

**Grid Enabled Specialist Data Environments:  
Forward Planning for GE\*DE Services for Specialist Data on  
Occupations, Educational Qualifications, and Ethnicity**

**Paul S. Lambert**  
**Vernon Gayle**  
**Koon Leai Larry Tan**  
**Jesse M. Blum**  
**Alison Bowes**  
**Simon Jones**  
**Kenneth J. Turner**  
**Guy Warner**  
**Richard O. Sinnott**  
**Erik Bihagen**

University of Stirling

University of Glasgow  
University of Stockholm

**23<sup>rd</sup> December 2008 [Edition 1.0]**

**DAMES Project Technical Paper 2008-1**

*Technical Papers of the DAMES Node: Data Management through e-Social Science, <http://www.dames.org.uk/publications.html>. DAMES is an ESRC Research Node, Ref: RES-149-25-1066, based at the Universities of Stirling and Glasgow (the National e-Science Centre, [www.nesc.ac.uk](http://www.nesc.ac.uk)). The DAMES Node is a component of the ESRC National Centre for e-Social Science ([www.ncess.ac.uk](http://www.ncess.ac.uk)).*

*Version history:*

*Edition 0.1 distributed internally on 22<sup>nd</sup> October 2008*

*Edition 1.0 published on 23<sup>rd</sup> December 2008*



## Contents:

<b>1. Objectives of DAMES and GE*DE</b> .....	3
<i>DAMES</i> .....	3
<i>GE*DE</i> .....	4
<b>2. GEODE: Expanding resources on occupational information</b> .....	5
<i>Occupational information resources</i> .....	5
<i>Occupational unit groups</i> .....	5
<i>Indexing occupational information resources</i> .....	6
<i>Existing facilities for occupational data</i> .....	9
<i>Existing facilities for occupational data at GEODE</i> .....	11
<i>Plans for further developments of GEODE portal</i> .....	13
<b>3. GEEDE: Establishing resources for data on educational qualifications</b> .....	17
<i>Educational Information Resources</i> .....	17
<i>Recording data on educational qualifications</i> .....	18
<i>Analysing data on educational qualifications</i> .....	18
<i>Envisaged GEEDE services</i> .....	20
<b>4. GEMDE: Establishing resources for data on ethnicity</b> .....	21
<i>Requirements for specialist data on ethnicity</i> .....	21
<b>5. Conclusions</b> .....	23
<b>References</b> .....	25

## 1. Objectives of DAMES and GE\*DE

In this paper we outline the activities and plans associated with the 'GE\*DE' ('Grid Enabled Specialist Data Environments') components of the 'DAMES' research Node. The DAMES Node (Data Management through e-Social Science, [www.dames.org.uk](http://www.dames.org.uk)), based at the Universities of Stirling and Glasgow, is following a three year research programme which commenced on 1<sup>st</sup> February 2008.

### *DAMES*

The DAMES Node covers a number of distinct activities united by the theme of using e-Social Science approaches to support social scientists in undertaking tasks of 'data management'. For the purposes of the DAMES Node, 'data management' is defined as covering tasks, usually performed by social scientists themselves, which involve accessing, manipulating, and linking related data resources, for the purposes of undertaking their own research and analysis<sup>1</sup>. A fuller introduction to the activities of the DAMES Node is available from the project's website.

DAMES is funded as an ESRC research Node comprising part of the ESRC's National Centre for e-Social Science ([www.ncess.ac.uk](http://www.ncess.ac.uk)). NCeSS and DAMES represent major investments by the UK funding councils in developing e-Science facilities for the benefit of social science research. Aspects of e-Science capabilities which are seen as offering contributions to social science research include developments in complex computational calculations on social science data (e.g. CQeSS, 2008); facilities for the collection, storage and display of large volumes of complex social science data (e.g. GeoVUE, 2008; ReDReSS, 2008); standards setting with regard to discovering, describing, and analysing data (e.g. Chorley et al., 2008); and facilities for coordinating secure access to confidential data (e.g. Sinnott et al., 2006). The e-Science contributions associated with DAMES are concentrated around standards setting in terms of using metadata to support access to distributed heterogeneous data resources; encouraging collaborative use of distributed data; and supporting secure access to distributed data.

