Instructions for Using the GEODE Portal, edition 1.1

Paul S. Lambert	University of Stirling
Koon Leai Larry Tan	University of Stirling

11th June 2007 [Edition 1.1]

GEODE Project Technical Paper No. 1

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 to the GEODE project occupational information resources

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 (see section 2).

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.

The GEODE project primarily deals with 'aggregate occupational information resources'. These are electronic data files which contain some descriptive information on a number of different occupational positions. Typical examples of this descriptive information include social class classification schemes (data on which social class particular occupational positions should be placed in); and statistics on occupational circumstances (such as statistics on gender segregation – the proportion of women within any particular occupational position).

Broadly, usage of the GEODE service falls into two categories[1]:

i) Accessing occupational information files (and linking them with other data)

Most users of GEODE are people who wish to access the wide range of occupational data resources indexed under GEODE.

Users may search across the occupational information resources stored in GEODE (see section 3). They can choose to immediately download relevant occupational information (section 3). However most often they will also wish to use the GEODE 'matching' service which automates the matching process linking the GEODE data resources to the user's own data files (see section 4).

Scenarios 1, 2 and 3 below illustrate typical applications in accessing occupational information.

The operations of searching, accessing and matching occupational information can be achieved by entering the GEODE portal as a 'guest' (see section 2).

[1] GEODE is generally oriented to analyses of occupational information which summarises a wider population – for instance, the respondents in a national survey. It is worth being aware that some types of occupational analysis in the social sciences are much more focussed, for instance with an interest in the specific details of particular occupational positions. At present GEODE does not particularly cater to such analytical approaches. Users with interests in the circumstances of particular named occupations would ordinarily be better consulting specialised resources on those occupational positions – for example, http://www.euroccupations.org/main/ .

ii) Depositing occupational information resources

Smaller numbers of users may use GEODE in order to distribute their own occupational information resources to other researchers (see section 5). This exercise is usually undertaken by social scientists with a specialist interest in occupational data analysis. Scenarios 4 and 5 below illustrate typical applications in depositing occupational information.

Depositing occupational information resources on GEODE requires logging into the GEODE portal with a personalised account (see section 2).

For information, in the social sciences this task has previously been achieved by supplying occupational data through dedicated webpages, for example:

Harry Ganzeboom's conversion tools: http://home.fsw.vu.nl/~ganzeboom/pisa/

CAMSIS project: http://www.camsis.stir.ac.uk/

Table 1: Selected scenarios in using GEODE

Usage 1: Access SOC-90 value labels

Scenario: A researcher (using SPSS) has obtained a survey dataset where occupations have been coded to the numeric values of the UK SOC-90 occupational unit group scheme. They wish to attach the textual descriptions for the relevant occupations to their data file.

Expert view: There is an SPSS file called 'UK1990socsubgpsandlabelsv1.sps', which is free to download from http://www.camsis.stir.ac.uk/occunits/distribution.html#UK that gives text value labels for all SOC-90 3-digit units. The user needs to download this file.

GEODE contribution: Login to GEODE as a named user or guest. Use the search engine to search for resources which cover the UK SOC-90 file. The search should reveal the SPSS file 'UK1990socsubgpsandlabelsv1.sps'. The user can immediately download the file from GEODE, and/or may visit the distributing website.

Usage 2: Translate SOC-90 to CASMIN social class scheme

Scenario: A researcher (using SPSS) has obtained a survey dataset where occupations have been coded to the numeric values of the UK SOC-90 occupational unit group scheme. They wish to attach CASMIN (aka. Goldthopre) class scheme values to the relevant occupations to their data file.

Expert view: There is an SPSS file called 'gb91soc90.sav' which is free to download from http://www.camsis.stir.ac.uk/Data/Britain91.html that links SOC-90 3-digit units, in combination with a 1-digit definition of employment status, to the CASMIN class scheme. The user should calculate employment status measures (if they can), and then process the SPSS file either themselves, or by using GEODE.

GEODE contribution: (Once the employment status data is prepared), login to GEODE as a named user or guest. Use the search engine to search for resources which cover the UK SOC-90 file. The search should reveal the SPSS file 'gb91soc90.sav'. The user now has two choices:

i) Immediately download the file from GEODE, and/or may visit the distributing website, and follow its own instructions for linking the data in SPSS with their own records.

ii) Use the GEODE occupational matching programme to process the linkage between their own data files and the GEODE indexed data file *[the programme on GEODE to do this is called 'gbsocukempst']*

Usage 3: Access gender segregation statistics for SOC-90 unit groups

Scenario: A researcher (using SPSS) has obtained a survey dataset where occupations have been coded to the numeric values of the UK SOC-90 occupational unit group scheme. They wish to attach data on gender segregation (the proportion of women nationally within each occupational unit group) to each occupational unit in their data file.

