An illustrative guide: Using GEODE to link data from SOC-2000 to NS-SEC and other occupation-based social classifications

Paul S. Lambert University of Stirling

11th June 2007 [Edition 1.1]

GEODE Project Technical Paper No. 2

Technical Papers of the GEODE project: Grid Enabled Occupational Data Environment, www.geode.stir.ac.uk/publications.html . ESRC Small Grant in eSocial Science, Ref: RES-149-25-1015. The GEODE project is affiliated with National Centre for eSocial Science (www.ncess.ac.uk) and National eScience Centre (www.nesc.ac.uk).

Contents:

1. INTRODUCTION

1.1 Background

In the GEODE project we have attempted to provide an online data index service which stores and supplies occupational information resources for the benefit of social scientists who work with occupational data. The service is accessed by logging into the 'GEODE portal', either with personalised details or as a guest.

GEODE stands for "Grid Enabled Occupational Data Environment". The GEODE project involves exploiting the computing technologies associated with the 'Grid', (also known as 'e-Science' and 'e-Social Science'). A full text introduction to the GEODE project is given in Lambert et al (2006). Tan et al (2006) adds further technical details to the description of the service. The project web-pages www.geode.stir.ac.uk also contain information on the GEODE project and its participants.

An extended introduction to the GEODE project's online services is given in Lambert and Tan (2007). Sections from that text are also reproduced at times within the text below.

1.2 Practical illustration

Exploiting the GEODE services does at present require some effort. There are a number of web-pages and documents available via the website – www.geode.stir.ac.uk – which offer instructions on using the GEODE portal, but all of these require some commitment from the user in following their instructions.

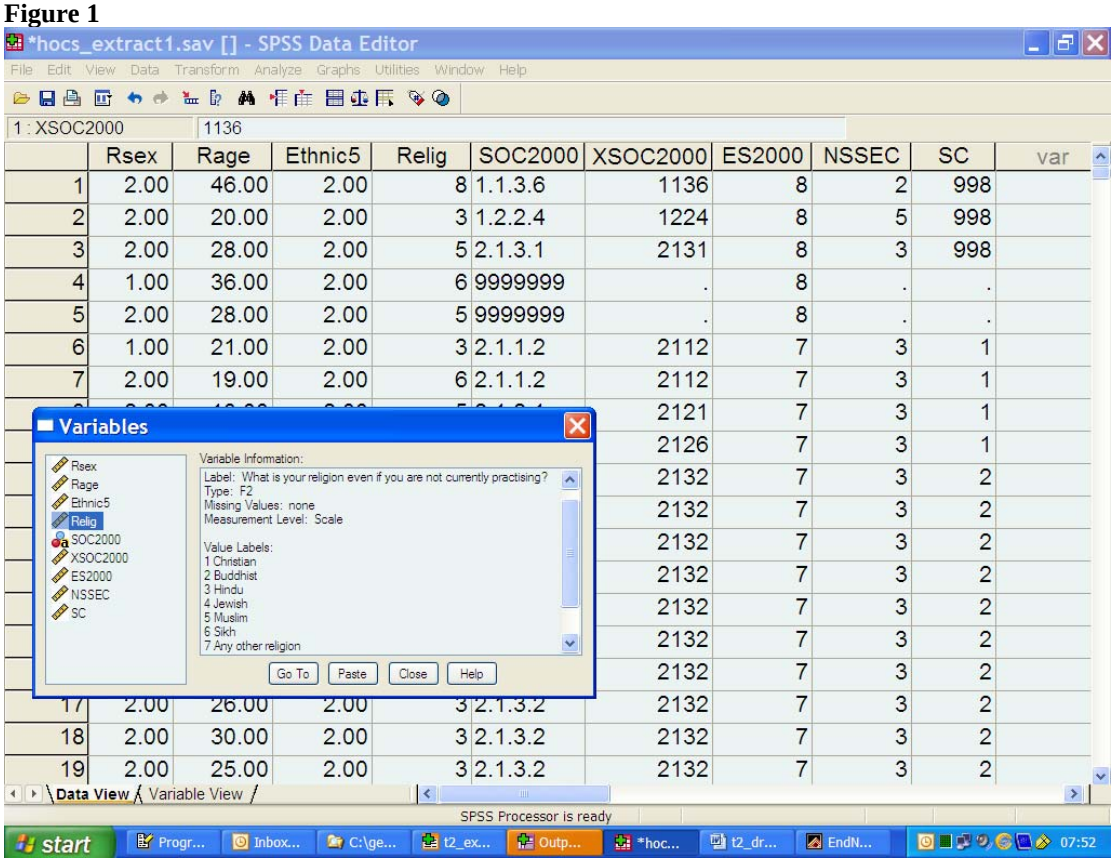
Our experience suggests that step-by-step examples are the easiest way for users to understand the facilities available at GEODE. With this in mind, this short document has been prepared in order to illustrate one of the most common examples of usages of the GEODE occupational information service. This is the scenario of a user with a social survey dataset which contains occupational information in a nationally standardised occupational unit group scheme, who wishes to produce occupation-based social classifications on their dataset.

In our illustration we will take the example of a micro-data extract from the Home Office Citizenship Survey (HOCS) of 2005 (Home Office, 2006). Figure 1 shows a screen shot of this data stored in the SPSS package[1].

If you wish to exactly replicate this data example, the full HOCS micro-data is free to download for academic researchers from the UK Data Archive http://www.data-archive.ac.uk/ (study number 5367). This example runs on a small extract from the full survey; the SPSS commands used to construct the extract file used here are reproduced in Appendix Table A1 below.

[1] This example is conducted using SPSS, though it generalise to the use of other similar packages. The GEODE website also includes selected notes on using Stata for similar operations.

Figure 1



In this example, we have taken a small extract from the HOCS sample to keep this illustration a little simpler. Our extract features micro-data on 9 variables (some basic demographic data and some occupational records) for 805 cases (for information, our subsample comprises those respondents from the 2005 HOCS whose ethnic group is 'Asian', who are aged between 18 and 50, and who contributed data on their employment status, variable 'es2000').

The user's requirement is to 'get some social class measures' onto this micro-data. In our experiences in the GEODE project, this has been the most common requirement of GEODE users.

1.3 Principles of using occupational micro-data

We will start this illustration by making some assertions. In social research reports, it is not uncommon to see instances where a social classification measure is assigned to occupational micro-data through an *ad hoc* classification method – such as a case-by-case allocation into 'social class' categories on the basis of the analysts own judgement. This is a very unsatisfactory approach – it prevents replication and comparability, and ignores the many suitable resources which could (and should) have been used for such data. This point has been made in many previous publications (for UK examples, see Armstrong, 1972; Bechhofer, 1969; Lambert, 2002; Lambert et al., forthcoming 2007; Marsh, 1986).

A better model features two stages of work in assigning occupation-based social classifications. Researchers should:

1) Preserve ‘source occupational data’ by using standardised occupational index schemes

- By ‘source occupational data’ is meant the original record describing the occupational position (such as was collected by the questionnaire interview). This data should be recorded in a detailed occupational unit group scheme (such as SOC-2000) on the relevant micro-data. Ideally it should also use additional variables to describe other relevant features of the occupational position, such as measures of employment status. For instance, in the UK the Office for National Statistics publishes guidelines on collecting and preserving source occupational data at:
  http://www.statistics.gov.uk/methods_quality/ns_sec/questions.asp

2) Use published resources to code source occupational data to an occupation-based social classification scheme on the basis of transparent algorithms

- That is, exploit published data which indicates what value certain occupational unit groups should be assigned within certain occupation-based social classifications (such as social classes, or scores on a stratification scale). The fact that such data is published and available to other researchers is crucial, since it guarantees the transparency and replicability of the classification used.

1.4 Reviewing the HOCS data example

As is standard with secondary social surveys, the HOCS data has been released with some ‘source occupational data’ for each case. The extract used in this example (Figure 1) features source occupational data in the ‘SOC2000’, ‘XSOC2000’ and ‘ES2000’ variables. All of these variables concern information about the current or last job held by each survey respondent (there are also some other source occupational variables available in the HOCS, but we have omitted them for clarity). The first two measures provide records of occupational position in the 4-digit SOC-2000 occupational unit group codes (ONS, 2000). In fact the two variables ‘SOC2000’ and ‘XSOC2000’ actually contain the same data, but in two different formats. The third variable is a measure of ‘employment status’, which has been recorded in a standardised scheme for employment status categories in the UK. In our extract the distribution of cases to the employment status variable is shown in Figure 2.

Figure 2

	Employment Status ES2000	
	Male	Female
	Cases	Cases
1 Self-employed : large establishment (25+ employees)	1	3
2 Self-employed : small establishment (1-24 employees)	14	33
3 Self-employed : no employees	23	71
4 Manager : large establishment (25+ employees)	10	26
5 Manager : small establishment (1-24 employees)	4	17
6 Foreman or supervisor	38	105
7 Employee (not elsewhere classified)	133	322
8 No employment status info given - for use in this program on	1	4
Total	224	581

Our task in this example will be to use these source occupational variables (SOC-2000 and Employment Status) in order to link the HOCS micro-data with appropriate occupation-based social classifications.

As well as the source occupational data, it may be apparent from Figure 1 that the extract already has some social classification data on it. In variable 'NSSEC' the extract features the occupation-based social classifications of the operational (also known as the 'full' or 'long') version of the NS-SEC scheme (e.g. Rose & Pevalin, 2003). In variable 'SC' the extract features a version of the Registrar General's Social Class scheme (e.g. Reid, 1998, Appendix B). These measures would have been derived during the processing of the HOCS data prior to its release as a secondary dataset, on the basis of Office for National Statistics algorithms for each scheme (ONS, 2002; OPCS, 1991). Their distributions are shown in Figures 3 and 4.

Figure 3

	NS-SEC - long version																													
	1	2	3.1	3.2	3.3	3.4	4.1	4.2	4.3	5	6	7.1	7.2	7.3	7.4	8	9	10	11.1	11.2	12.1	12.2	12.3	12.4	12.5	12.6	12.7	13.1	13.2	13.3	13.4
Male	1	2	10	5	3		17		2	12	4	13	3	2	2	12	19	15	5	3	13	12	5	10		4	2	1	14	11	12
Female	2	19	44	28	11	1	40	3	7	26	13	14	11	3	3	26	60	29	8	8	25	45	17	32	1	4	2	7	32	35	16
Total	3	21	54	33	14	1	57	3	9	38	17	27	14	5	5	38	79	44	13	11	38	57	22	42	1	8	4	8	46	46	28

Figure 4

	SC - Social Class (old scheme)					
	1	2	3.1	3.2	4	5
1.00 Male	13	54	41	46	53	7
2.00 Female	55	170	67	140	123	14
Total	68	224	108	186	176	21

We don't show the value labels for these social class schemes to save space, but they are available online:

- NS-SEC www.statistics.gov.uk/methods_quality/ns_sec/analytic_operation_cat_subcat.asp
- SC: www.statistics.gov.uk/methods_quality/ns_sec/continuity.asp *(in roman numerals)*

We have said our aim in this analysis is to 'get some social class data'. Given this aim, it might not be unreasonable to ask why we should not stop here, and use these NS-SEC and SC variables? In many projects this really is as far as a researcher need go in this field – if there is a well-documented classification available for the relevant data resource, the most efficient thing to do is to use it. However, there are three reasons for reading on.