The objectives of DAMES are, in many ways, prosaic. Few of the data management tasks to which DAMES is oriented cannot already be undertaken by sufficiently skilled social science researchers. However, DAMES is motivated to address a perceived shortfall in UK social science research capacity concerned with making good use of existing data resources. It is argued that few researchers are, in practice, sufficiently confident in undertaking the full range of data manipulation tasks (many of which involve engaging with specialist data resources) which might enhance their analysis. Consequences can be that the process of manipulating data can be unduly slow, bothersome, and motivates researchers to avoid more advanced manipulations and instead resort to the most convenient options. The research projects of the DAMES Node are therefore intended to develop services and resources which will facilitate social science researchers in undertaking data management tasks which

might otherwise have been seen as challenging or prohibitive, or may not even have been thought feasible.

There are already several extant projects and publications which discuss the benefits of integration between related data resources and services (e.g. UK Data Forum, 2007; CESSDA-PPP, 2008; Hagenars, 2008). The DAMES Node seeks to integrate its facilities with other resources related to handling social science data. These include facilities which distribute data resources and associated support and documentation (large scale resources include major data archive services such as the UK Data Archive, [www.data-archive.ac.uk](http://www.data-archive.ac.uk) and the Council of European social science data Archives, [www.CESSDA.org](http://www.CESSDA.org); examples of smaller scale resources are websites maintained by academic researchers for the purpose of disseminating specialist data, such as the SPSS format occupational coding files distributed by Ganzeboom, 2008). There are also external initiatives in research capacity building which are oriented towards the analysis of social science data with areas of provision related to the interests of the DAMES project (such as the UK's Economic and Social Data Service, [www.esds.ac.uk](http://www.esds.ac.uk)).

### ***GE\*DE***

A major component of the DAMES Node is a project theme concerned with 'Grid Enabled Specialist Data Environments', abbreviated as 'GE\*DE'. The asterisk symbol is deliberately used as a 'wildcard' indicator to denote that this theme is concerned with several distinctive specialist data resources. The current Node is concerned with specialist data resources relevant to occupations, educational qualifications, and data about ethnicity. Its services are being designed to allow extensibility to other specialist areas.

The fields of data on occupations, educational qualifications and ethnicity are all areas where there have been significant volumes of previous methodological endeavour. In particular, research on occupations already benefits from numerous existing data facilities (e.g. Leiulfstrud et al., 2005; Eurooccupations, 2008; Ganzeboom, 2008; IISH, 2008). However it does not follow that all social scientists routinely make good use of data on occupations – on the contrary, we have in the past been critical of how empirical researchers typically process data on occupations (Lambert et al., 2007). Thus, much of the effort of GE\*DE is concerned with exploiting, and improving the accessibility of, existing guidance and resources in the field.

The specialist data services related to occupational data began with a project known as GEODE (Grid Enabled Occupational Data Environment, [www.geode.stir.ac.uk](http://www.geode.stir.ac.uk)) which commenced in 2005. Within the DAMES Node, work is planned to expand and improve the services made available through GEODE, and to develop new comparable specialist data services concerned with educational qualifications (Grid Enabled Educational Data Environments, GEEDE); and ethnicity and minority status (Grid Enabled ethnic minority Data Environment, GEMDE).

To sections below lay out the remit of the three GE\*DE services in greater detail, and indicate strategies undertaken and planned within the DAMES programme of research.

## **2. GEODE: Expanding resources on occupational information**

### *Occupational information resources*

‘Occupational information resources’ (OIR’s) can be defined as datasets containing structured records on a set of occupational positions. OIR’s are used in social science research for several purposes, including supplying statistical data about the properties of occupations, and supplying information (‘conversion keys’) to enable the coding of occupational titles into popular ‘occupation-based social classifications’. There are a great many examples of published OIRs, which usually take the form of small electronic files published in a flat format (e.g. Leiufrud et al., 2005; Rose et al., 2007; Eurooccupations, 2008; EUROSTAT, 2008b; Ganzeboom, 2008; Guveli, 2008; ILO, 2008b; Lambert, 2008). The nature, contributions, and documentation of OIR’s was discussed by Lambert et al. (2007).