Expert view: There are several sources of gender segregation statistics. One publication (Hakim, 1998) uses data on UK 1991 census to present gender segregation values for each SOC-90 unit. This data has been transcribed into SPSS format and stored at GEODE.

GEODE contribution: Login to GEODE as a named user or guest. Use the search engine to search for resources which cover the UK SOC-90 file. The search should reveal the SPSS file 'soc90seg_hakim1998.sps'. The user can immediately download the file from GEODE. The could also is wanted to use the GEODE portal to undertaking a file matching exercise which links their own data with these statistics *[the programme on GEODE to do this is called 'hakimsoc']*

Usage 4: Supply a data file on occupational positions to GEODE

Scenario: A researcher has prepared some descriptive data on the average income levels held by women in different occupations in the United States in 2004, using the US SOC-2000 occupational unit group scheme. They would like to make this data available to other researchers so that they may attach this information to their records in SOC-2000 units.

Expert view: The file could be deposited to GEODE by uploading it into the index service whilst filling out a small number of questions on the origins of the resource. Once deposited it will be registered with the search engine on GEODE and will then be available to other users of GEODE for download from there. Members of the GEODE project may subsequently enhance its accessibility by extending the data curation process (cf. usage 5).

GEODE contribution: Login to GEODE as a named user (this will require email registration with the GEODE project contacts). Use the 'deposit data' tab to upload the data file or files, providing information on the name of the data producer and a short description of the files. GEODE project members will further curate the data after it has been uploaded to GEODE.

Usage 5: Prepare enhanced meta-data on an occupational information file supplied to GEODE

Scenario: [This process would ordinarily be undertaken by members of the GEODE project]. A data resource has been supplied to GEODE but is currently only available for download by other users in its original format. There is a desire that the data should also be available via the GEODE matching service.

Expert view: The data will only be available for file matching processes after it has been fully curated to the GEODE-M metadata standard. This is a short manual operation which can be undertaken by the data depositor or members of the GEODE project (usually the latter). This operation involves making edits to an xml format data file which contains information on the occupational data file.

GEODE contribution: Login to GEODE as a named user (this will require email registration with the GEODE project contacts). Use the 'deposit data' tab links to edit the metadata file associated with an existing resource. *[This service became available to public users on 8.1.07. Instructions on this are in section 5]*

Further orienting issues:

Occupational index schemes

The GEODE service organises aggregate occupational information data files according to the occupational 'index scheme' to which each file refers. Index schemes include published dictionary style definitions of different occupational titles, known as 'occupational unit group' schemes (OUGs). Also, other index schemes exist to record the occupational position, such as index measures of 'employment status'. When social scientists collect data on the occupational positions of their subjects (for example - the occupations of those who completed a survey questionnaire), the data is usually recorded as a location in such an 'index scheme'. There is a further discussion of index schemes, including a listing of all known occupational index schemes used in GEODE, on the GEODE web-pages at: http://www.geode.stir.ac.uk/ougs.html).

Occupational information meta-data

The indexing facilities associated with the GEODE service hinge upon exploiting appropriate meta-data about the relevant occupational information resources (meta-data is data describing the data resources themselves, such as the author(s) of the resource and the date it was produced). The GEODE service looks for specific metadata entries which are organised in terms of an internationally standardised metadata protocol, the Data Documentation Initiative (http://www.icpsr.umich.edu/DDI/). The GEODE component of this protocol is known as the 'GOEDE-M' metadata standard. Meta-data issues in GEODE are also described in Lambert et al (forthcoming 2007), and on our webpage http://www.geode.stir.ac.uk/geode_m_curation.html . We refer to the process of adding appropriate meta-data to an occupational information resource as the process of 'data curation'.

Data file formats

The GEODE data resources are available either for direct download in their original format, or else, after being fully curated, in the form of a plain text data file which can be processed automatically during the file matching operation. In the former case, resources are often supplied the formats of proprietary packages such as SPSS or Stata (they are simply supplied in whatever formats their original creator developed them in). In the latter case, of file matching, users will need to prepare their own data in plain text format. This is usually a simple process of saving out the data file from the package of choice, but choosing a specialist option in order to save it out into plain text format (which may also easily be read by the programme). Instructions on undertaking these linkages are posted on our webpage http://www.geode.stir.ac.uk/file_convert_info.html .