- Firstly, it is useful to know what to do with survey micro-data when such measures (social classifications) are not available in a pre-prepared format. In particular, any type of comparative analysis between different datasets may well require manipulation of derived variables on at least some of the micro-data (since it is unlikely that the same derived variables will be available for every dataset). This paper is a brief methodological illustration on the HOCS, which is designed to be applicable to the many other datasets which do not initially include derived variables.

- Secondly, there are significant pragmatic limitations to the variables summarised above (NS-SEC and SC), which restrict their value as analytical measures, and provoke many researchers to use variant versions of them. For instance, both schemes have a nominal level of measurement; both have many different categories, some of which are sparse; and both have strongly gendered distributions, meaning that certain categories tend to be sparse either for male of female populations only. Thus, conventional statistical approaches tend to encourage the collapsing together of some of the social class categories of either scheme, and/or consideration of using their categories in an ordinal or metric framework (for instance, variants of the NS-SEC scheme are often analysed by assigning metric scores to each category, e.g. Hendrickx & Ganzeboom, 1998). In either case, it is less desirable to undertake such operations without any previous background information; a more scientific approach is to exploit categorisations and derivations which have also been published in other outputs (one example being the instructions on recoding which can be found on the web-pages cited above for the NS-SEC).

- Lastly, it is valuable to consider the numerous alternative occupation-based social classifications available to social scientists, rather than relying only on those already available. Indeed, though there has been published support for both the NS-SEC and SC schemes (e.g. Reid, 1998; Rose & Pevalin, 2003), there have also been many theoretically and empirically based critiques of both classifications(e.g. Prandy, 1998, 2002; Weeden & Grusky, 2005). If occupational circumstances are to be explored in any depth, it is informative to review a variety of alternative measures. Indeed, it is often argued that an unfortunate feature of previous survey data analysis projects has been that many researchers have ignored more plausible alternative occupation-based measures, in favour of the measures which are immediately available on existing datasets (e.g. Lambert, 2002). A particular aspect of this concerns problems of categorisation – the occupation-based social classifications most likely to be found on existing secondary surveys are categorical in nature, yet there is much evidence to suggest that alternative metric scales of occupational circumstances are more robust analytical tools and more amenable to a wider variety of appropriate data analytical strategies than are class categorisations.

1.5 Introducing the GEODE portal

In this illustration we will show examples of using the GEODE 'portal' to access and exploit occupational information resources. The portal is a specialist web-page service which involves logging in with a username and password. Many users of GEODE will have been supplied with a personalised username and password, which may be obtained on email request to the GEODE project authors (see Lambert & Tan, 2007). Alternatively, it is also possible for any user to login to the GEODE portal as a 'guest', with the details:

Username: guest
Password: geode

Login access to the GEODE portal is linked from: www.geode.stir.ac.uk

Guest level access to the portal allows users to search for occupational information and to use the GEODE file matching facility. Personalised login accounts support the same facilities, plus in addition they allow users to deposit their own occupational information resources at the GEODE server.

The layout of the GEODE portal is slightly different between a 'guest' and a personalised login. Therefore in the following examples we often show two images of the portal according to the type of login access – in the example below, Figure 5 shows the front page seen after a guest level login, and Figure 6 shows the front page visible after a personalised login (note that the layout of tabs, also known as 'portlets', is different in these two logins).

Figure 5

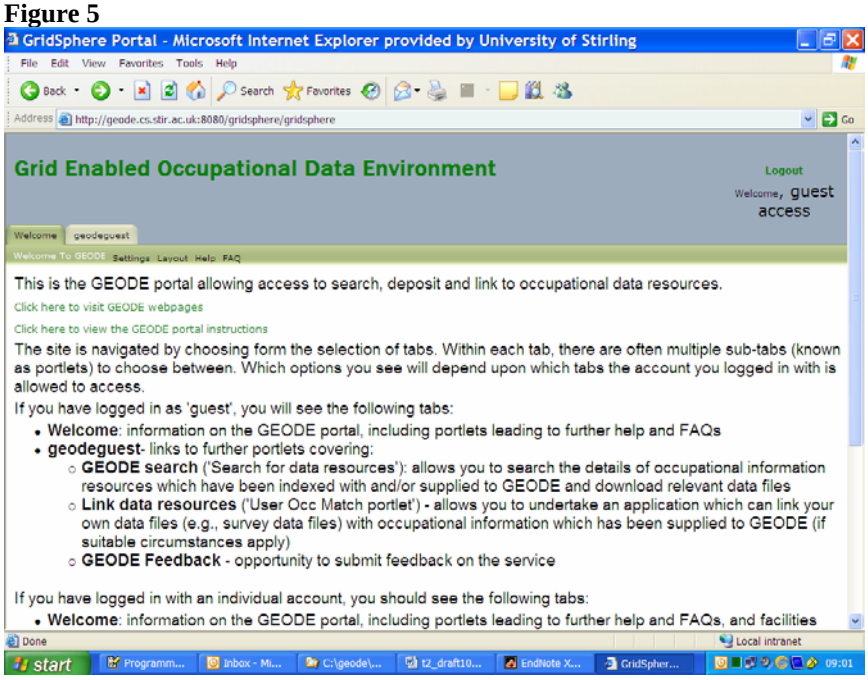
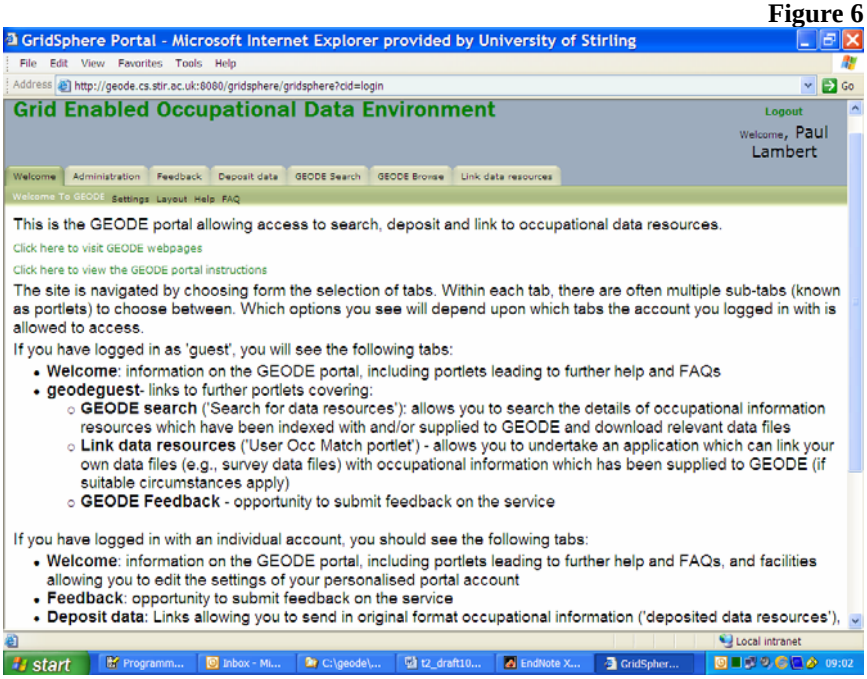


Figure 6


GridSphere Portal - Microsoft Internet Explorer provided by University of Stirling
File Edit View Favorites Tools Help
Address http://geode.cs.stir.ac.uk:8080/gridsphere/gridsphere?cid=login
Go
Grid Enabled Occupational Data Environment
Logout
Welcome, Paul Lambert
Welcome
Administration
Feedback
Deposit data
GEODE Search
GEODE Browse
Link data resources
Welcome To GEODE
Settings
Layout
Help
FAQ
This is the GEODE portal allowing access to search, deposit and link to occupational data resources.
Click here to visit GEODE webpages
Click here to view the GEODE portal instructions
The site is navigated by choosing form the selection of tabs. Within each tab, there are often multiple sub-tabs (known as portlets) to choose between. Which options you see will depend upon which tabs the account you logged in with is allowed to access.
If you have logged in as 'guest', you will see the following tabs:
Welcome: information on the GEODE portal, including portlets leading to further help and FAQs
geodeguest- links to further portlets covering:
GEODE search ('Search for data resources'): allows you to search the details of occupational information resources which have been indexed with and/or supplied to GEODE and download relevant data files
Link data resources ('User Occ Match portlet') - allows you to undertake an application which can link your own data files (e.g., survey data files) with occupational information which has been supplied to GEODE (if suitable circumstances apply)
GEODE Feedback - opportunity to submit feedback on the service
If you have logged in with an individual account, you should see the following tabs:
Welcome: information on the GEODE portal, including portlets leading to further help and FAQs, and facilities allowing you to edit the settings of your personalised portal account
Feedback: opportunity to submit feedback on the service
Deposit data: Links allowing you to send in original format occupational information ('deposited data resources'),
Local intranet
start
Programm...
Inbox - Mi...
C:\geode\...
t2_draft10...
EndNote X...
GridSpher...
09:02

2. USING GEODE: SEARCHING FOR RESOURCES

To reiterate, the scenario here is that we have access to micro-data from the Home Office Citizenship Survey extract, which features some source occupational data in terms of SOC-2000 unit groups and employment status details, and we wish to link this data with some appropriate social classifications. There are two ways in which we could use the GEODE portal to achieve this (also see Lambert & Tan, 2007). The first involves exploiting GEODE as a searchable database of occupational information resources (covered in this section). The second usage involves a facility provided through GEODE which allows us to link together micro-data with other occupational information files (sections 3 and 4 below).

On logging into the GEODE portal, there are two tabs ('portlets') which allow us to search for suitable resources. These are the 'browse' and 'search' portlets. If logged in as a guest, these two resources are available from links under a portlet called 'geodeguest' (see Figures 7 and 8 for the search and browse links respectively). If logged in with a personalised user account, these are visible as distinctive tabs (see Figures 9 and 10 respectively).