### *Occupational unit groups*

A critical concept in the definition of occupational information resources is the ‘occupational unit group’. This refers to an externally published listing of categories of occupational positions. Such listings, which are usually but not necessarily numerical in nature, are often referred to as ‘standard indexes’ or, in the terminology of information sciences, ‘standard categories’. Occupational unit groups also have the characteristics of simple taxonomies. They ordinarily identify how any particular occupation fits into one and only one category within a given listing. Locations in the listing are often (but not necessarily) hierarchically organised, such as in the examples of the 1988 and 2008 versions of the International Standard Classification of Occupations, whereby a numerical 4-digit index incorporates 4-digit ‘unit groups’, nested within 3-digit ‘minor groups’, which are nested within 2-digit ‘submajor groups’ which in turn are nested within 1-digit ‘major groups’ (ILO, 1969, 1990, 2008a).

Occupational unit groups comprise listings of categories of occupational titles. These are often published by national and international statistics agencies for the purposes of harmonisation and standardisation in data collection (e.g. ONS, 2007; ILO, 2008a). Statistical agencies’ listings of occupational titles typically define around 300 to 500 different occupational titles. However listings of occupational titles of other orders of magnitude are also common, such as academic studies which define more curtailed taxonomies of occupations and give information on them (e.g. Chan & Goldthorpe, 2004), and listings of job and occupational titles which cover many more units (e.g. Davies et al., 2003; Eurooccupations, 2008).

Tools for coding textual descriptions of occupational titles into well know statistical agency classifications are commonly employed by researchers such as social survey data collectors (e.g. IER, 2008). Accordingly, much social science research data on occupations is available in categories of a well known occupational unit group. In Britain, for instance, most survey micro-data available from the UK Data Archive

contains occupational data coded to the UK Standard Occupational Classification as appropriate for the time period, and/or to one of the ILO's ISCO taxonomies (ILO, 2008a).

Besides listings of occupational titles, other forms of occupational unit groups are also relevant to social science research on occupations. Widely used examples are listings of categories of 'employment status' or employment relations (e.g. Elias, 2000, employment status categorisations typically identify between 2 and 10 distinct categories); and listings of categories of the industrial sector of occupations (e.g. EUROSTAT, 2008a, popular schemes of industry categories may cover between 3 and 500 categories). In many applications, the unique combination of occupational title, occupational employment status and industry is of research interest.

As defined above, the GEODE project provides data on Occupational Information Resources (OIR's). There are therefore navigated in relation to occupational unit groups: an OIR is any dataset which systematically provides information structured around any published occupational unit group taxonomy (or combination of taxonomies). Since there are many different occupational unit group taxonomies (for examples of listings see <http://www.geode.stir.ac.uk/ougs.html>), there are, correspondingly, many thousands of OIR's which are potentially relevant to the GEODE research. Hitherto, a few hundred OIR's have been deposited at or linked with the existing GEODE service, with priority having been given to OIRs which use major British and international occupational unit group taxonomies.

### *Indexing occupational information resources*

Specialist social scientists often undertake work which leads to the generation of new occupational information resources. Non-specialists are typically interested in finding and accessing relevant OIRS to their own research project. Effective mechanisms for undertaking both activities already exist. For instance, specialist data is typically distributed through websites maintained by academic researchers themselves (e.g. Ganzeboom, 2008), or by national statistics agencies (e.g. ONS, 2007). Non-specialists researchers have the opportunity to access suitable data from these sources, often exploiting bespoke guidance documentation associated with the resources.

Nevertheless there are a number of ways in which previous models for the distribution of OIRs may be sub-optimal (see esp. Lambert et al., 2007). The manner in which data resources are distributed tends not to be standardised in terms of software formats, user requirements, or systematic metadata. The extent to which relevant social science research projects actually exploit the full range of occupational information which may be suitable to their analyses is more limited than is desirable (Lambert et al., 2008). The GEODE project is therefore intended to improve the distribution and use of OIR's for social research projects. A number of illustrative scenarios of how GEODE services may contribute in this manner were laid out by Lambert and Tan (2007), and are reproduced below in Table 1.