Micro-data access

The GEODE facilities for linking occupational information with micro-social data on occupations are a unique feature of the GEODE service. However they do apply only to scenarios when the user has access to occupational micro-data on their own machine. This is the most common scenario, though there are some models for accessing survey micro-data where this doesn't hold (for instance the NESTAAR service provided by the UK's Economic and Social Data Service, http://nesstar.esds.ac.uk/webview/, and the LIS project allowing access to cross-national survey data, www.lisproject.org).

2. Entering the GEODE portal

You may enter the GEODE portal by logging in from the entry page (linked from www.geode.stir.ac.uk).

Help on working with the GEODE portal is available from several sources:

- this technical paper (http://www.geode.stir.ac.uk/publications.html)
- the links available from the portal front page
- textual description entries within the portal

You can login to the GEODE portal either as a named user or as a guest.

Named users can:

- Search data
- Download data
- Use the GEODE data matching service
- Deposit data
- Manage those data resources that have been deposited previously
- Edit personal account space settings (e.g. display language and password)

- Named user access requires a personalised account: email Paul Lambert (paul.lambert@stirling.ac.uk) requesting a GEODE account, and login details will be manually created and sent to you at the earliest opportunity.

Guest users can:

- Search data
- Download data
- Use the GEODE data matching service

- Guest user access requires logging in with generic login details (username: guest; password: geode).

The GEODE portal runs in a 'Grid middleware' (i.e. software) environment known as 'Gridsphere'. The resources which you will see after logging in are dependent on which account you logged in with.

Gridsphere arranges resources under a series of tabs and sub-tabs known as 'portlets'. Each portlet within the GEODE portal allows a user to undertake a different operation with the GEODE project's occupational information resources. A site map of the portlets is given below.

Users should beware that using the 'back' key on an internet browser is not a reliable way to navigate through Gridsphere – it is better practice to follow specific links when moving through the site.

Logging out:

- It is preferable to log out manually when leaving the GEODE portal.
- An automatic logout from the GEODE portal occurs after 30 minutes of inactivity.

Table 2: GEODE Portal: Portlet site map

* = not available to guest users

Welcome -> Welcome statment -> Settings* -> Layout* -> Help -> FAQ	 Textual description of the portal Allows changes to settings, e.g. password, time zone (may only be edited by named users) Allows changes to visual layout (may only be edited by named users) Further resources offering help with the GEODE service Further help resources: users' frequently asked questions
-Feedback	Opportunity to convey feedback on the GEODE service
GEODE Search -> Search data GEODE Browse -> Browse data	 Allows searching of GEODE resources (see section 3) Allows browsing of GEODE resources (see section 3)
Deposit data* -> Manage my data resources* -> Manage my G1 data resources*	 Allows upload of new resources (see section 4) and management of resources previously uploaded (see section 5) Allows curation of data which has already been deposited to GEODE – requires xml files to be specified, giving metadata on the occupational information, and subsequently allowing file matching linkages to be deployed
Link Data Resources: -> Manage my occ resources	*(called 'User Occ Match Portlet' if logged in as a guest)* Allows implementation of file matching exercises

3. Searching the GEODE data index service

Guest and named users may use the GEODE portal to search those occupational information resources which have been indexed under GEODE.

This requires use of the ‘search’ portlet and/or the ‘browse’ portlet.

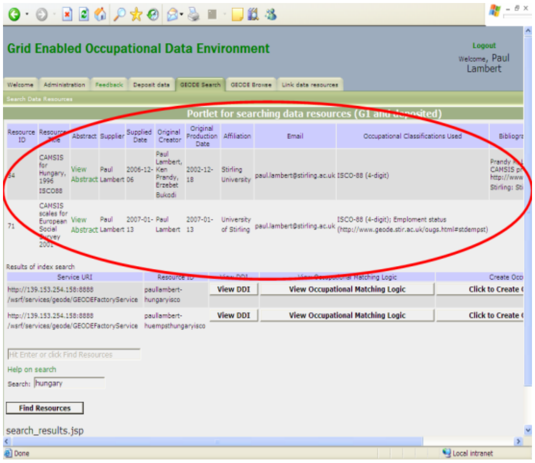
[We are still working on improving the usability of the search engine and browse facilities, 11.6.07].

Search Portlet:

The search portlet accepts any terms, but its usage requires some consideration:

- **(A) There are two groups of resources which your search may uncover.** These are **(1) ‘uncurated’ resources**, which 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), but are not fully integrated into GEODE. There are also **(2) ‘curated’ resources**, which have been integrated into the GEODE file matching process. These are the links that appear towards the end of the file searching linkage. 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.