Figure 7

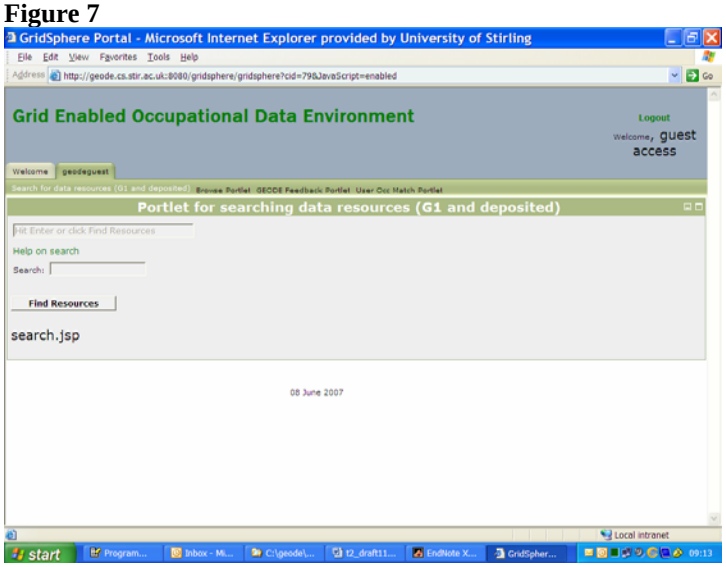


Figure 8

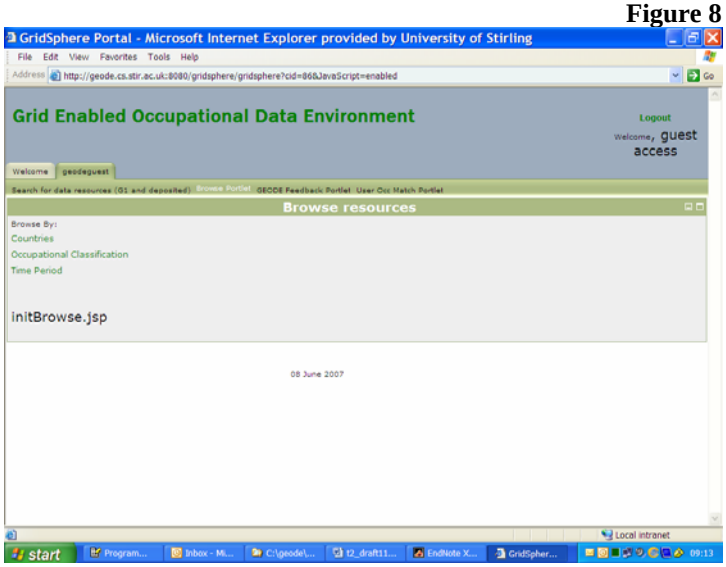


Figure 9

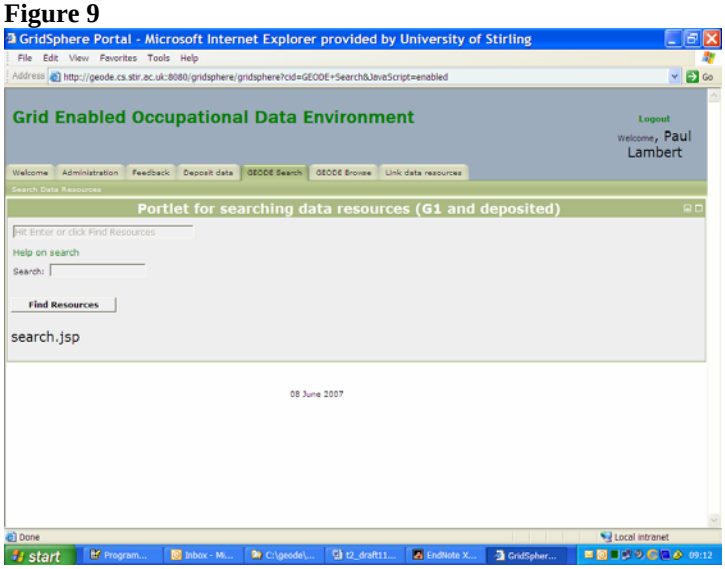


Figure 10

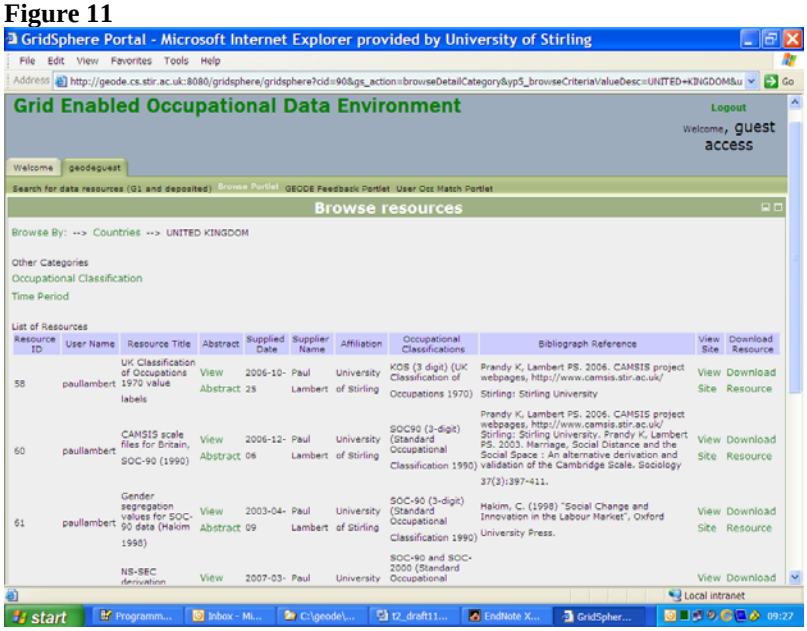
The question now arises of what we should search or browse for? In this example, we want to browse or search for occupational information resources which are related to the country, occupational classification or time period covered by the HOCS.

It is important to appreciate that within GEODE, there are two types of occupational information resource which may be detected by a user. These are known as 'uncurated' and 'curated' resources. They are described more fully in section 2.3 below, as well as in Lambert and Tan (2007). We are usually more interested in 'uncurated' resources.'Curated' resources are best thought of as a subset of 'uncurated' resources, with extra information added to them.

2.1 Browsing

We will start with the 'browse' option, and what might be the more obvious option, of browsing by countries (Figure 11).

Figure 11

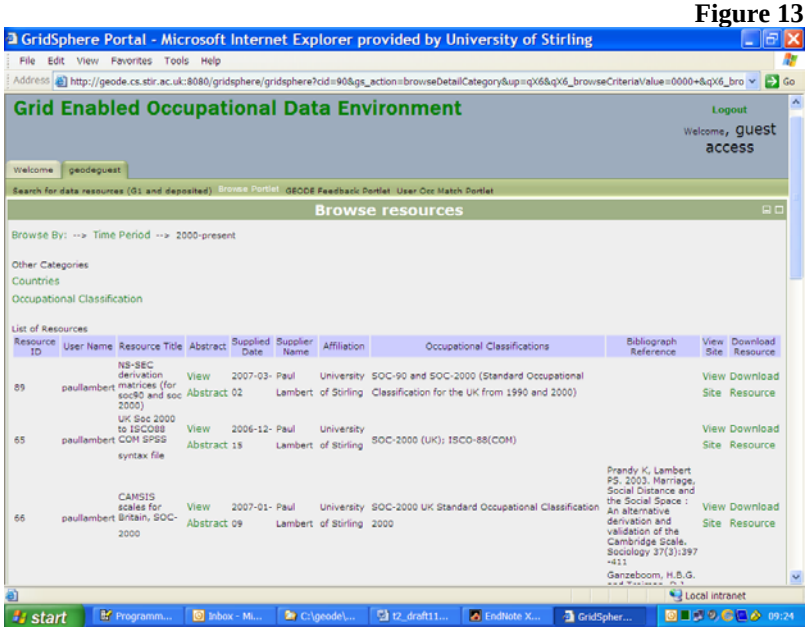


If we click on the 'countries' link of the browse menu, and then follow the link for United Kingdom, we see a number of hits (occupational information resources registered at GEODE which are relevant to the UK). In fact, some of these hits are relevant to SOC-2000 data (for instance the last visible resource (number 81) in Figure 11. We will shortly expand on dealing with results from browsing and searching for files.

On the browse option, we could also have tried browsing by either of the other fields, namely 'occupational classification' and 'time period'. However, when we browse by occupational classifications (e.g. Figure 12), we see that there is no differentiation between different national-level classifications (such as SOC-2000) within the schemes used (this method of browsing is mostly relevant when searching for cross-national classifications such as ISCO). On the other hand, when we browse by time periods, and select the period 2000-present, we do see several apparently relevant resource (e.g. Figure 13). However a downside of this searching method is that we also see lots of irrelevant hits such as resources from other countries.

Figure 12

Figure 13



An important feature of the 'browse' facility in GEODE is that is only applies to 'uncurated' occupational information resources (see section 2.3).

2.2 Searching

Next we will comment on searching GEODE on certain keywords. It is useful to know a little bit about the way that the 'search' facilities implemented within the GEODE portal work. There is more detail on the GEODE webpages and in Lambert and Tan (2007).

Like many internet search engines, the GEODE search facility has some tricky features which are not always user-friendly. In particular the existing search algorithms have some features which can catch out a careless user. Some important features include:

- search terms are not case sensitive
- search rules operate with some differences for 'curated' and 'uncurated' occupational information resources stored at GEODE (cf. section 2.3)
- search rules applicable only to 'uncurated' resources
    - hyphens are interpreted as characters
        - *e.g. SOC-2000 is a different word from SOC2000*
    - double quotes are used to search for exact expressions
    - the character '*' can be used to indicate unspecified or 'wildcard' text at the start or the end of a word, but it cannot be used to indicate the middle section of a word, for example:
        - *soc* returns any record with words beginning with 'soc'*
        - **2000 returns any record with words ending with 2000*
        - *soc*2000 returns any record with words beginning with 'soc' (it ignores the 2000)*
    - Additional symbols can be used to specify combinations of requirements for multiple words (examples are given on the portal link 'help on search'), for example:
        - *two words entered without any other text is an 'or' search, returning records which feature either one or the other word*
        - *Two words entered with a + sign in front of both words is an 'and' search, returning records which feature both words*

The purpose of searching the GEODE service would usually be to identify any possible occupational information resources which could be relevant to the study in hand. Here, we may search GEODE across a wide range of possible terms which might be relevant to our interests in working with the HOCS. We know that we are interested in data from the UK / Britain, and we have occupational data in SOC-2000 unit groups and in employment status categories. We also know in advance of some occupation-based social classifications which we may be relevant in (NS-SEC and SC). We also have additional background data on measures such as gender and religion. Some possible search terms are:

SOC-2000 SOC2000 SOC SOC-2000 SOC2000 +Employment +status Class NSSEC NS-SEC "SC" "Social class" "Social classification" "Standard occupational classification"	Britain UK England "United kingdom" Gender Religion Ethnic*	"David Rose" Rose Ganzeboom Goldthorpe Lambert

