Unlocking the geospatial potential of survey data

Thomas Ensom[1] and Veerle van den Eynden[1]

[1]UK Data Archive, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ
Tel. +44(0)1206 874973, tensom@essex.ac.uk,
geo.data-archive.ac.uk

Summary: National survey data presents a hugely valuable, but currently underused resource for researchers, learners and teachers using geospatial methodologies. Spatial units are the fundamental method of georeferencing survey data and the integrity of any analysis or visualisation relies on their appropriate usage. Building on the UK Data Archive's expertise in the preservation and dissemination of social science data, we looked in-depth at ways of increasing the value of UK Data Archive services to GIS users. Our developments include a novel web application which assists in the location, interpretation and linkage of spatial unit variables in datasets.

KEYWORDS: spatial units, social sciences, survey data, georeferencing, data services, microdata

1. Introduction

The UK Data Archive, based at the University of Essex, curates and provides access to the largest collection of digital data in the social sciences in the UK. Over 5,000 datasets from government departments, public authorities, research institutes, private companies and academic researchers are made available online to researchers, learners and teachers. Examples of major studies for which the Archive holds data include the Labour Force Survey, British Crime Survey and Understanding Society. GIS and georeferenced data is of ever increasing value to social scientists (Goodchild 2009). To make the most of new methods and tools as they emerge, it is crucial that valuable data resources are presented in a way that enables them to be utilised to their maximum potential. The recent ESRC review of geospatial resource needs (Owen et al. 2009) recognised that despite the growth in availability of geospatial data, the full potential of geospatial data is not being realised by economic and social scientists. The study consulted 512 researchers, finding that two thirds of them consider access to more detailed georeferenced data as the most important data service improvement needed, followed by geospatial linking services.

Many data in the Archive collection have geospatial potential, particularly microdata (that concerning individuals) from large-scale national social surveys and longitudinal surveys. Typically a dataset will be georeferenced using a spatial unit; that is, a set of discrete divisions of a larger space. For example, the electoral wards that subdivide the UK. Problems with disclosure risk however, mean that spatial unit variables have often not been included, or even removed before the dataset is deposited with an archive. There are also an array of potential problems with the quality of the spatial units provided, which if disregarded have the potential to result in flawed analyses, particularly in the hands of the untrained. This is particularly significant given the ready availability of shape files for everything from census boundaries to unit postcodes through services such as UKBORDERS, and more generally the popularity of spatial visualisation. Formal explorations of spatial metadata and data quality challenges have been limited and standardised approaches slow to emerge. This is perhaps in part due to the need for discipline specific strategies.

As part of a JISC-funded project based at the UK Data Archive, we set out to ascertain wherein the problems lie, both by talking to those using the data and looking in-depth at the datasets. We then set out to implement a series of service developments to meet these challenges at the UK Data Archive.

2. Information gathering

Our first challenge was to better understand the needs of users harnessing Archive data in GIS. These fundamental requirements would form an essential foundation to our strategy. We approached this with the idea to build on previous work and focus in on the detail of the methodologies used in handling, preparing and linking survey data in GIS. We carried out interviews with five researchers working on the border between social science and GIS, exploring the problems they encounter and potential improvements that might improve their workflow. Results inform Table 1 below, while full details of this consultation can be found in the published report (Ensom & Van den Eynden 2011).

The second part of our assessment was a major information gathering task, examining the current quality and quantity of geospatial variables in national survey data at the UK Data Archive. This was initiated by querying the UK Data Archive catalogue to identify a set of national survey datasets for the period 1974-2010. Priority was given to time series data from the most popular datasets available through Archive services (known as the 'major studies'). We reviewed 308 datasets, covering a broad sweep of sub-disciplines. We also included datasets held under varying access methods, including special end user license agreements and the Secure Data Service. We proceeded to evaluate and validate the georeferencing variables within individual data files to be able to identify spatial unit definitions used. In consultation with EDINA we then verified the variable quality and ascertained to which readily available boundary and point geography resources the unit could be mapped and to what degree of accuracy. With minimal metadata on which to build, doing this retroactively was at times challenging, a catalogue record typically including only a non-standardised list of those spatial units reported to have been included. This exercise feeds into Table 1 (below), and is also now available through the U·Geo Browser interface (see section 3).

Table 1. Data service requirements for spatial social scientists, structure by 'problems' and an associated 'requirement' from the data service.