Uncurated resources:



Grid Enabled Occupational Data Environment

Portlet for searching data resources (G1 and deposited)

Resource ID	Resource Title	Abstract	Supplier	Supplied Date	Original Creator	Original Production Date	Affiliation	Email	Occupational Classifications Used	Bibliogra
54	CAMSIS for Hungary, 1996 ISCO88	View Abstract	Paul Lambert	2006-12-06	Paul Lambert, Ken Prandy, Erzebet Bukodi	2002-12-18	Stirling University	paul.lambert@stirling.ac.uk	ISCO-88 (4-digit)	Prandy K, L CAMSIS pr http://www Stirling: Sti
71	CAMSIS scales for European Social Survey 2001	View Abstract	Paul Lambert	2007-01-13	Paul Lambert	2007-01-13	University of Stirling	paul.lambert@stirling.ac.uk	ISCO-88 (4-digit); Emploment status (http://www.geode.stir.ac.uk/ougs.html#stdempst)	

Results of index search

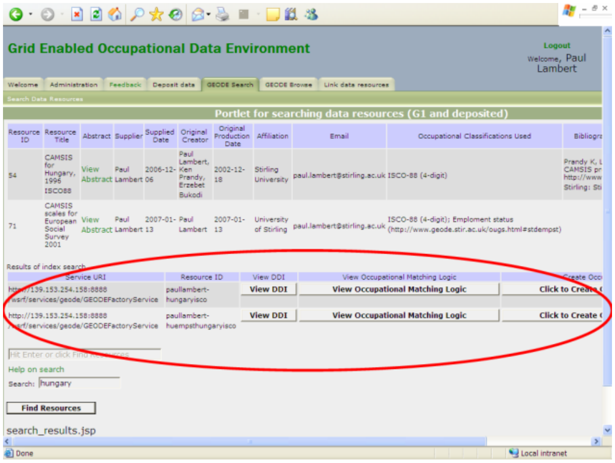
Service URI	Resource ID	View DDI	View Occupational Matching Logic	Create Occ
http://139.153.254.158:8888 /wsrf/services/geode/GEODEFactoryService	paullambert-hungaryisco	View DDI	View Occupational Matching Logic	Click to Create (
http://139.153.254.158:8888 /wsrf/services/geode/GEODEFactoryService	paullambert-huempsthungaryisco	View DDI	View Occupational Matching Logic	Click to Create (

Hit Enter or click Find Resources

Search: hungary

search_results.jsp

Curated resources:


Grid Enabled Occupational Data Environment

Portlet for searching data resources (G1 and deposited)

Resource ID	Resource Title	Abstract	Supplier	Supplied Date	Original Creator	Original Production Date	Affiliation	Email	Occupational Classifications Used	Bibliogra
54	CAMSIS for Hungary, 1996 ISCO88	View Abstract	Paul Lambert	2006-12-06	Paul Lambert, Ken Prandy, Erzebet Bukodi	2002-12-18	Stirling University	paul.lambert@stirling.ac.uk	ISCO-88 (4-digit)	Prandy K, L CAMSIS pr http://www Stirling: Sti
71	CAMSIS scales for European Social Survey 2001	View Abstract	Paul Lambert	2007-01-13	Paul Lambert	2007-01-13	University of Stirling	paul.lambert@stirling.ac.uk	ISCO-88 (4-digit); Emploment status (http://www.geode.stir.ac.uk/ougs.html#stdempst)	

Results of index search

Service URI	Resource ID	View DDI	View Occupational Matching Logic	Create Occ
http://139.153.254.158:8888 /wsrf/services/geode/GEODEFactoryService	paullambert-hungaryisco	View DDI	View Occupational Matching Logic	Click to Create (
http://139.153.254.158:8888 /wsrf/services/geode/GEODEFactoryService	paullambert-huempsthungaryisco	View DDI	View Occupational Matching Logic	Click to Create (

Hit Enter or click Find Resources

Search: hungary

search_results.jsp

- **(B) The rules which generate search results operate differently for the curated and uncurated resources** *[Jun 2007: this situation is likely to be revised in the future]*. At present, uncurated resources are searched using an array of search logic terms which are described on the portal link ‘help for search’. Curated resources however are searched only for exact matches of terms.

The following extract features a focussed description of using the GEODE search portlets, related to a scenario which is described in the GEODE project technical paper 2 (LAMBERT 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.

Interpreting search results:

- 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

Browse Portlet

The ‘browse’ function provides an alternative means to ‘search’ to locate occupational information resources which have been indexed at GEODE.

It organises occupational information resources in groups according to the country, time period, and type of occupational index unit (and combinations thereof).

[At time of writing [11.6.07], the ‘browse’ facility in GEODE only applies to the ‘uncurated’ occupational information resources, and only to those examples of uncurated information for which the appropriate ‘browse’ categorisations were specifically declared by the data depositor].