Implementing some of the above terms will retrieve a wide range of different results. We will comment on how to deal with search results in section 2.3, but first we will describe some features of searching on the terms above:

- Occupation-related terms
    - Note that the search for 'SOC2000' and for 'SOC-2000' leads to different hits, because the hyphen symbol is interpreted as if it was a letter. Our advice is usually to try both searches in such situations.
    - The search for 'SOC2000 SOC-2000' generates several hits for uncurated resources (all entries with either term in it) but no hits for curated resources (all entries with exactly both terms)
    - Some terms generate more hits than is useful – for example 'SOC' and 'class'
    - The search for "SC" probably generated no hits. In fact, there are several resources at GEODE which are coded to the social class scheme which in the HOCS data is labelled as SC – but the resources on GEODE refer to this scheme by other names, including the 'registrar general's social class scheme' and 'rgsc'.
- Other terms
    - Gender and ethnicity: You may recall that our example dataset (from the Home Office Citizenship Survey) included data on gender and religious group (Figure 1). Since there are a number of diverse occupational information resources available at GEODE, it is worthwhile considering if there is any suitable occupational information which would engage with the substantive themes of our analysis. In fact, at time of writing, there is one data resource which may be relevant to ethnic differences by occupational groups (Blackwell, 2001), and one concerned with gender inequalities (the gender segregation indexes at SOC90 units associated with Hakim's (1998) study. Either of these resources might be fruitfully exploited for an analysis based on occupational data in the HOCS.
    - Author names: Using author names can sometimes be a useful way to locate specific occupational information resources, though there are also some difficulties here. For instance, David Rose and Harry Ganzeboom have authored several outputs related to national and international social class schemes, and searchers for 'David Rose', 'rose' and 'ganzeboom' all generate various several suitable records. However, examples which don't work as well are the searches for 'goldthorpe' and for 'lambert'. Goldthorpe has also authored many relevant occupational information resources, but searches on 'goldthorpe' reveal fewer hits than expected (one reason is that the schemes associated with Goldthorpe have sometimes been released by other authors, and may be named by other phrases such as 'EGP' and 'Casmin'). By contrast, searching for 'lambert' generates a great many more hits than the number of occupational information resources originally created by that author . Here, the issue is that many resources published by other authors have been supplied to the GEODE service by Lambert, meaning that his name is on most of the GEODE resources in some location or other.

In summary, searching across GEODE is at present a challenging process, since numerous alterative search terms are likely to be relevant to any particular requirement, but all will tend to generate slightly different results. Future work is planned to further enhance the GEODE searching service. Nevertheless it is hoped from the above that readers will see that searching the GEODE service can lead to productive results.

2.3 Interpreting the results after searching and browsing

To return to our practical example, we wish to identify resources at GEODE which may be used to attach some social class measures to our HOCS extract. The searches suggested above all generate some relevant data, but there are too many hits to navigate easily.

Two groups of occupational information resources

It is important to understand that there are two classes of resources stored at GEODE which are searched by the GEODE search engine. These are referred to as 'uncurated' and 'curated' resources (see also Lambert & Tan, 2007).

'Uncurated' resources (Figure 14, panel 1) are data files, or links to data files, in any format, which have been notified to the GEODE portal, but haven't been subjected to any further treatment (curation). They are available for other users to access in their original format (e.g. to download).

'Curated' resources (Figure 14, panel 2) are occupational information resources which have been integrated into the GEODE file matching process. These are the links that appear towards the end of search results. These resources, which connect to data files available from corresponding 'uncurated' resources, are data files which have had sufficient metadata added to them that they can be used to run the GEODE portal file matching procedure on them.

Broadly, uncurated resources are the sort of occupational information files which are already available through open access internet sites. These resources are generally easier to understand, and are the first point of interest for most users of the GEODE service. Curated resources, on the other hand, are occupational information files which have had specialist information added to them within the GEODE service. All curated resources necessarily link to one or more uncurated resource – meaning that curated resources constitute a subset of uncurated resources.

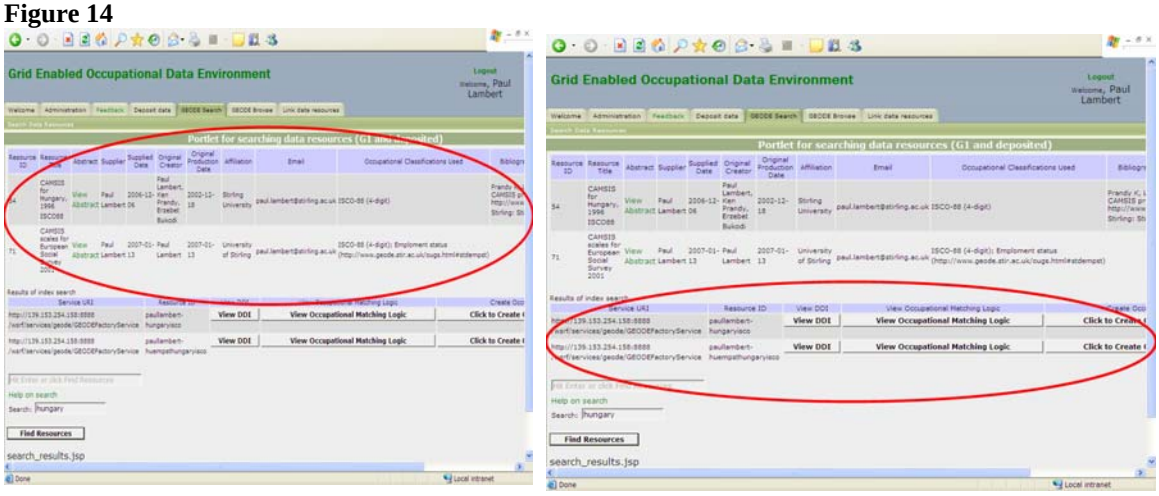
Now, in terms of searching at GEODE, it is useful to understand that the GEODE search engine searches over entries in the 'metadata' of occupational information resources (metadata is descriptive information which describes the data resource, see esp. Lambert et al., forthcoming 2007). When GEODE searches across relevant occupational information resources, it searches different volumes of metadata for each group of files:

- With 'uncurated' resources, the search is performed across a small range of descriptive data about the relevant occupational information resource (its metadata), such as its title and abstract.
- Searches on uncurated resources use a wider range of search rules (e.g. section 2.2)
- With 'curated' resources, the search is performed across a much wider range of metadata, which includes extended 'xml' format information files pertaining to the relevant resources.
- Search algorithms across 'curated' resources apply different (stricter) rules than those across uncurated data

- Some consequences of these two types of resources are that
    - Simpler search terms often result in more hits from curated resources than uncurated resources, because there is more metadata associated with the latter.
    - More complex search terms often result in more hits from uncurated than curated resources, because the rules applied to the search terms are more flexible for uncurated resources

For example, at time of writing a search for “hungary” generates the results shown in Figure 14.

Figure 14



Panel 1: Uncurated resources **Panel 2: Curated resources**

Exploiting results from searching / browsing

A good search term for the HOCS data requirement would be to enter: **SOC2000**

This leads to two useful uncurated resources and two relevant curated resources – see Figure 15 and Figure 16 respectively.

Figure 15

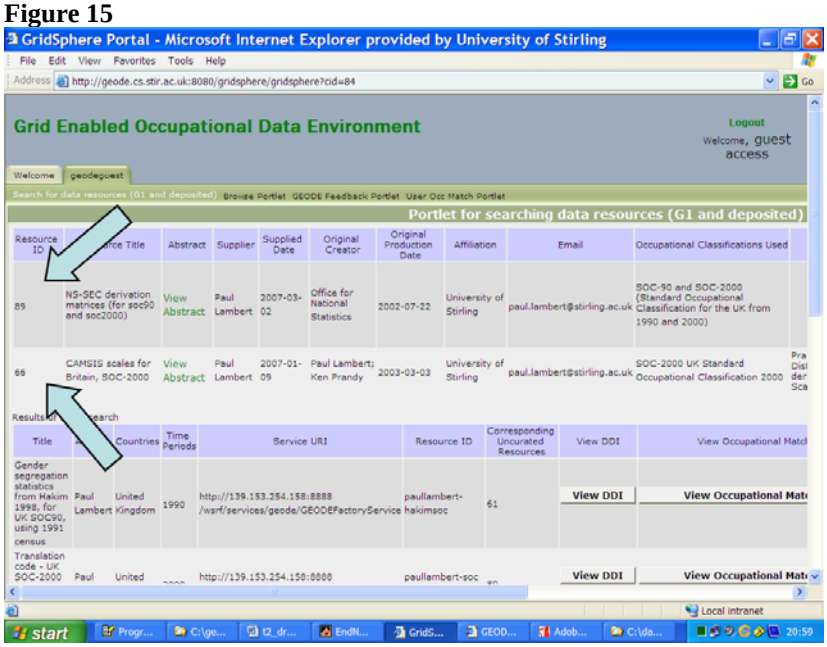
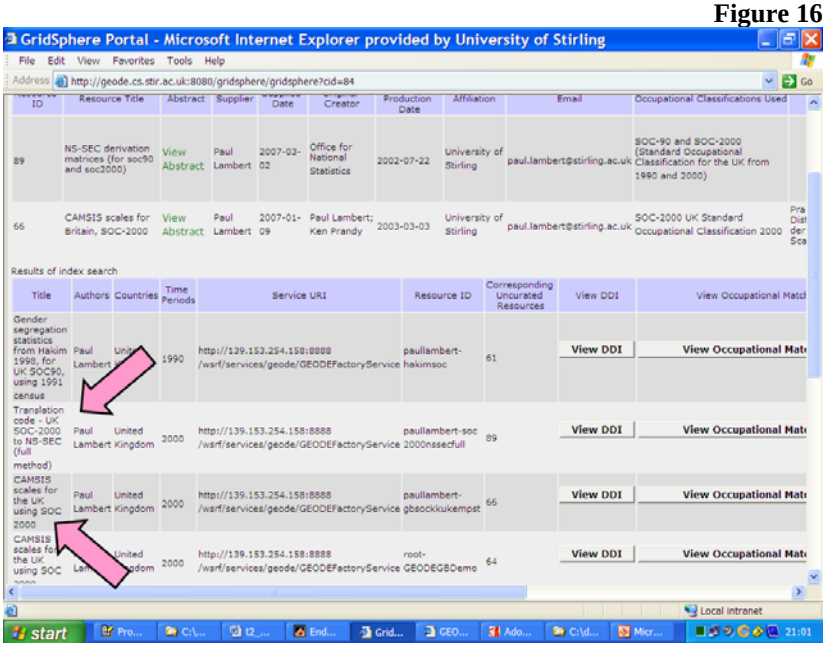


Figure 16



We will ignore the 'curated' resources (Figure 16) until Section 3.