### Table 1: Selected scenarios in using GEODE

[Source: Lambert and Tan, 2007]

#### 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 'UK1990socsubgpsandlabels1.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 'UK1990socsubgpsandlabels1.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 [ <i>the programme on GEODE to do this is called 'hakimsoc'</i> ]
<b>Usage 4: Supply a data file on occupational positions to GEODE</b>	
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.
<b>Usage 5: Prepare enhanced meta-data on an occupational information file supplied to GEODE</b>	
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. [ <i>This service became available to public users on 8.1.07. Instructions on this are in section 5</i> ]

The GEODE project is therefore oriented to resources which themselves convey systematic information about occupational positions. This marks a subtle distinction between the GEODE project and other projects which have focussed upon the details of individuals' occupational and employment experiences (e.g. Committee on Techniques for the Enhancement of Human Performance: Occupational Analysis, 1999; McGovern et al., 2007; Eurooccupations, 2008; O\*NET, 2008). These latter approaches are oriented towards conveying descriptive information about certain selections of occupational positions, but may not necessarily engage systematically

with a clearly defined occupational unit group taxonomy. GEODE focuses on coordinating data on defined groups of occupational positions, typically for the purposes of statistical analysis across the groups rather than the inspection of data on particular cases.

### *Existing facilities for occupational data*

We have already cited a number of examples of Occupational Information Resources. In particular there are several popular online sources of OIR's, selectively listed below in Table 2. These resources are essentially static locations where databases may be downloaded. There are few standardised protocols governing how data is distributed from such sources, although the RePeC archive does specify a number of formatting and metadata requirements.

Listed within Table 2, the example of the RePeC archive is important. RePeC is a much larger online repository for information resources associated with Economics, which includes within it some relevant Occupational Information Resources. RePeC's hosting of OIRs could in principle offer a framework for coordinating access to occupational data files. However the RePeC model puts the burden of data curation on the academic researcher. Despite citation incentives associated with RePeC, current practices suggest that imposing discipline on data depositors to adhere to the RePeC standard is unlikely to lead to large volumes of OIR's being deposited at RePeC (only a small minority of existing OIR's have been deposited at RePeC).

Aside from the online sources listed in Table 2, there are also numerous examples of relevant occupational information resources being distributed through text based publications. Numerous research papers and books, for example, which include appendices lists tables of data which have the characteristics of OIR's (e.g. Ganzeboom & Treiman, 1996; Hakim, 1998; Chan & Goldthorpe, 2004; Weeden & Grusky, 2005; Oesch, 2006). In certain instances a contribution of the GEODE project is simply to transcribe such resources into a format more suited to online distribution.

<b>Table 2: Selected online Occupational Information Resources</b>	
<a href="http://home.fsw.vu.nl/~ganzeboom/pisa/">http://home.fsw.vu.nl/~ganzeboom/pisa/</a>	Harry Ganzeboom's SPSS format tools for coding ISCO occupational unit groups into occupation-based social classifications (including ISEI, SIOPS and EGP). Users are requested to cite one from Ganzeboom et al. (1992; 1996; 2003)
<a href="http://home.fsw.vu.nl/hbg.ganzeboom/ISMF/ismf.htm">http://home.fsw.vu.nl/hbg.ganzeboom/ISMF/ismf.htm</a>	Harry Ganzeboom's coding information and documentation from the International Social Mobility File, focussing upon data on national standard occupational unit groups over the period 1946-2005.
<a href="http://www.camsis.stir.ac.uk/">http://www.camsis.stir.ac.uk/</a>	CAMSIS project website providing SPSS, Stata and plain text data files linking various occupational unit group schemes with occupation-based social classifications (CAMSIS scale scores and selected other measures)
<a href="http://www.camsis.stir.ac.uk/occunits/distribution.html">http://www.camsis.stir.ac.uk/occunits/distribution.html</a>	CAMSIS project website providing additional data on national occupational units such as category value labels and links to further resources
<a href="http://www2.sofi.su.se/~ebi/">http://www2.sofi.su.se/~ebi/</a>	Popular set of SPSS format conversion keys linking Swedish occupational unit groups with other occupational unit groups and occupation-based social classifications (see e.g. Bihagen & Ohls, 2007).
<a href="http://ideas.repec.org/c/boc/bocode/s425802.html">http://ideas.repec.org/c/boc/bocode/s425802.html</a>	Stata formatting of Harry Ganzeboom's ISCO conversion tools published by John Hendrickx at the REPEC online archive
<a href="http://www.ons.gov.uk/about-statistics/classifications/current/index.html">http://www.ons.gov.uk/about-statistics/classifications/current/index.html</a>	ONS harmonisation resources providing data on value labels for the SOC-2000 occupational unit group; data and value labels on the Standard Industrial Classification occupational unit group scheme; and information on locating SOC2000 categories within the NS-SEC occupation-based social classification.
<a href="http://www.iser.essex.ac.uk/esecc/">http://www.iser.essex.ac.uk/esecc/</a>	ESeC translation matrices: SPSS and MS Excel format codes for converting from ISCO occupational unit groups to the ESeC occupation-based social classification
<a href="https://international.ipums.org/international-action/codes.do?mnemonic=OCC">https://international.ipums.org/international-action/codes.do?mnemonic=OCC</a>	IPUMS dataset summary data on occupational unit groups and their labels from 35 countries
<a href="http://www.ayseguveli.nl/research.aspx">http://www.ayseguveli.nl/research.aspx</a>	SPSS format command files for coding from ISCO occupational unit groups to the occupation-based social classification advocated by Guveli (2006)
<a href="http://historyofwork.iisg.nl/coding.php">http://historyofwork.iisg.nl/coding.php</a>	HISCO project for historical classifications of data, with information on coding from HISCO occupational unit groups to the HISCLASS occupation-based social classification