Comment: Browsing by Countries, Time periods or Occupational Units

The Browse option sorts Occupational Information Resources according to groups of Countries, Time periods and Occupational Units. In our experience, the country classifications are the most useful to the majority of social scientists (it is often useful to rapidly check all uncurated resources associated with a particular country). Browsing by Time period or Occupational Units tends to be more appropriate for specialist interests. For instance, historians are most likely to find the time period search of value (because there are too many resources from recent time periods for this to be a useful index term). Similarly the organisation by occupational units only differentiates between a small number of named international classifications, and a generic group called ‘national specific classifications’ – this browsing facility is thus more likely to be useful to those who are seeking data on cross-national classifications.

4. Linking occupational data resources

The GEODE portal portlet ‘Link data resources’ allows for automated merging of a user’s original data (e.g. social survey micro-data) with relevant occupational information resources on GEODE.

This linkage service is a core provision of GEODE. It is felt that a leading reason for the under-exploitation of existing occupational information resources has been that non-specialist users find it difficult to undertake the data management tasks involved in linking their own data with occupational data resources.

Procedures for implementing the linking process, which involves deploying a JAVA application, are documented on our webpage:

http://www.geode.stir.ac.uk/matching_occupational_data.html

There is also an extended description of linking occupational data with regard to a particular example scenario within the GEODE project technical paper number 2 (LAMBERT 2007).

[At time of writing, 11 June 2007, we are still working on improving the usability of the GEODE occupational matching data service].

5. Depositing occupational data with GEODE

The GEODE portal is improved on every occasion that users index additional occupational information resources with it. The GEODE index service offers coordinated access to occupational information resources from a wide range of countries and time periods, and from a number of different analytical perspectives. Members of the GEODE project themselves actively index all occupational data which they have access to, but other users are keenly invited to index further data files.

The indexing of occupational information files is a two stage process (also described at http://www.geode.stir.ac.uk/geode_m_curation.html).

- Stage 1: Initial supply of only the most crucial data necessary to define occupational information resource
- Stage 2: Further updates to the xml files of any further available metadata

The data required at **stage 1** is quite limited, covering the name of the data file, details on the supplier and context of the study (such as which country and time period it applies to).

To make an initial supply of data, portal users should:

- Login to the portal as a named user
- Enter the 'deposit data' portlet
- Fill out the entries on the data entry form, detailing the website location of the resource, and/or uploading the data resource from the relevant location.

(Note that the data supply process makes a distinction between the supplier of the data to GEODE, and the originator of the data resource, who may or may not be the same people).

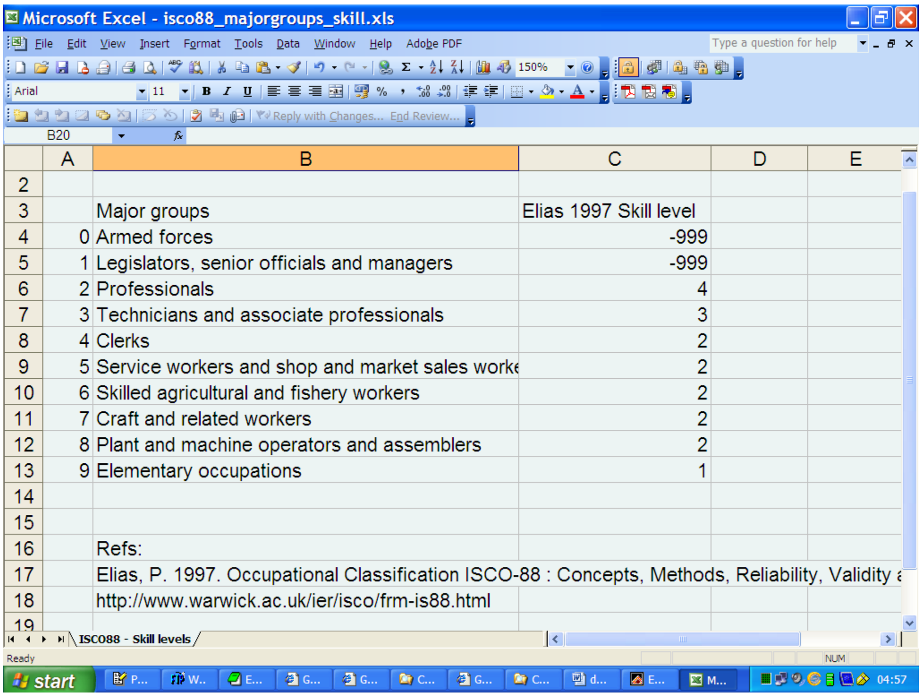
After a user has supplied data in this way, the data is listed on their GEODE account, and they may edit the content of the online record at a later time.