The 'uncurated' resources of most interest are listed in Figure 15 as those with GEODE reference numbers 66 and 89. Both of these records feature a brief description of an occupational information resource, and further links, including a URL to an external page where more on each resource can be found, and a link to directly download the data files referred to. These links lead you to resources (external to the GEODE project) which can be used for linking SOC-2000 occupational data to a number of occupation-based social classifications.

Ordinarily, users of GEODE would search (or browse) across resources and review the descriptions of any relevant ‘uncurated’ resources to see if they may be of value. Accessing these uncurated resources may be for many users the full extent of interest in the GEODE service. In this instance, GEODE has acted merely as a ‘library’ facility providing summary data on external occupational information resources which may be of help to our analysis.

3. USING GEODE: LINKING DATA THROUGH SOC-2000

As well as providing summary information on occupational information resources, GEODE also tries to offer further services to assist social scientists working with occupational information. One of these involves allowing users to deposit occupational information resources at GEODE (not discussed in the example application of this paper). The second provides facilities for directly linking occupational data with other data files, which is elaborated in this and the next section.

Facilities for linking data files are the main purpose of the 'curated' data resources seen in Figure 16[2].

In the HOCS scenario, the useful curated resources are called 'soc2000nssecfull' and 'gbsockkukempst' – see Figure 16. Use of the curated resources at GEODE is also described on the project web-pages http://www.geode.stir.ac.uk/ and in Lambert and Tan (2007), where it is noted that these resources are not especially user friendly. The text below gives a step-by-step example of using both of these resources in order to link in new social classifications to the HOCS data.

Q: How do we know that these two resources, 'soc2000nssecfull' and 'gbsockkukempst', are relevant / useful here?

A:

There is a small clue in the 'title' given to each of these resources. According to this, they both seem to be facilities for using with SOC-2000 data, in order to generate the 'NSSEC (full method)' and 'CAMSIS codes'. There is also, ultimately, a lot more information about the resources available in the 'View DDI' linkage, but this data is of a specialist nature and is best ignored at this stage. In fact, the critical source of information for non-specialist users is the column 'corresponding uncurated resource'. Every 'curated' resource originates from an uncurated resource (secton 2.3). From this column, it is possible to trace the origins of each of these two resources in a more interpretable form. In these cases:

- ***soc2000nssecfull*** *– is a data file based upon the matrices described in 'uncurated resource number 89'. These are matrices published by the UK's Office for National Statistics which detail how combinations of SOC-2000 and employment status measures should be assigned to NS-SEC categories. In fact the GEODE file soc2000nssecfull includes four different derivations of the NS-SEC scheme. Those using the full method to produce operational NS-SEC categories with and without the use of employment status data, and those using the full method to produce analytical NS-SEC categories with and without the use of employment status data. Employment status data must be available in a certain format and coded to 0 if it is not known (see below, Figure 25).*

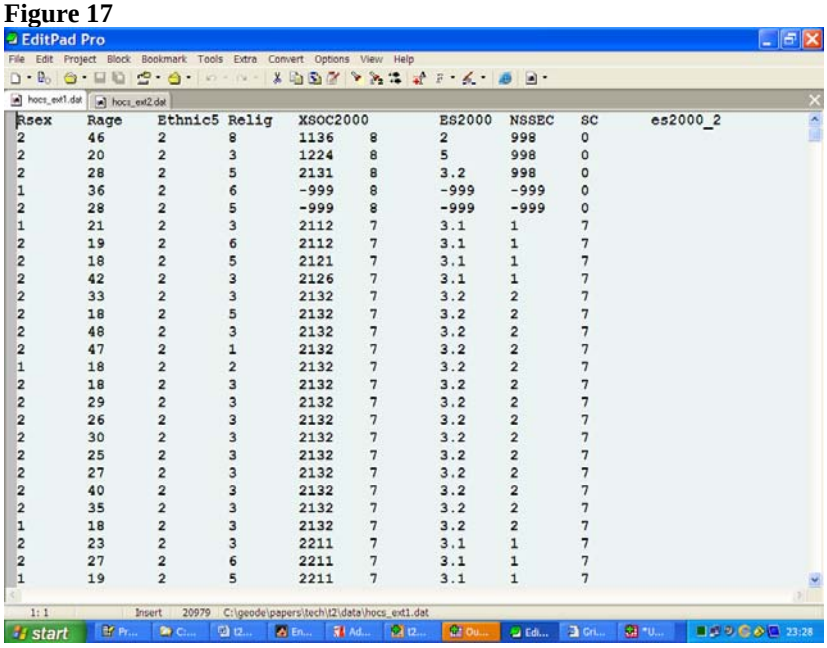
[2] Although primarily geared to the 'linking' services described below, these resources may in some circumstances also be of interest to information scientists and to specialists in working with occupational data, since they feature a greater volume of documentation on the occupational information resources.

- ***gbsockkukempst*** *– is a data file based upon the data described in 'uncurated resource number 66'. This is an index file published by the CAMSIS project (www.camsis.stir.ac.uk) which details how combinations of SOC-2000 and employment status measures should be assigned to a number of different occupation-based social classifications, including CAMSIS scale scores for men and women, and social class categories in a number of popular schemes. Employment status data must be available in a certain format and coded to 0 if it is not known (see below, Figure 25).*

To exploit these files:

- The first requirement is to export the SPSS format HOCS extract file into a plain text version of the same data. This is necessary because of the structure of the GEODE linking service, which runs on plain text files[3]. (In future revisions of the GEODE service, this may be eliminated). SPSS syntax to achieve this transformation is shown in Appendix Table A2. A plain text variation on the existing data, after recoding the employment status variable to the values required by the two resources above, is shown in Figure 17.

Figure 17



- The next stage involves using the GEODE portal to run a linking process with this plain text data file. In fact, both of the two curated resources shown above in Figure 16 can be used in this way. To implement the linkages, we need to do the following:

- Click on the link buttons 'click to create occupational resource' (see Figure 18) for both of the two curated resources (it is necessary to repeat the search for soc2000 after the first creation, in order to have the option for the second resource. Then click on the GEODE portal's link for 'user occ match portlet' (if you are logged in as a guest – see Figure 19) or alternatively for 'link data resources' (if you are logged in as a named user – see Figure 20).

[3] Notes on this requirement are given on http://www.geode.stir.ac.uk/file_convert_info.html .

Figure 18

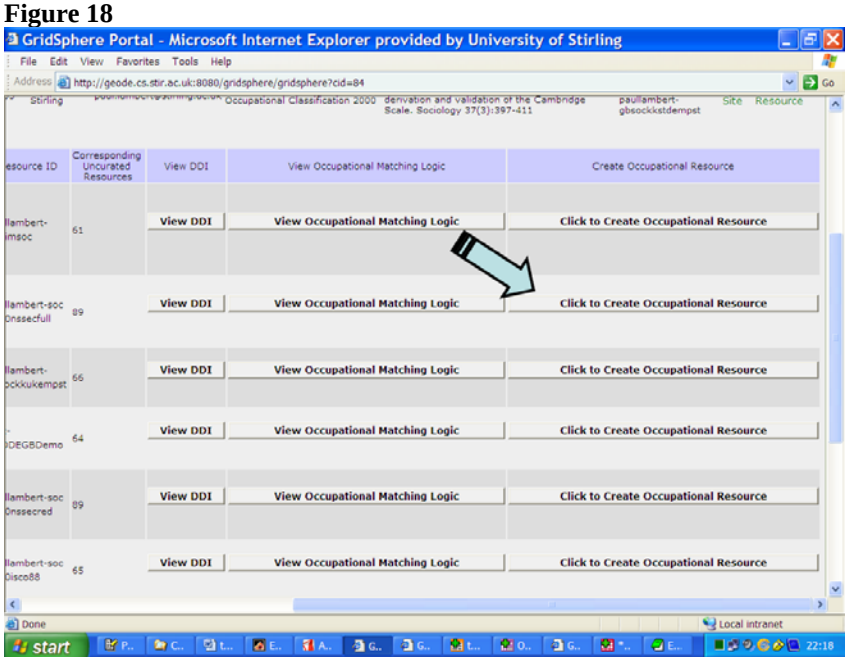


Figure 19

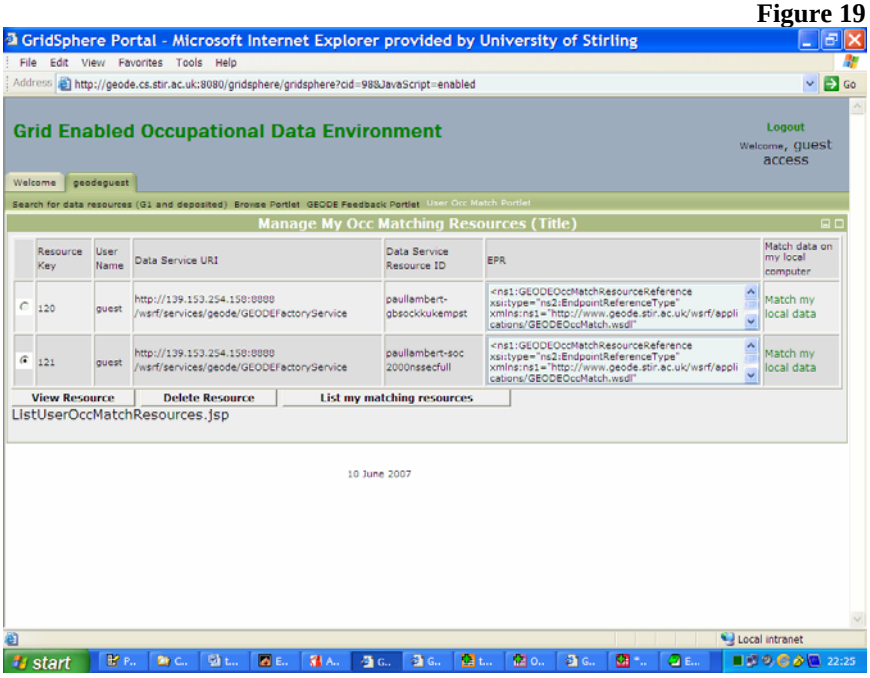
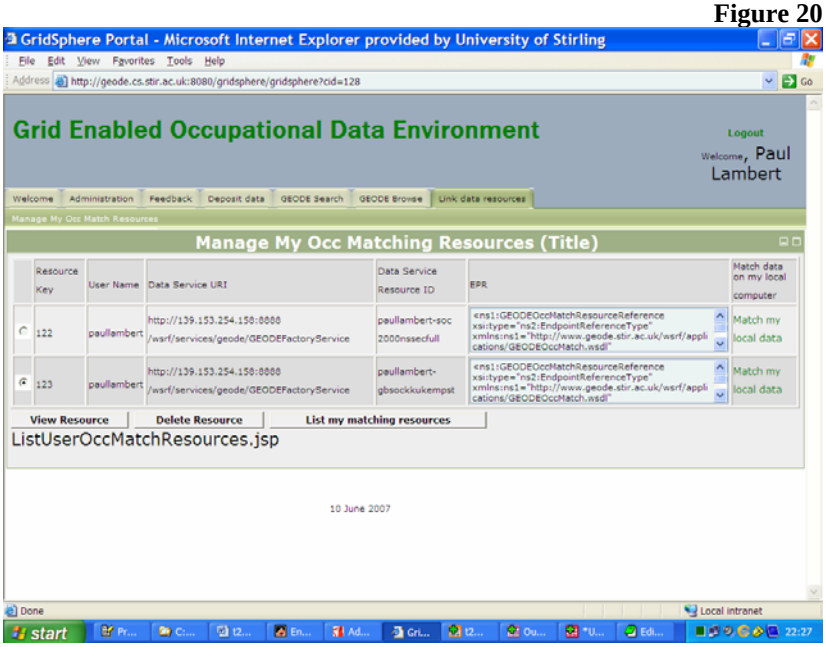


