U·GEO: REPORT ON USER NEEDS

UNLOCKING THE GEOSPATIAL POTENTIAL OF SURVEY DATA

PUBLIC
09 November 2011
Version: 01.00

THOMAS ENSOM
VEERLE VAN DEN EYNDEN

T +44 (0)1206 872001
E info@data-archive.ac.uk
www.data-archive.ac.uk

UK DATA ARCHIVE
UNIVERSITY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX, CO4 3SQ

Work funded by the Geospatial strand of the **JISC Infrastructure for Education and Research Programme**

1. Introduction

This report summarises information gathered to assess the needs of UK Data Archive users in better utilising the geospatial element of our data holdings. The information compiled here is taken from multiple sources and is a combination of original findings and synthesis of existing materials. At the core of the consultation were a series of discussions with users of social science survey data available through core UK Data Archive services: the Economic and Social Data Service (ESDS) and the Secure Data Service (SDS). We focused in on how they work with geospatial data and what they require from data providers in order to do so most effectively. While some are specific to the Archive, most of our findings are broadly applicable to the handling of geo-referenced survey data. The definitions of terms given in the table below are a reference for this document, and are not necessarily comprehensive.

Table 1. Definition of terms.

Term	Definition
Economic and Social Data Service (ESDS)	A data service delivered by the UK Data Archive, providing access to research data from the social, political and economic sciences.
Government Office Region (GOR)	A **spatial unit** providing a broad level statistical and (until its recent abolition) administrative regions for England. Additional 'pseudo-regions' for Wales, Scotland and Northern Ireland are often included.
Government Statistical Service (GSS) Coding and Naming	This scheme provides a basic standard for referencing spatial units with codes. The most commonly used UK spatial units are covered.
National Statistics Postcode Directory (NSPD)	A series of lookup table for UK postcodes, allowing for various spatial units to be extrapolated from a single postcode. They are time-referenced.
Output Area / Super Output Area	A **spatial unit** created to provide a stable low level geography for statistical analysis. Super Output Area aggregates Output Areas for coarser grained analysis.
Secure Data Service (SDS)	A UK Data Archive delivered data service providing secure access to sensitive and/or disclosive data.
Spatial Unit	A spatial unit provides a reference for the location of a particular entity – be it to within a bounded space or a specific point location. Examples: county; parliamentary constituency
Special License	An end user license with conditions preventing disclosure of sensitive data. Can be seen as a less strict alternative to the **Secure Data Service**.

Previous consultations form useful background to this report, most significantly an ESRC review of geospatial data in the social sciences[1], and an associated workshop series. Complementing this is a report published on geospatial support services in higher education institutions[2]. In addition, Census.ac.uk user interaction at workshops and other events provides some further indication of geospatial 'problem areas', which have particular significance with regards users with lower levels of expertise. Our interviewees were selected based on their involvement with spatial research using social science survey data. The discussions were semi-structured (undertaken in person, over telephone and via email) allowing us to engage users with targeted questioning while allowing for some flow in the conversation. The typical workflow for a data user is divided into sections, and it was these that we structured our interviews around and are the headings used below. Processes and problems discussed were common to many users and the points detailed below reflect the generality of the message. Each section ends with a set of recommendations for the Archive based on the comments of users.

2. Data selection and availability

The datasets used vary considerably with research specialisation. For example, economic datasets are more likely to contain Travel to Work Areas, while political datasets will often contain a constituency unit. This is a product of varied research practise and conventions associated with particular disciplines. It should be noted

that these conventions seem to generally be dictated by data producers rather than data re-users. Despite differences however, there is a common reliance on Census geographies across sub-disciplines.

The availability of low level geographies (such as postcode or grid reference) via special licence and the newly launched SDS is extremely important to researchers who work with spatial data. Output Area (and the related Super Output Area) is becoming an increasingly common choice of unit to attach to data, for data providers and users, as it provides a unit ideally suited (and indeed, tailored) to statistical analysis: minimal boundary changes and a consistent unit size. Despite recent innovations such as the SDS, users expressed frustration at licensing and disclosure prevention measures which mean they cannot carry out the research they want to. The main issues are the provision of only very coarse resolution spatial units, and the re-coding/scrambling (a semi-anonymisation technique) of low level geographic variables.

A sometimes severely limiting factor in spatial analysis is sample size, as sufficient numbers of individuals are not always available to permit sound statistical analysis. Typically this factor is prohibitive to the extent that local authority is the smallest recommended variable for analysis. Some users (particularly those involved in further education) were keen that guidance might be provided for those new to spatial analysis on the suitability of data based on sample size.