The growing popularity of Stata (StataCorp, 2008) in many social science disciplines raises one significant issue for the distribution of occupational information. Stata is an accessible, general purpose package for data manipulation and analysis which is well suited to structured documentation of its data procedures (Scott Long, 2008). Of particular relevance to the distribution of social science data, Stata supports a facility to allow the direct reading of data and/or command files from online locations (most other general purpose packages do not share this collaborative feature). Within the field of occupational information resources, several researchers have taken advantage of these features to distribute data files and command files directly online. This facilitates succinct access to relevant OIRs. For instance, in order to exploit a CAMSIS data file from (for instance) the UK in 1991 using the SOC90 unit group scheme, an SPSS user would be required to:

- download a zip file from the CAMSIS project website at <http://www.camsis.stir.ac.uk/versions.html#Britain>;
- extract the component SPSS data file from that archive to a defined directory (e.g. 'c:\temp');
- then run an SPSS command identifying the file and the path to which it has been extracted, such as:  

```
get file= "c:\temp\gb91soc90.sav" .
```

By contrast, to achieve the same in Stata would simply require running the command:

```
use "http://www.camsis.stir.ac.uk/downloads/data/gb91co80.dta" , clear
```

The popularity of Stata and its online facilities thus raise exciting possibilities for data management tasks within the remit of DAMES and GEODE. Nevertheless, in UK social science only a minority of researchers are currently fluent users of Stata (for instance a survey by the Oxford e-Science Institute identified that around 10% of social science data analysts favoured Stata, see [www.oii.ox.ac.uk/microsites/oess/survey/](http://www.oii.ox.ac.uk/microsites/oess/survey/)).

As suggested above, the extensive existing resources for distributing occupational information resources (as listed in Table 2) have some limitations. Of most significant, they are neither as widely used, nor as consistently deployed, as they could in principle be. Poor levels of uptake suggest poor quality scientific research, failing to enhance data resources to their full capacity (c.f. UK Data Forum, 2007). Inconsistencies in the way resources are exploited raises serious challenges for models of replicability, transparency and harmonisation between research projects (cf. Dale, 2006; Freese, 2007).

### ***Existing facilities for occupational data at GEODE***

A number of facilities have already been developed on the GEODE service now incorporated in DAMES (available at [www.geode.stir.ac.uk](http://www.geode.stir.ac.uk)). Most of these facilities were developed in a previous period of funding for this project, between 2005-2007. Broadly, the existing facilities comprise open access online resources, such as

guidance on occupational information and publications, and a 'portal' service which has partially restricted access (see esp. Lambert & Tan, 2007). The portal service, run through Gridsphere, offers facilities for depositing data and/or metadata on occupational information resources, and for users to search through data and metadata to identify and link with suitable occupational information resources.

The portal service is accessible to all using a guest login (which is published on the open access website). Users entering the site on this account have the ability to search existing resources and download materials linked from the search resources. However, only users with a personalised username and password have the facility to deposit data at the GEODE portal. Currently, management of personalised user accounts involves the project staff creating and supplying accounts on personal request. This process has hitherto been easily managed for the small numbers of users who have so far requested personalised accounts.