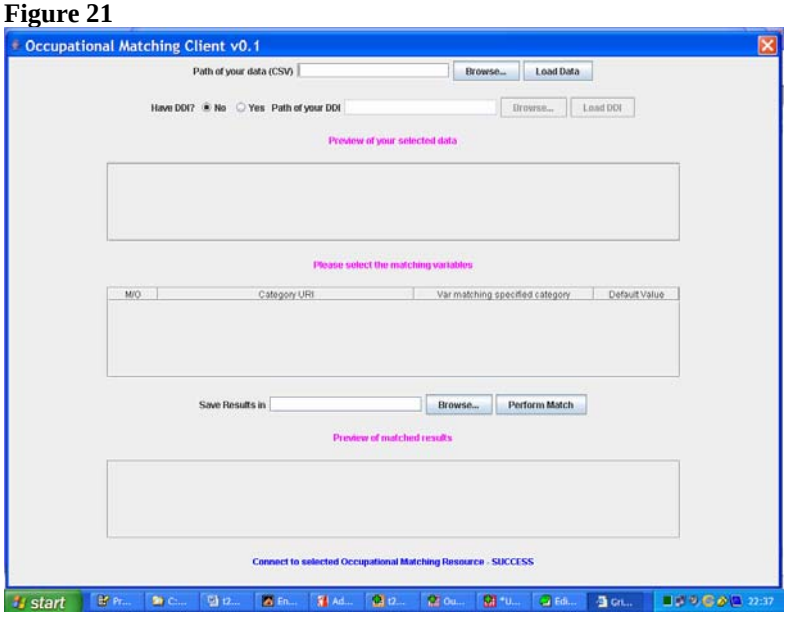
Figure 20



Comment: The portlets illustrated in Figures 19 and 20 show temporary 'resource keys' which are intermittent to the process of matching occupational data. Their details would usually be ignored.

Next, within the relevant 'link data resources' (Figure 20) or 'user occ match portlet' (Figure 19) page, you need to click on the 'match to my local data' link. This launches a Java application (Figure 21) which will perform the linkage. There is further description of this application at the GEODE project webpages http://www.geode.stir.ac.uk/ . *(Depending on how Java is installed on your machine, you may need to dismiss -clicking 'run' or 'ok' - certain pop-up error messages which may appear when you launch this application).*

Figure 21



Once launched, the key screen looks as shown in Figure 21. It is useful to remember that any screen you see in this application is referring strictly only to the particular occupational information file from which it was launched (whichever of the two resources seen in Figures

19 and 20). In this example we will begin with the Java application seen for the resource called ‘soc2000nssecfull’.

- It is next necessary to identify the micro-data which is going to be linked to the relevant occupational data. This is done by clicking on browse and then load data – see Figures 22 and 23.

Figure 22

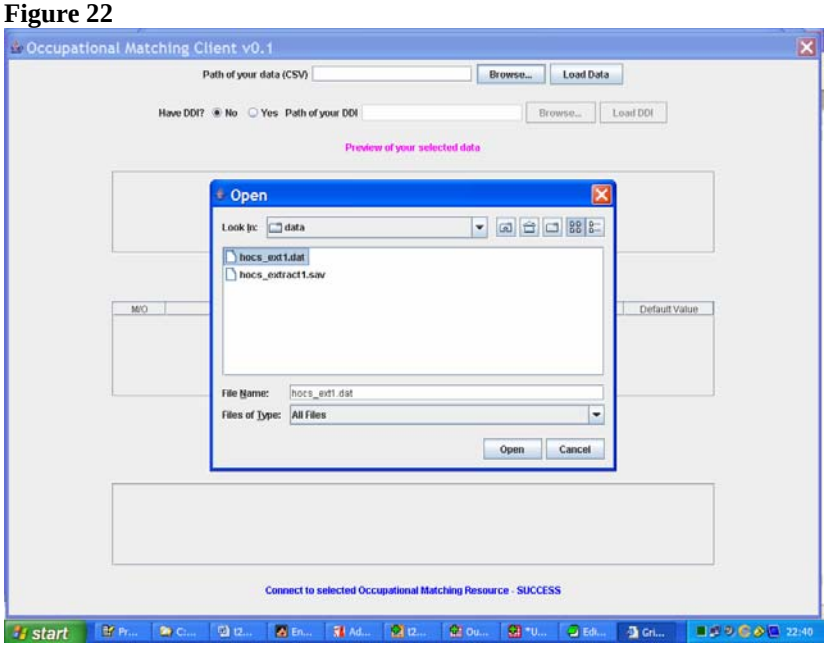
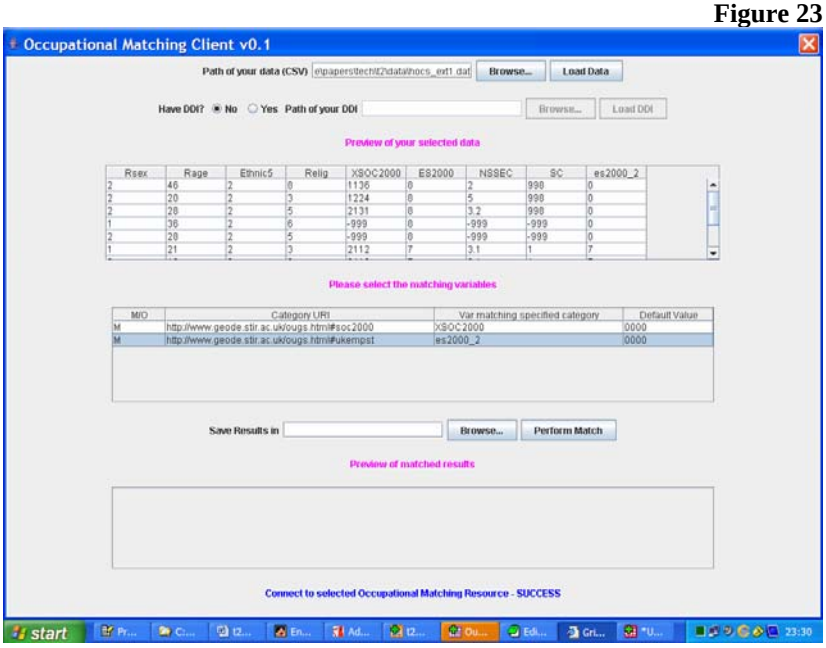


Figure 23



- Once data is entered into the portal in this fashion, it is possible to see two new options under the heading ‘please select your matching variables’. **These are critical to using the GEODE file linking programme.** These tell you what sort of occupational information is required by the occupational information resource. In this example the resources is the file called ‘soc2000nssecfull’, which is the SOC-2000 to NS-SEC linkage provided by the Office for National Statistics. This file requires a

standardised variable for SOC-2000 and for employment status. To use the GEODE linking service, it is necessary to state which variables in the user's own file correspond to the format of those variables.

- A useful feature of the Java application is that a link to descriptive information about the relevant resources is available. We can click on either to find a webpage describing what is required of the 'soc2000' (Figure 24) and 'ukempst' (Figure 25) variables. Fortunately in this case, we have both variables present in our micro-data file (in the columns labelled 'xsoc2000' and 'es2000_2' – the latter was a minor recoding of the 'es2000' variable). We therefore fill out the drop down list to connect with those variables.

Figure 24

Figure 25

Comment: It may often be the case that not all of the required matching variables are available in the micro-data file. It is sometimes useful to create appropriate versions of the variables manually, for instance, calculating a 'ukempst' format variable from multiple variables with information on employment status. Also, it is possible to indicate missing values on the matching variables.

- The last actions necessary are to specify the (new) output file to which results will be sent - this is possible by clicking the 'browse' option at 'save results in' – then to click 'perform match'. This will invoke the file matching process (including a preview of the outputs as they are created in blocks of cases (Figure 26). The final output is a new plain text file with additional occupational data (Figure 27).

Figure 26

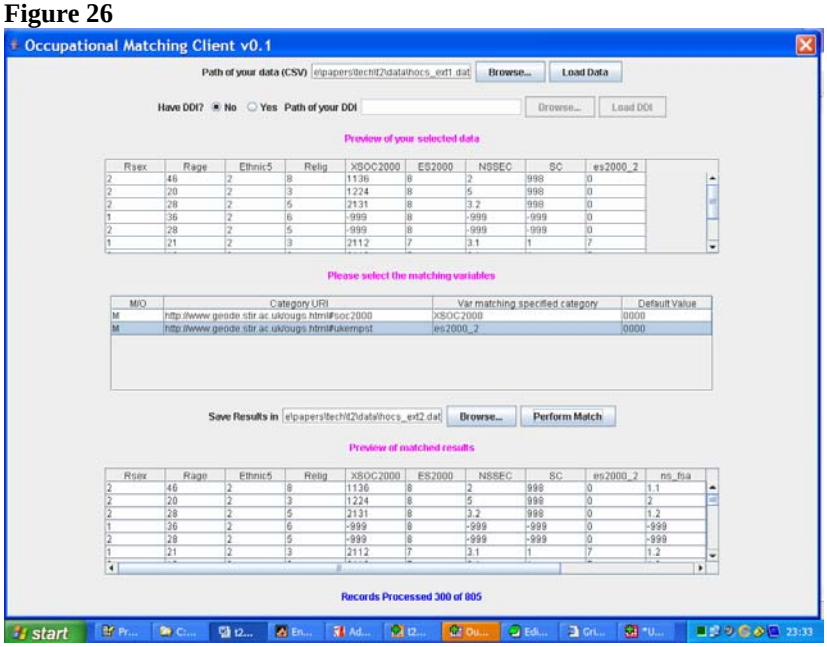
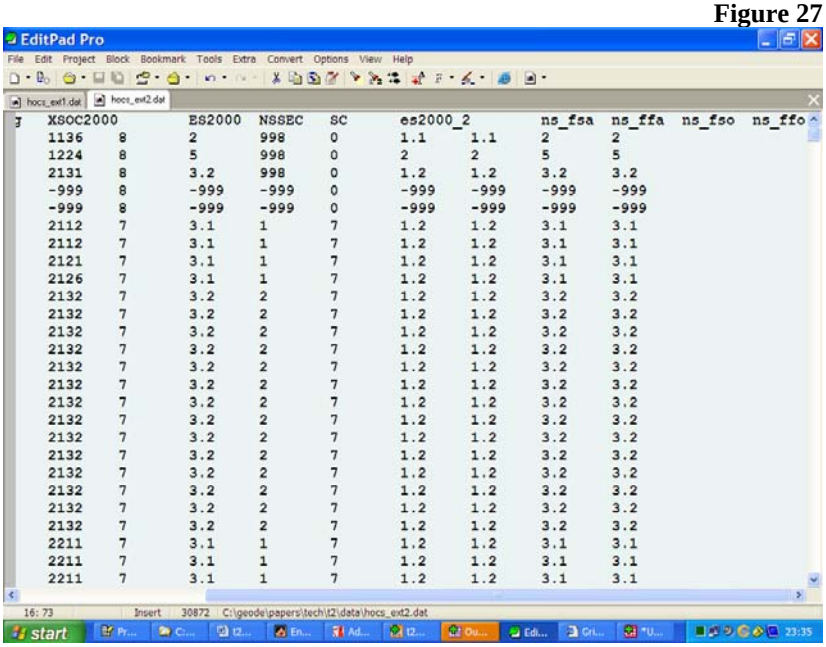