After the data has been exposed, it is now eligible to be curated in **stage 2** of the data supply process. The addition of a fuller set of metadata to the resource is a complex process as there is a great deal of data which may be entered. Illustrations of the relevant metadata are published at http://www.geode.stir.ac.uk/geode_m_curation.html . Reflecting the complex data, the stage 2 data curation is usually only undertaken by members of the GEODE team.

Illustration: Stage 2 data curation

In the following paragraphs we describe the processes of curating occupational information at GEODE in terms of a simple example file.

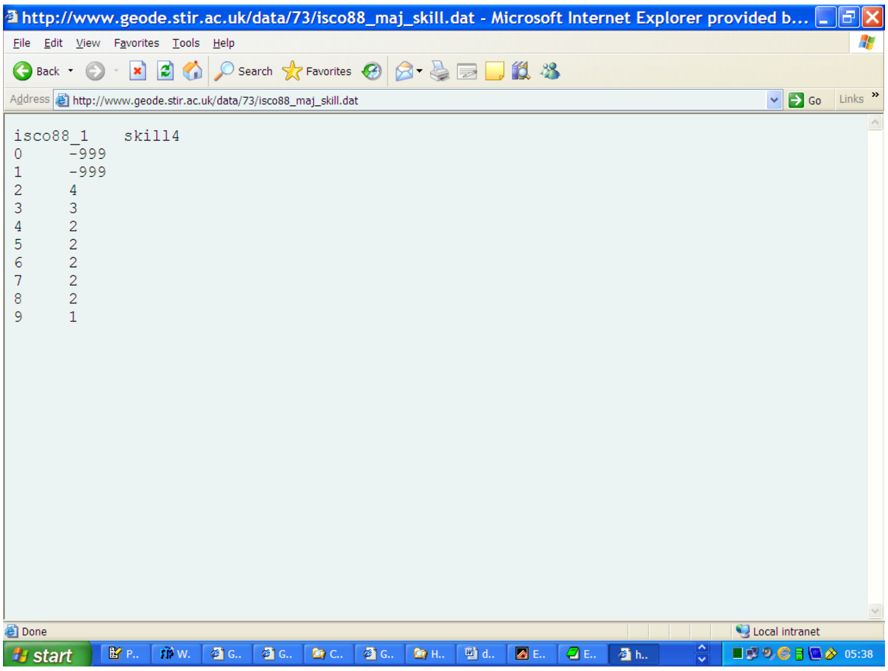
The example we use is a simple tabular data file which describes how the 10 major groups of the ISCO-88 occupational units can be assigned to skill levels. This resource has been deposited to GEODE in an 'uncurated' form in the substance of a single Microsoft Excel file called 'isco88_majorgroups_skill.xls':


Microsoft Excel - isco88_majorgroups_skill.xls
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question for help
Arial 11 150%
Reply with Changes... End Review...
B20
ISCO88 - Skill levels
Ready
NUM
start
04:57

	A	B	C	D	E
2					
3		Major groups	Elias 1997 Skill level		
4	0	Armed forces	-999		
5	1	Legislators, senior officials and managers	-999		
6	2	Professionals	4		
7	3	Technicians and associate professionals	3		
8	4	Clerks	2		
9	5	Service workers and shop and market sales worke	2		
10	6	Skilled agricultural and fishery workers	2		
11	7	Craft and related workers	2		
12	8	Plant and machine operators and assemblers	2		
13	9	Elementary occupations	1		
14					
15					
16		Refs:			
17		Elias, P. 1997. Occupational Classification ISCO-88 : Concepts, Methods, Reliability, Validity a			
18		http://www.warwick.ac.uk/ier/isco/frm-is88.html			
19					

In order for a data resource to be fully curated, to the extent that it may be used in the GEODE occupational matching programme, **it is necessary to create a second (curated) resource associated with the data on the GEODE portal, and to curate three new files associated with the resource.**

The first step involves creating a simple format tab delimited data file representing the data. Instructions on deriving such files are on http://www.geode.stir.ac.uk/file_convert_info.html . This file must be posted on an open access webpage (at any location). For instance, the file associated with the above resource is shown below:

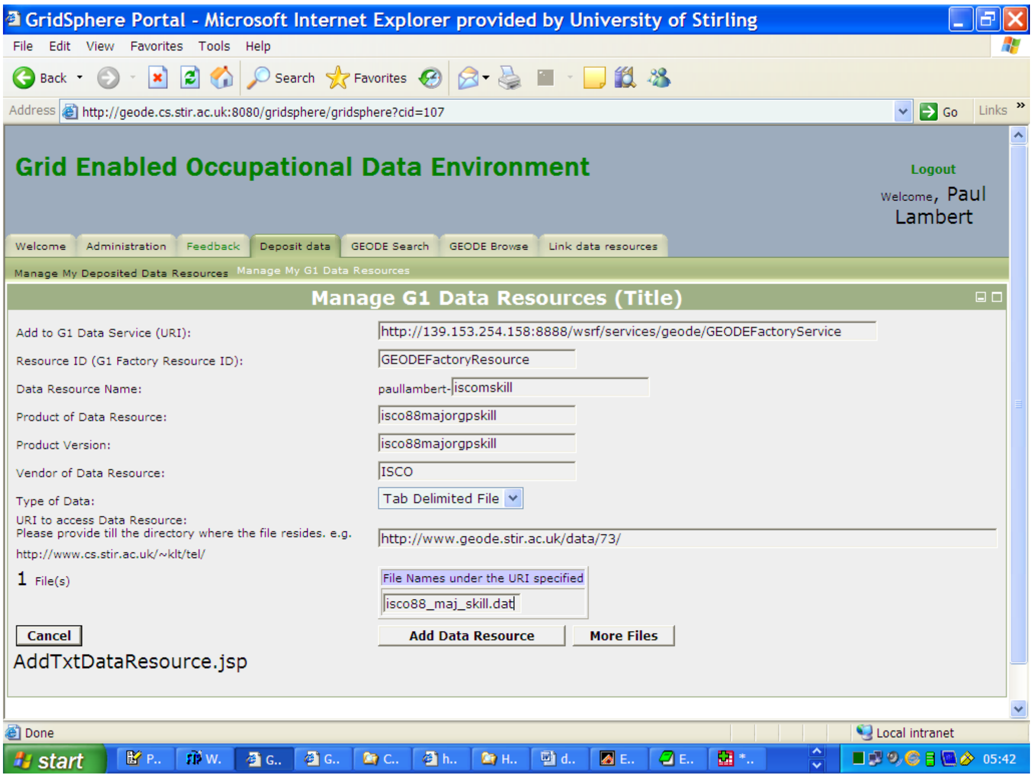

http://www.geode.stir.ac.uk/data/73/isco88_maj_skill.dat - Microsoft Internet Explorer provided b...
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.geode.stir.ac.uk/data/73/isco88_maj_skill.dat
Go Links
Done
Local intranet
start
05:38

```
isco88_1    skill4
0       -999
1       -999
2       4
3       3
4       2
5       2
6       2
7       2
8       2
9       1
```

Once the plain text file has been posted in an appropriate location, it is necessary to create, within GEODE, a resource which describes the location of this new file, and makes the linkage with its metadata documentation. The creation of this resource is done by:

- Clicking on the 'deposit data tab'
- Clicking on 'list resources'
- Clicking 'add new text data resource'
- Filling out the online form, giving a name for the resource and identifying a unique uri at which the resource file in plain text format is located.
- Saving the resource with 'Add data resource'

An image of the online form for the above resource, ready to be added, is shown below:


GridSphere Portal - Microsoft Internet Explorer provided by University of Stirling
File Edit View Favorites Tools Help
Back Search Favorites
Address http://geode.cs.stir.ac.uk:8080/gridsphere/gridsphere?cid=107
Go Links
Logout
Welcome, Paul Lambert
Welcome Administration Feedback Deposit data GEODE Search GEODE Browse Link data resources
Manage My Deposited Data Resources Manage My G1 Data Resources
Done
Local intranet
start P.. W. G.. G.. C.. h.. H.. d.. E.. E.. *..
05:42

Grid Enabled Occupational Data Environment

Manage G1 Data Resources (Title)

Add to G1 Data Service (URI):	http://139.153.254.158:8888/wsrf/services/geode/GEODEFactoryService
Resource ID (G1 Factory Resource ID):	GEODEFactoryResource
Data Resource Name:	paullambert-iscomskill
Product of Data Resource:	isco88majorgpskill
Product Version:	isco88majorgpskill
Vendor of Data Resource:	ISCO
Type of Data:	Tab Delimited File
URI to access Data Resource: Please provide till the directory where the file resides. e.g. http://www.cs.stir.ac.uk/~klt/tel/	http://www.geode.stir.ac.uk/data/73/
1 File(s)	File Names under the URI specified isco88_maj_skill.dat

AddTxtDataResource.jsp

Once this resource has been created – here the resource has been called 'iscomskill' – it is next necessary to populate the resource with data on the occupational information itself. This is done by supplying two xml format information files. In practice it is therefore necessary to create two new files associated with the resource:

1) An xml format file containing metadata on the document

This is the DDI format metadata described above. Instructions with examples on creating DDI metadata are published on http://www.geode.stir.ac.uk/geode_m_curation.html .