Uptake of existing GEODE services amongst social science researchers is at present very low. Only handfuls of users have deposited materials at the site, and feedback from those searching and browsing the site has been limited to dozens. Feedback received has often been critical of the accessibility of the existing service. As described in Lambert and Tan (2007), a significant amount of user effort is required to access or deposit data in the existing portal, whilst there are also certain system instabilities, such as the server speed, which can impinge upon use of the portal.

Many of the facilities available through the GEODE portal may appear to be comparable to other information services such as the open access websites listed in Table 2. However, the current GEODE services differ from one-way online provisions in several ways. Firstly, they involve searching facilities on distribution data resources. Second, they allow registered users to upload resources 'live' on the portal site. Third, they automatically assign standard format metadata to all OIRs registered with the service. Fourth, they can be facilitated with potentially dynamic datasets (for instance an OIR could be registered at the GEODE service whilst remaining deposited at an external website; if the content of that data file were later updated, the GEODE service can automatically incorporate the revised file. Lastly, the GEODE service features a unique 'occupational matching service' designed to link a user's secure micro-data with relevant occupational information.

The 'occupational matching service' is a convenience service which offers automatic linking between a user's own micro-data, on the basis of a relevant OUG, and the data stored on a relevant OIR registered with GEODE. Whilst users can achieve these deterministic data linkages through their own means (such as using 'file matching' commands within mainstream statistical data analysis packages such as SPSS and Stata), many social scientists are not comfortable with the relevant programming required to join together separate data files. Therefore the GEODE occupational matching services offers a resource to link OIR data to micro-data in a secure manner. The current implementation works through a java application which links a plain text data file on the user's machine with the relevant OIR, subject to the user identifying the columns in their own data which correspond to the appropriate OUG scheme to which the OIR refers.

### *Plans for further developments of GEODE portal*

During the DAMES Node, developments of the GEODE portal service are being implemented in a two stage process. Firstly, a ‘public’ version of the portal is to be maintained on a purpose built server associated with the ‘e-Infrastructure for the Social Science’ project (NCeSS, 2008) at University of Manchester. This has the attraction of long-term sustainability of the portal and integration with other e-social science resources (and also overcomes some problems associated with server speed and access to server types which arise in the existing pc based portal server run from the University of Stirling). Secondly, developmental versions of the GEODE server are developed and maintained from servers at the University of Stirling.

It is hoped to implement a great many further developments in how the GEODE server appears to users and presents results over the course of the DAMES project, with the objective that within the lifetime of the DAMES Node, the portal will move from its initial ‘proof of concept’ status, with low usability, to a robust, high usage and high quality social science data resource. Table 3 lists currently envisaged and desired updates to the portal. In many instances these emerge from responses to consultations with social science users. The process of ‘service delivery’ associated with implementing these changes and testing the revised services will clearly be a substantial activity within the DAMES Node.

<b>Table 3; Prospective usability extensions to the GEODE portal, 2008-2011</b>	
<b>Version histories</b>	
0.1	Published October 2006, University of Stirling, currently accessible from <a href="http://www.geode.stir.ac.uk">www.geode.stir.ac.uk</a>
1.1	Public version hosted University of Manchester, Summer 2009 <i>(incorporating updates 1-3)</i>
1.2	Developmental version published from University of Stirling, Spring 2009 <i>(incorporating updates 4-19 and others)</i>
2.1	University of Manchester, Public version <i>(a copy of 1.2)</i>
	· · <i>Further updates between developmental and published versions as required</i> · ·
<hr style="border-top: 1px dashed black;"/>	
<b>Envisaged improvements since version 0.1</b>	
1.	Expansion of URI space on OIR resource entry form
2.	Removal of countries with no relevant data from the ‘browse’ list by countries
3.	Improvement of search engine to identify terms throughout data record and to identify ‘fuzzy’ terms (e.g. isco88 cf. ISCO-88)
4.	Automated DDI3 curation.  <i>The current version requires construction of DDI2 data files which is currently conducted manually. For the DDI3 version tools for automating metadata entry</i>