Figure 27



In this example we have identified two curated occupational information resources of interest. We next repeat the above process using the second data file – called gbsockkukempst – by again clicking the relevant 'match occupational information' link (Figures 19 or 20) and filling out the same responses. Note that the input file may now be the output file created above (Figure 28), and that the final output data now adds additional variables from the second occupational information resource (Figure 29).

Figure 28

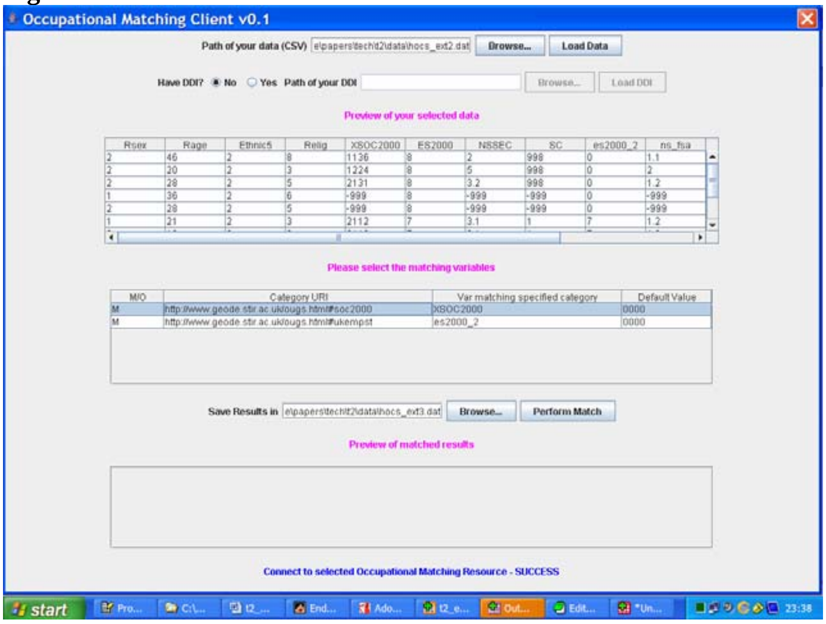
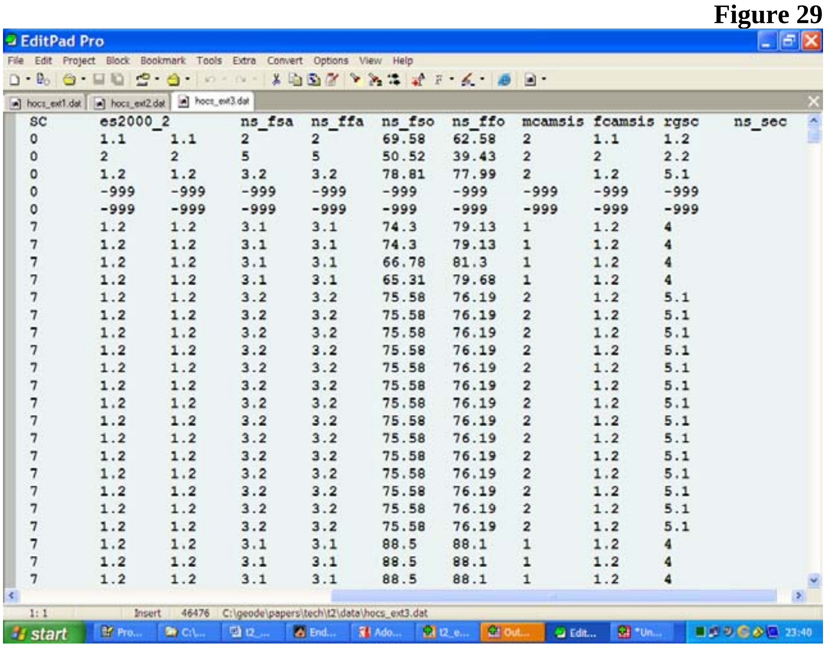
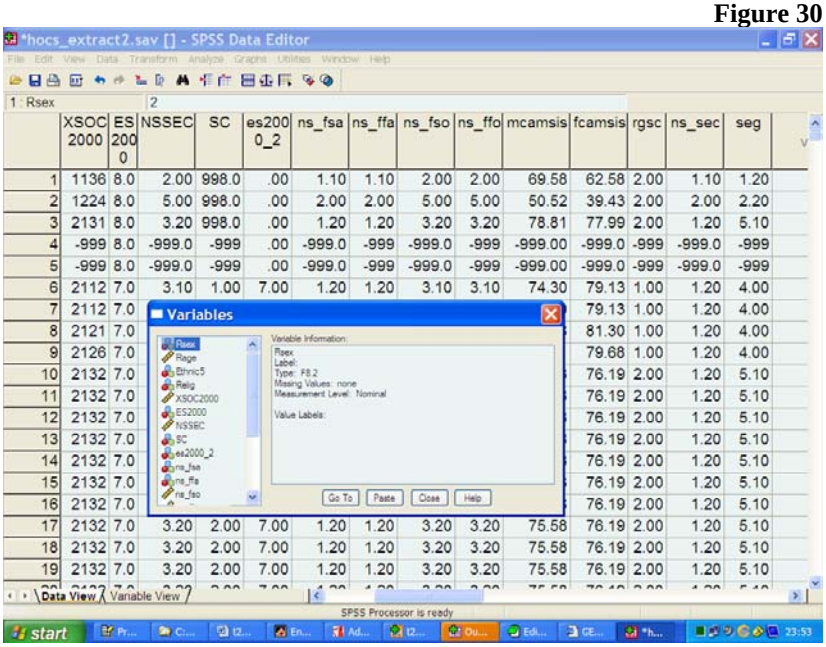


Figure 29



For most users, the final stage of linking data would now be to read these values back into an SPSS format data file. If this is done using the example syntax file shown in Appendix Table A3, the resulting output is seen in Figure 30.

Figure 30



Summary of linkage

We have started with the following occupational data:

XSOC2000	*Standard Occupational Classification 2000*
ES2000	*Employment status (in standardised categories)*

By exploiting externally published occupational information files indexed at GEODE, we have added to our micro-data file the following new variables:

ns_fsa	*National Statistics Socio-economic Classification (NS-SEC) full method simple analytical version (uses fewer class categories, ignores employment status data)*
ns_ffa	*National Statistics Socio-economic Classification (NS-SEC) full method analytical version (uses fewer class categories, uses employment status data)*
ns_fso	*National Statistics Socio-economic Classification (NS-SEC) full method simple analytical version (uses more class categories, ignores employment status data)*
ns_ffo	*National Statistics Socio-economic Classification (NS-SEC) full method simple analytical version (uses more class categories, uses employment status data)*
mcamsis	*Male CAMSIS scale scores*
fcamsis	*Female CAMSIS scale scores*
rgsc	*Registrar General's Social Class category*
ns_sec	*National Statistics Socio-economic Classification (NS-SEC) full method simple analytical version (uses fewer class categories, uses employment status data)*
seg	*Socio-Economic Group (class categories)*

Note that in some cases we have generated the same data on more than one variable. One feature of this is that we have undertaken a useful check on the reliability of different operationalisations of the same social classifications.

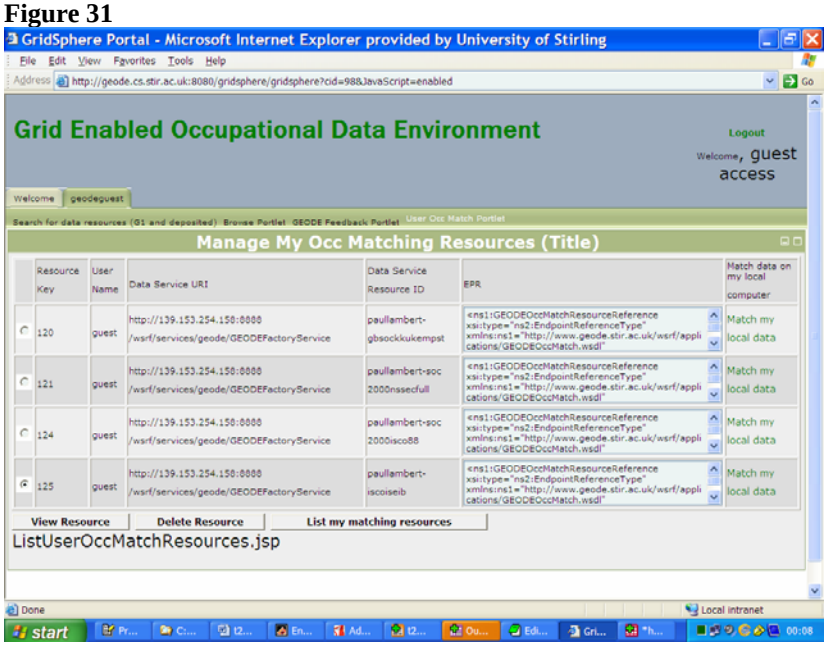
4. USING GEODE: LINKING DATA THROUGH ISCO-88

Starting from the data file shown in Figure 30, there are in fact many more occupation-based social classifications which could be linked with SOC-2000 data. One resource is to search for further occupational information files which connect other SOC-2000 measures with social classifications – there are other resources available. However, another approach which it is useful to illustrate concerns using the resources at GEODE first to translate from SOC-2000 into another occupational unit group scheme, then to exploit an occupation-based social classification for such another unit group. An obvious example would be to use the influential International Standardised Classification of Occupations 1988 (ISCO-88), and its published linking with the International Socio-Economic Index stratification scale (Ganzeboom, 2007; Ganzeboom & Treiman, 2003).