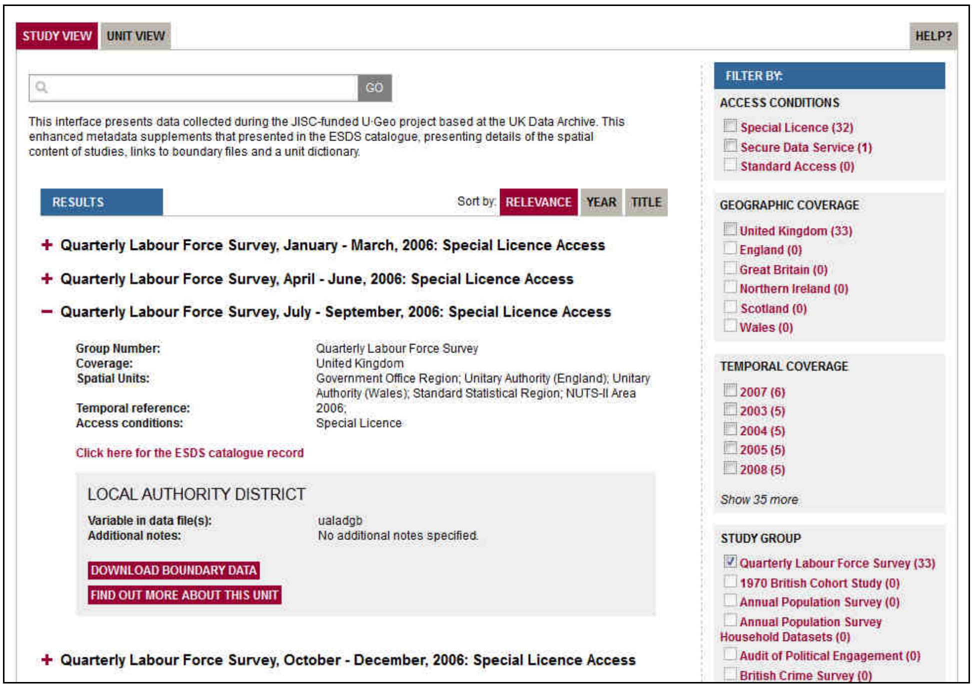
Problem	Requirement
Inconsistent variable coding and labelling. Laborious manual matching of datasets.	Coding variables to conform with standardised definitions e.g. Office for National Statistics administered GSS Coding and Naming scheme. Also consider clear variable naming.
Not clear which definition of a spatial unit is being used. No time referencing.	Provide standardised unit definitions with timestamp metadata.
Differing requirements among users; may need access to different kinds of unrelated spatial unit.	Where possible provide selection of spatial units to suit different users and/or flexibility allowing for these to be derived with lookup tables. E.g. grid reference and postcode.
Not enough spatial units available, and not at a fine enough level of detail.	Encourage deposit of high quality spatial variables with archives. Encourage use of disclosure solutions such as special license agreements and the Secure Data Service. Supply more studies with 'output area' unit.
Finding metadata/documentation for a spatial unit often challenging; linking to boundary data even more so.	Present enhanced metadata for major studies through an appropriate interface. Particularly clarification of changes/inconsistency across longitudinal data. Also point users to appropriate boundary files.

It is worth mentioning that there is encouraging evidence of a continuing upward trend in the quantity, resolution and quality of geospatial variables in data deposits. However, while it is becoming an increasingly common choice, the uptake of the Output Area spatial unit has been limited. We feel this is a shame given that it provides a non-disclosive, statistically sound unit for fine grain analysis, and would encourage its further adoption.

3. Service development

While standards such as the EU mandated INSPIRE directive provide a framework (EC 2008) for making outward facing metadata more useful to geospatial practitioners, current standards are not able to capture sufficient detail on spatial units to allow for unambiguous capture of issues pertinent to the social sciences. INSPIRE compliant metadata is soon to be implemented at the UK Data Archive, forming part of resource discovery developments that should much improve quality of life for spatial social scientists. However, we wanted to look outside the scope of INSPIRE and showcase the value of augmented metadata based on the additional spatial unit quality information captured during the dataset review (see section 2). In addition to helping experts simplify potentially time consuming 'data excavation' tasks, the site could act as an enabler for the less experienced, providing essential information for valid data integration.