	<i>will be accessed (e.g. SPSS to DDI routines) and tools for making DDI3 manual entry accessible will be prepared and integrated with the portal service.</i>
5.	Organisation of search engine to differentiate outputs between OUGs and other terms (see also comments on linkage with 'ougs.html' (14) below).
6.	Reformatting of search outputs for improved navigation by clearer linking between curated and uncurated results.
7.	<p>Organisation of search outputs according to quality criteria</p> <p><i>At present, results from searches of the GEODE portal are not ordered, but merely list all OIRs with relevant hits in their metadata from the search term used. This leads to difficulty in differentiating between major, popular resources, and less widely used or peripheral resources. One strategy would be restricting the depositing of data to a known group of information providers. However the GEODE project takes a pluralistic approach to providing occupational information and does not seek to censor information contributions. Therefore alternative quality monitoring data is intended to be attached to OIRs and integrated with search results. This will be defined by the number of users (downloads) and matches (invoking of file matching programme) per resource; expert inputs from members of GEODE; and (potentially) live user feedback and ratings. Search results will be sorted in order of ratings results.</i></p>
8.	<p>Definition of types of OIR and organisation of portal and portal searches according to those types</p> <p><i>All OIRs covered by GEODE summarise an input and output datum. The input is, by definition, always an OUG (or a combination of 2 or more OUGs). The output may be another OUG, or some other information about occupations such as locations within an occupation-based social classification or summary descriptive data. At present, searches of GEODE do not differentiate between the types of output data provided by an OIR.</i></p> <p><i>Navigation of the portal is likely to be significantly enhanced if clear pointers to the type of information provided can be made available. Accordingly, we plan that output results will be structured (with a navigable interface) between categories of output types. These categories will be OUG; and other distinctions to be confirmed, likely to be occupation-based social classifications, and other data. An example output would be that a Search on OUG's -&gt; 'Links to other OUG's / 'links to other data'.</i></p> <p><i>For instance, one widely used OIR is the UK ONS's 'crosswalk' from the SOC2000 OUG to the ISCO-88 OUG (this gives data on how SOC2000 categories are linked into ISCO-88 positions). Here, both the input and the output are OUGs. Another widely used OIR is the UK ONS's linkage from the SOC2000 OUG to 'NS-SEC' (this tells users where SOC2000 categories fit within the NS-SEC occupation-based social classification). Here, the input is an OUG but the output is an occupation-based social classification.</i></p>
9.	<p>Navigation of OIRs in terms of 'paths' through OUGs</p> <p><i>A common user desire is to get occupational information from a data resource which is available for a certain OUG which the user does not in fact have direct access to. For instance, the popular OIRs published by Ganzeboom provide 'ISEI' scale scores for ISCO-88 (and ISCO-68) OUGs, but many researchers may wish to attach ISEI scores to their data which is not coded to a suitable ISCO scheme but is coded to a national standard OUG scheme (such as the UK</i></p>

	<p>SOC2000). In many, though not all, instances, a suitable linkage is possible via one or more intermediate occupational information resources. In the example above, a separately published OIR (known as a ‘crosswalk’) can be used to coded SOC2000 categories into suitable ISCO-88 units; and these ISCO-88 units could then be linked with ISEI scores. Such multiple-step ‘paths’ through data are often well know to specialists in occupational data, but would not be obvious to a novice user. Searchers of the existing GEODE database are limited to ‘1-step’ linkages of published OIRs. An improvement to the GEODE portal would therefore be achieved by showing subsequent resources available from multiple step paths through data. This is likely to be invoked in the form of subsidiary linkages from new OUGs generated from original data, which would be facilitated by change (X) above. Quality monitoring of these paths may also be necessary. The may be achieved through workflow modelling of popular paths through OIRs.</p>
10.	<p>Attention to missing data formats in the occupational matching service</p> <p><i>Currently, missing data default codes in SPSS and Stata are problematic when recoded to plain text format data, users having to be instructed to ensure all missing values are given a distinct numeric code.</i></p>
11.	<p>Extension of occupational matching service to allow naming of new variables</p> <p><i>The GEODE ‘occupational matching service’ is described briefly in the text above. If using the service to match data from the OIR onto the users own dataset, one limitation of the current service (v0.1) is that the user has no control over the names given to the newly created variables attached to their dataset. This can be problematic, especially if more than one column of occupational data is to be linked with the OIR (this is a common requirement – for instance the social scientist may want to link occupational data for both the husband’s job and the wife’s job on the same row of their dataset</i></p>
12.	<p>Extension of occupational matching service to automatically link textual OUG titles with numerical OUG indexes whenever available</p> <p><i>This extension arises from user feedback which identifies that data on occupational titles proves a highly effective means for researchers to both understand, and identify errors or inconsistencies, when exploiting OIRs. For instance, if an OIR supplies information on the ISCO-88 category number 2146 has an ISEI score of 71 (see Ganzeboom et al 2003), this information is greatly enriched if the textual description of ISCO-88 category 2146 (‘Chemical engineers’) is also added to the record.</i></p>
13.	<p>Extension of ‘occupational matching’ service to run directly on SPSS and Stata format data</p>
14.	<p>Automated publishing of open access OUG information and direct linkage with GEODE portal resource .</p> <p><i>Currently the website <a href="http://www.geode.stir.ac.uk/ougs.html">http://www.geode.stir.ac.uk/ougs.html</a> lists descriptions of occupational unit groups on which GEODE has relevant information. This page is currently maintained manually, with no interlinking with live updates to the portal resources. This page was originally created solely for the internal use of the GEODE researchers (in order to define unique URI’s for each OUG</i></p>