To do this in GEODE, it is necessary to identify the two curated occupational information resources which support these translations, then replicate the matching procedures used above to achieve this. This is shown in the following examples.

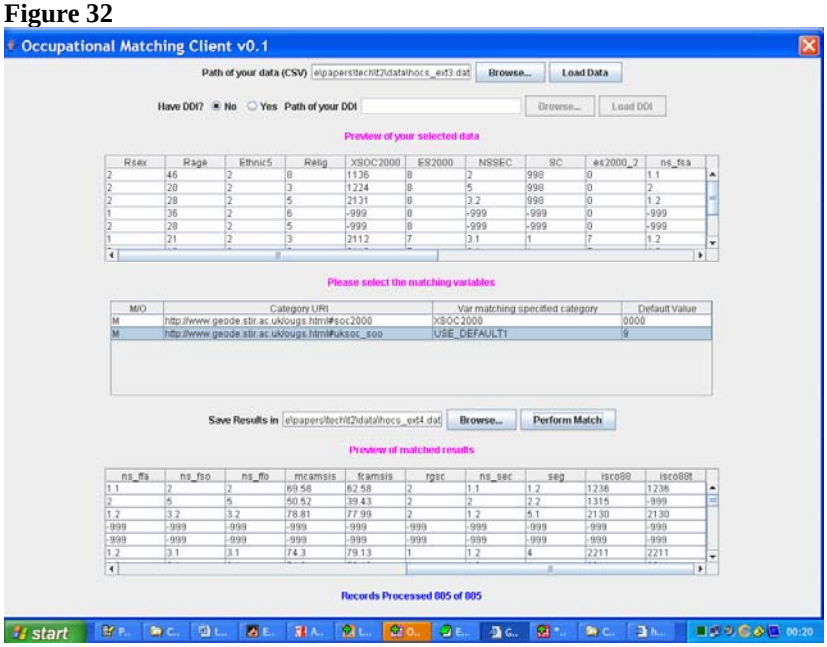
The two 'curated' resources required for this procedure are called 'soc2000isco88' and 'iscoiseib' (Figure 31).

Figure 31



The occupational matching Java interface for the SOC2000-ISCO88 conversion looks as is shown in Figure 32. Note that this match would, if available, use a variable measuring 'size of establishment' (uksoc_soo)'. This variable is not easily available in our HOCS extract, so it is set to a default value with the value 9 (for unknown).

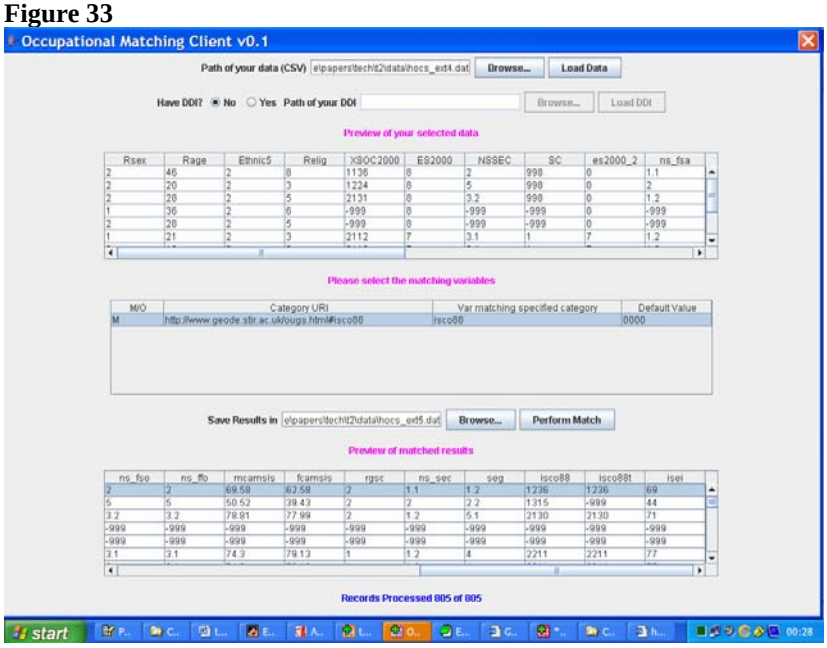
Figure 32



We have now added on ISCO-88 codes to the SOC2000 categories. For instance, the first row in our data was coded to SOC2000 category 1136 ("Information and communication technology managers"); it has been recoded into ISCO88 category 1236 ("Computing services department managers").

The occupational matching Java interface for the ISCO-88 to ISEI linkage looks as is shown in Figure . The end result is a new dataset with SOC2000 linked to ISEI scores, as well as other social classifications listed above (e.g. Figure 2). This illustration indicates the extended analyses that could in principle be brought to bear on occupational data files using the resources indexed at GEODE.

Figure 33



Appendix

Table A1: SPSS Syntax used in extracting the micro-data file used in this example, from the original data HOCS data downloaded from the UK Data Archive.

```
get file="c:\data\hocs\2005\hocs_2005_data_archive_version.sav".

select if (es2000 ge 1 & es2000 le 8).
select if (ethnic5=2).
select if (rage ge 18 & rage le 50).
sort cases by es2000 (d) soc2000 (a).

sav out="c:\geode\papers\tech\t2\data\hocs_extract1.sav"
   /keep = rsex rage ethnic5 relig
              soc2000 xsoc2000 es2000 nssec sc   .
```

Table A2: SPSS Syntax used in exporting the HOCS extract file to plain text format *(This example recodes missing data indicators to numeric values, and eliminates the string version of the SOC variable).*

```
get file="c:\geode\papers\tech\t2\data\hocs_extract1.sav"
     /drop=soc2000 .

recode all (missing,sysmis=-999).
missing values all (-777).
descriptives var=all.

save translate /outfile="c:\geode\papers\tech\t2\data\hocs_ext1.dat"
      /type=tab /fieldnames /replace .
```

Table A3: SPSS Syntax used in reading plain text data files generated by GEODE matching service to SPSS format.

```
get translate /file="c:\geode\papers\tech\t2\data\hocs_ext3.dat"
       /type=tab /fieldnames.
* (with NS-SEC and CAMSIS data now added).

sav out="c:\geode\papers\tech\t2\data\hocs_extract2.sav".



get translate /file="c:\geode\papers\tech\t2\data\hocs_ext5.dat"
       /type=tab /fieldnames.
* (with ISCO88 and ISEI codes now added).

sav out="c:\geode\papers\tech\t2\data\hocs_extract3.sav".
```

References

Armstrong, W. M. (1972). The use of information about occupation. In E. A. Wrigley (Ed.), *Nineteenth Century Society: Essays in the use of quantitative methods for the study of social data* (pp. 191-310). Cambridge: Cambridge University Press.

Bechhofer, F. (1969). Occupations. In M. Stacey (Ed.), *Comparability in Social Research* (pp. 94-122). London: Heinemann (in association with British Sociological Association / Social Science Research Council).

Blackwell, L. (2001). *1991 Census Ethnic Group Occupations [computer file]*. Colchester, Essex: UK Data Archive [distributor], SN: 4357.

Ganzeboom, H. B. G. (2007). Tools for deriving status measures from ISKO-88 and ISCO-68. Retrieved 1 June, 2007, from http://home.fsw.vu.nl/~ganzeboom/PISA/

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standarised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnick & C. Wolf (Eds.), *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables* (pp. 159-193). New York: Kluwer Academic Press.

Hakim, C. (1998). *Social Change and Innovation in the Labour Market : Evidence from the Census SARs on Occupational Segregation and Labour Mobility, Part-Time work and Student Jobs, Homework and Self-Employment*. Oxford: Oxford University Press.

Hendrickx, J., & Ganzeboom, H. B. G. (1998). Occupational Status Attainment in the Netherlands, 1920-1990. A Multinomial Logistic Analysis. *European Sociological Review, 14*, 387-403.

Home Office. (2006). *Home Office Citizenship Survey, 2005 [computer file]*. Essex: UK Data Archive [distributor], SN: 5367 (Home Office Communities Group and National Centre for Social Research).

Lambert, P. S. (2002). Handling Occupational Information. *Building Research Capacity, 4*, 9-12.

Lambert, P. S., & Tan, K. L. T. (2007). *Instructions for Using the GEODE Portal, Edition 1.1*. Stirling: GEODE Project Technical Paper No. 1, University of Stirling, and http://www.geode.stir.ac.uk.

Lambert, P. S., Tan, K. L. T., Turner, K. J., Gayle, V., Prandy, K., & Sinnott, R. O. (2006). *Developing a Grid Enabled Occupational Data Environment*. Paper presented at the Second International Conference on e-Social Science.

Lambert, P. S., Tan, K. L. T., Turner, K. J., Gayle, V., Prandy, K., & Sinnott, R. O. (forthcoming 2007). Data Curation and Social Science Occupational Information Resouces. *International Journal of Digital Curation*.

Marsh, C. (1986). Occupationally Based Measures. In A. Jacoby (Ed.), *The Measurement of Social Class* (pp. 1-47). London: Social Research Association.

ONS. (2000). *Standard Occupational Classification 2000, Volume 1: Structure and description of unit groups*. London: Office for National Statistics.

ONS. (2002). *The National Statistics Socio-economic Classification User Manual (Version 1 April 2002)*. London: Office for National Statistics and http://www.statistics.gov.uk/nsbase/methods_quality/ns_sec/.

OPCS. (1991). *Standard Occupational Classification, Volume 3: Social Classifications and Coding Methodology*. London: Office for Population Censuses and Surveys.

Prandy, K. (1998). Deconstructing classes: Critical comments on the revised social classification. *Work Employment and Society, 12*(4), 743-753.

Prandy, K. (2002). Ideal types, stereotypes and classes. *British Journal of Sociology, 53*(4), 583-601.

Reid, I. (1998). *Class in Britain*. London: Polity.

Rose, D., & Pevalin, D. J. (Eds.). (2003). *A Researcher's Guide to the National Statistics Socio-economic Classification*. London: Sage.

Tan, K. L. T., Gayle, V., Lambert, P. S., Sinnott, R. O., & Turner, K. J. (2006). *GEODE - Sharing Occupational Data Through the Grid*. Paper presented at the 5th UK e-Science All Hands Meeting.

Weeden, K. A., & Grusky, D. B. (2005). The Case for a New Class Map. *American Journal of Sociology, 111*(1), 141-212.