The application we have developed, called the U·Geo Browser, will make spatial units more accessible, transparent and usable than ever before; presenting extensive enhanced metadata including standardised and time-referenced definitions for spatial units. It will also help users easily identify which survey data can be linked to digital boundary data available through the EDINA service UKBORDERS. The interface is built on open-source Solr technology, and allows for intuitive, on-the-fly filtering of search results by relevant criteria.



This interface presents data collected during the JISC-funded U·Geo project based at the UK Data Archive. This enhanced metadata supplements that presented in the ESDS catalogue, presenting details of the spatial content of studies, links to boundary files and a unit dictionary.

+ Quarterly Labour Force Survey, January - March, 2006: Special Licence Access

+ Quarterly Labour Force Survey, April - June, 2006: Special Licence Access

− Quarterly Labour Force Survey, July - September, 2006: Special Licence Access

Group Number:	Quarterly Labour Force Survey
Coverage:	United Kingdom
Spatial Units:	Government Office Region; Unitary Authority (England); Unitary Authority (Wales); Standard Statistical Region; NUTS-II Area
Temporal reference:	2006;
Access conditions:	Special Licence

Click here for the ESDS catalogue record

LOCAL AUTHORITY DISTRICT

Variable in data file(s):	ualadgb
Additional notes:	No additional notes specified.

+ Quarterly Labour Force Survey, October - December, 2006: Special Licence Access

ACCESS CONDITIONS
□ Special Licence (32)
□ Secure Data Service (1)
□ Standard Access (0)

GEOGRAPHIC COVERAGE
□ United Kingdom (33)
□ England (0)
□ Great Britain (0)
□ Northern Ireland (0)
□ Scotland (0)
□ Wales (0)

TEMPORAL COVERAGE
□ 2007 (6)
□ 2003 (5)
□ 2004 (5)
□ 2005 (5)
□ 2008 (5)

STUDY GROUP
☑ Quarterly Labour Force Survey (33)
□ 1970 British Cohort Study (0)
□ Annual Population Survey (0)
□ Annual Population Survey Household Datasets (0)
□ Audit of Political Engagement (0)
□ British Crime Survey (0)

Figure 1. A screen-capture of the U·Geo Browser (beta version) front page. Facets (which function like filters) appear in the right hand column. A list of Quarterly Labour Force Survey results are visible in the left section, with filters along the right hand side which allow for further refinement of the results. Links take the user to download pages for the dataset and boundary files, and to a generic definition of the spatial unit (in the example this is Local Authority District). The boundary data download option is only available for data with a well-defined spatial unit.

5. Conclusions

Social science data archives can contain an incredible wealth of material of potential interest to GIS users in a broad array of vocations and disciplines. The UK Data Archive has undertaken a first foray into making this data more accessible to those users. As geo-referenced data is increasingly made available through data services, it should be a primary concern that both those generating and preserving any kind of geo-referenced survey data take into account the issues we have raised.

A secondary take away message is that innovative methods can provide powerful tools in addressing data challenges. We present one here, building a powerful application on technology perhaps more familiar as part of a shopping experience than a research process. While much innovation in GIS data services focuses on visualisation and mapping, we would encourage people to go back to basics and ensure that consideration is given to fully realised geospatial metadata.

6. Acknowledgements

This project was funded by the JISC Geospatial programme.

Many thanks to EDINA for their contributions throughout the project as technical consultants, particularly James Crone. Thanks also to all those at the UK Data Archive who offered contributions.

7. References

Ensom, T. & Van den Eynden, V. (2011). *Report on user needs: unlocking the geospatial potential of survey data.* Colchester, UK Data Archive, University of Essex.

European Comission (2008). Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (Text with EEA relevance). *Official Journal of the European Union* L 326, p. 12–30

Goodchild, M. F. (2009). Social Sciences: Interest in GIS Grows. Chapter in *GIS Best Practises: Social Sciences*. ESRI, New York.

Owen, D., Green, A. & Elias, P. (2009). *Review of geospatial resource needs.* ESRC, Swindon.

8. Biography

Thomas Ensom is a biologist by training, having particular interests in biodiversity informatics and taxonomy. Since joining the UK Data Archive for the Rural Economy and Land Use programmes Data Support Service, he has explored in depth the challenges of managing and preserving research data, particularly geospatial issues and user interface design.

Veerle Van den Eynden manages the Research Data Management Support Services at the UK Data Archive, supporting researchers with good data practices and optimising data sharing opportunities. She previously studied people-plant and people-environment interactions, using GIS in natural resource management research.