Key recommendations for the Archive are:

- **Where possible provide a selection of spatial units to suit different users – for example, political research may require Parliamentary Constituency while a teacher may be happy with Government Office Region (GOR). When possible, grid reference and postcode should be provided as these allow the user to move between different units using lookup tables.**
- **Provide encouragement and support to those who are not submitting high quality spatial variables, or are holding back low level geographies for fear of disclosure risk. This should include additions to the deposit forms and procedure, and the offer of solutions such as Special License and SDS access.**
- **Pursue the relaxation of licensing rules on spatial data. The obvious disclosure prevention measures have been mentioned, but also copyright issues attached to boundary data. Postcode and grid reference are always preferred units – lookups such as the National Statistics Postcode Directory (NSPD) can then be used to derive other units.**

3. Locating and interpreting spatial units

There were many comments relating to the consistency of variables within time series and between collections. This problem can be separated into two major issues: that the spatial unit name can be recorded incorrectly, incompletely or ambiguously; and that poorly documented variables lack metadata indicating the time period the spatial unit was defined. For example, Government Office Region has been redefined three times, in 1996, 1998 and 1999. It is crucial to know which version of this unit has been referenced, as the boundaries may vary between the versions. With other spatial units changes may be more regular and larger in scale. These changes have the potential to negatively impact the integrity of analyses using the data in question. There was near-unanimous agreement across sources that this problem is an area of particular priority.

Key recommendations for the Archive are:

- **Provide time referenced unit definitions (the single most emphasised issue by users).**
- **Take measures to improve the detail of deposited documentation materials on spatial variables, through additions to the deposit forms and guidelines.**
- **Add to cataloguing procedures and input programs to ensure catalogue spatial metadata is compiled to as high a standard as possible.**
- **Variable names used across time series may not always be consistent, a fact which must be reflected in documentation and guidance.**
- **Providing information on the frequency of changes and the authority that manages them would be beneficial.**

4. Preparation and linkage of geospatial variables

Users showed concern over the coding (or value labelling) of variables. Often arbitrary, non-standard codes are used to notate the geospatial variable, which forms a major obstacle when preparing data for analysis, particularly with heavily subdivided geographies, such as ward. A lack of appropriate coding leaves users with the arduous process of matching the free text entries to a coding scheme themselves. For example, a list of wards used as the basis of a survey geocoding may have been sorted into alphabetical order and numbered from 1 to 9,434. In the GIS boundary files a coding scheme may have been used to give each Ward a 9 digit code.

One user raised the significance of alleviating the sample size problem by combining data sets, thus boosting the population sample. There are groups of common questions shared among surveys, but these require data that is linkable on comparable spatial units, at the desired resolution. Harmonisation efforts should be careful to consider these issues.

EDINA services such as UKBORDERS and Unlock seem to be widely used by researchers, suggesting these could be appropriate resources to link to from our data. There is some demand for guidance on appropriate resources, as there is a paucity of this information in the social sciences.

Key recommendations for the Archive are:

- **Complete coding of variables to conform with definitions using some kind of standardised system. We would recommend the ONS administered GSS Coding and Naming scheme.**
- **Consistency with look up file codes, and information on appropriate lookup files, particularly when postcode or grid reference is provided.**
- **Provide links to useful resources for social science researchers, preferably in a way that links specific data to directly relevant tools.**
- **Further consideration of the comparability of spatial units across survey datasets as harmonisation efforts continue.**

5. Conclusions

Our consultation very much supported views established by the team during the U·Geo project. This included auditing the geospatial content of survey datasets and reviewing extent of metadata provided. The Archive will systematically tackle the issues raised through the activities of the U·Geo project with the following approaches:

- **Developing a comprehensive controlled vocabulary for spatial units to sit under our catalogue metadata.**
- **Improving Archive metadata through the implementation of the EU INSPIRE directive, augmented with additional documentation of relevance to social science data. There will be a particular priority placed on provision for temporal referencing of spatial units.**
- **Implementing a specialised resource discovery tool presenting enhanced metadata for major studies in a search/browse interface and links to appropriate boundary files. This will provide an intuitive way of exploring our geospatial data resources with relevance to both the beginner and advanced GIS user. This will form a pilot and demonstration of technologies which will be integrated with full ESDS catalogue at a later date.**
- **Improving the guidance given to data depositors; ensuring they provide data and metadata of the highest possible quality.**
- **Acquiring more Output Area, postcode and grid reference variables for major studies with access through Special License and SDS.**

Geospatial data provision in the social sciences has been steadily improving, a trend highlighted not only by work preceding ours but also through looking at the datasets themselves. However expertise is fragmented across and within institutions – so collaboration is a must. There seem to be two extremes in the GIS–social science user base. There are the expert users who have become very used to working with messy data and developing their own ways of dealing with the associated problems. As detailed above, these users share many of the associated frustrations of these methods. At the other end of the spectrum are the beginners; those wishing to start using GIS in their work, whose use of geospatial methods is being limited or even

prevented by the complexity of the processes required. It is crucial that both types of user are considered when developing data services; and indeed, should logically move forward together. Among some users there is a desire for more guidance in preparing and linking data for spatial analysis. With expertise scattered, data service providers must play a part in this. A long term recommendation of this report is the development of Archive resources to support these activities. The first crucial step however is providing a solid foundation of good geospatial data and metadata.

A1. Acknowledgements

Thanks to the researchers who generously offered us their time to discuss their experiences.

The combined work of the various geospatial consultations preceding this one has formed a valuable ground for this report, and we are grateful for the work that went into producing these.

A2. Referenced publications

1. Owen, D., Green, A. & Elias, P. (2009) Review of geospatial resource needs.
2. Sutton E., Chisholm, H. & Armitage, T. (2010) Survey of Support for Geospatial Resources within Higher and Further Education