Most of those instructions are directed to relatively complex data resources, and, subsequently, the DDI metadata files given as examples tend to be quite long. However, many resources do not need such extensive curation. An example of xml format DDI metadata which would be suitable for the resource described above is downloadable from:

http://www.geode.stir.ac.uk/data/73/ddi_isco88_1_skill.xml

2) An xml format file containing the ‘occupational matching logic’

This ‘logic’ is used to link together the index file with the local (micro-social) data. The content required for the ‘logic’ file can be derived on the basis of information from the DDI metadata. The ‘logic’ file constitutes a short xml file requiring statements which identify the ‘input’ and ‘output’ variable(s) (corresponding to the occupational index unit(s) and the social classification information).

It is possible for GEODE users to view examples of existing ‘occupational matching logic’ xml files by searching ‘G1’ resources and clicking the ‘View occupational logic’ tab. For information, the simple logic required for the above file is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<geodeOccMatchLogic targetNamespace="http://www.geode.stir.ac.uk/wsrf/applications/occ_match_logic.xsd"
xmlns="http://www.geode.stir.ac.uk/wsrf/applications/occ_match_logic.xsd">
  <inputs>
    <var required="true">
      <alias>tar_isco88_1</alias>
      <category>http://www.geode.stir.ac.uk/ougs.html#isco88_1</category>
       <valueType>integer</valueType>
      <defaultValue>0000</defaultValue>
       <req_format>9999</req_format>
      <remedy formula="TRAIL_PAD" value="0"/>
     </var>
 </inputs>

  <matches indexFileTable="isco88_maj_skill.dat">
    <match>
      <varInput>
        <var>
          <alias>tar_isco88_1</alias>
        </var>
      </varInput>
      <varIndex>
        <var>
         <name>isco88_1</name>
        </var>
      </varIndex>
    </match>
  </matches>

  <outputs>
    <varOutput>
      <name>skill4</name>
     <var>
      <name>skill4</name>
      </var>
    </varOutput>
  </outputs>

</geodeOccMatchLogic>
```

The curation of the resource therefore requires the uploading of these two XML files (the DDI documentation, and the occupational matching logic). This is achieved by clicking the links under ‘deposit data’, and the subsequent links ‘Modify DDI’ and ‘Modify Occupational Matching Logic’ which are available under the ‘Manage G1 Data Resources’ portlet.

Once these files are in place, it is possible for users to individually create the linkages which allows for the processing of occupational matching exercises (see section 4 which refers to the webpage instructions http://www.geode.stir.ac.uk/matching_occupational_data.html).

6. Further resources

Readers are encouraged to consult the GEODE webpages for further information:

- Front page: http://www.geode.stir.ac.uk/
- Publications: http://www.geode.stir.ac.uk/publications.html (includes links to text of conference papers authored by GEODE project members)
- File matching notes: http://www.geode.stir.ac.uk/matching_occupational_data.html
- Data curation notes: http://www.geode.stir.ac.uk/geode_m_curation.html
- Listing of occupational index schemes: http://www.geode.stir.ac.uk/ougs.html
- File format conversion notes: http://www.geode.stir.ac.uk/file_convert_info.html

References

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

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.

Lambert, P.S. 2007. *An illustrative guide: Using GEODE to link data from SOC-2000 to NS-SEC and other occupation-based social classifications, Edition 1.1*. Stirling: GEODE Project Technical Paper No. 2, University of Stirling, and http://www.geode.stir.ac.uk.

Lambert, P.S., Tan, K.L.T., Turner, K.J., Gayle, V., Prandy, K. and Sinnott, R.O. 2006. 'Developing a Grid Enabled Occupational Data Environment' *Second International Conference on e-Social Science*. Manchester, and http://www.ncess.ac.uk/events/conference/2006/

Lambert, P.S., Tan, K.L.T., Turner, K.J., Gayle, V., Prandy, K. and Sinnott, R.O. forthcoming 2007. 'Data Curation and Social Science Occupational Information Resouces'. *International Journal of Digital Curation*.

Tan, K.L.T., Gayle, V., Lambert, P.S., Sinnott, R.O. and Turner, K.J. 2006. 'GEODE - Sharing Occupational Data Through the Grid' *5th UK e-Science All Hands Meeting*. Nottingham, and http://www.allhands.org.uk/