondents information to prevent identification by unauthorised users and allowing researchers access to the data in new and powerful ways.

INTRODUCTION

The largest repository of Australian research community's social science data is the Australian Data Archive (ADA), including all outputs from ARC funded research. Much of this data has far richer information than has previously been explored by social scientists. To explore one component of this we have developed a visualisaton data tool that extracts the spatially-coded data from ADA and translates it into a form more amenable for online GIS visualization and exploration. Throughout the workflow, we have used open-source tools and open standards such as WMS for web maps delivery.

ARCHITECTURE

A primary driver of the architecture was to enable visualization of spatial data in a time critical manner for online web mapping. The decision was made to post-process survey data with the required geographic markers recorded into a separate data store that was more suited to the task. The data store chosen was PostgreSQL with PostGIS extensions, both open-source products.

One important aspect of the social science data is that the geographic identifiers within survey data poses a risk of breaching confidentiality. We have take steps to minimize the risks that an individual may be identified through the services we provide. This is implemented by using two separate databases for the storage of the necessary data;

1. A "private" database which is accessed only by ADA personnel that contains the geo-located unit-record data, and
2. A "public" database which is accessed by web services that contains only survey responses aggregated to geographic areas, eg mean of response to a particular survey question in each Federal Electorate.

As an additional user protection, we ensure that there are at least 3 individuals aggregated to each geographic region. If there is less than this then that particular region will show no results for that particular geographical region.

The "private" database also includes a library of geographic boundaries that have been accumulated over the project (as of writing this includes approximately 120 different boundaries from 1981 – 2010). In addition the ADA data holdings, we also include some freely available data from the Australian Bureau of Statistics for comparison purposes, eg comparing unemployment rate to responses to survey questions in different areas.

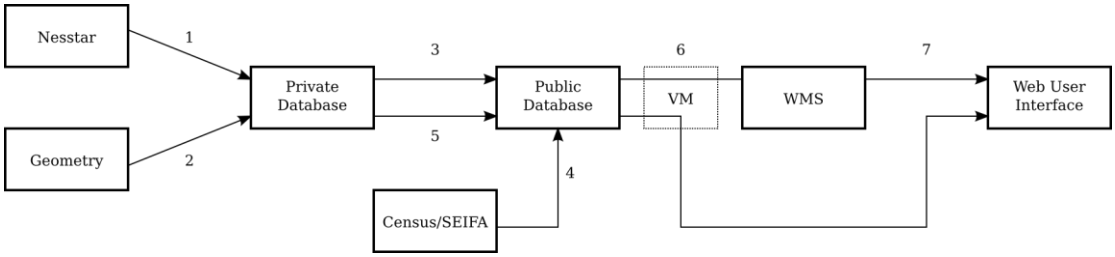

Figure 1:The GIS data framework architecture and workflow

WORKFLOW

The workflow of the ADA GIS data framework, outlined in Figure 1 is as follows:

1. Survey data is extracted from ADAs unit-record data held in Nesstar and stored into the private database,
2. Geographic boundaries are accumulated into the private database geometry library,
3. Unit-record data is aggregated to a given geographic boundary and ingested into the public database,
4. ABS data, such as Census and SEIFA is ingested into the public database,
5. Geographic boundaries are post-processed to be more amenable to online web mapping,
6. The public database is replicate to one or many (as required) virtual machines coupled to a WMS for serving web map image,
7. The Web User Interface provides an exploratory tool for visually exploring the data to enable researchers to identify any interesting spatial characteristics in the data.

ABOUT THE AUTHORS

Rhys Hawkins is a programmer at the ANU Supercomputer Facility, specializing in visualization.
Ben Evans is the Head of the ANU Supercomputer Facility at the Australian National University. He leads projects in HPC and Data-Intensive analysis, working with the partners of NCI and the research sector.
Deborah Mitchell is the Director of the Australian Data Archive.
Steven McEachern is the National manager of the Australian Data Archive.