	<p><i>scheme), but this page has in practice proved a popular open access resource to data on occupational unit group taxonomies.</i></p> <p><i>We propose two contributions to improve information provision. Firstly, new OIRs entering the GEODE system will generate an automatic entry on this website if the OUG to which they refer is not already listed on this page (that entry may subsequently be updated manually with expert input by members of the GEODE project). Secondly, direct linkages between entries on this page and records in the GEODE portal will be developed. If an external user has indentified an OUG within the webpage, a hyperlink should be available which automatically logs them into the GEODE portal (as a guest) and searches the portal for resources corresponding to that OUG. Secondly, any entry deposited to the GEODE portal will necessarily be associated with a location on the OUG's webpage (this linkage was not automatically invoked in the past, but relies on manual entry of the OUGs page).</i></p>
15.	Addition of a Wiki covering FAQs and user experiences of using OIRs within the portal.
16.	<p>Attention to facilities for monitoring (dynamic) external websites.</p> <p><i>This point arises from user feedback. Many resources on the GEODE portal point to external uri's which may themselves change or have their content substantially revised. Mechanisms for monitoring the resources would be desirable.</i></p>
17.	<p>Linkage to direct download of Stata format data (reflecting Stata's advantageous facilities for reading online files, e.g. "use <a href="http://www.dame.org.uk/uk/gb91soc90.dta">http://www.dame.org.uk/uk/gb91soc90.dta</a>")</p>
18.	Allow for multiple files to be uploaded with a single resource.
19.	<p>Allow for an uploaded file to be replaced</p> <p><i>The current mechanism for revisions to files that have been uploaded to the GEODE server ordinarily requires deletion of the original record and re-posting of the revised resource (this is not necessary if the resources is merely pointing to an externally hosted data file, but is required if the resource has been manually uploaded to the GEODE server).</i></p>

### **3. GEEDE: Establishing resources for data on educational qualifications**

#### ***Educational Information Resources***

Social science research data about educational qualifications shares many of the characteristics of data on occupations. First, there is interest in databases on qualification levels, such as descriptive statistics on the profiles of those achieving different qualifications, and the components of qualifications themselves (e.g. OECD, 2008). Second, there is interest in data which can be used to operationalise numerous alternative measures of educational attainment.

From the GE\*DE perspective, the various alternative measures of educational attainment have a similar status to occupation-based social classifications, insofar as they are a major focus of interest for social researchers working with data on educational qualifications, without being the only relevant type of data. Resources oriented to measures of attainment may involve prescriptions on (or academics' own records of) how to collect data in categories of educational qualifications and recode it into measures of educational attainment (e.g. ONS, 2005). Relevant data also often involves information from comparative research on harmonising educational qualifications into 'comparable' measures of educational attainment over time or between countries (e.g. Brauns et al., 2003; Hoffmeyer-Zlotnick & Warner, 2005; Schneider, 2008).

A variety of 'Educational Information Resources' (EIRs) can be identified which have some parallels with the 'Occupational Information Resources' described above. EIRs provide information on educational qualifications, whether it is summary statistical data, or information organised for the particular purpose of deriving measures of educational attainment. The GEEDE service is concerned with providing facilities to organise and distribute such EIR's in a manner which is effective and productive to social science researchers.

However, in comparison with occupational data, relatively few EIR's are currently distributed through electronic files, whereas relatively more EIR's are distributed through textual publication (e.g. Brauns & Steinmann, 1999). Moreover, a critical difference between data on occupations and on educational qualifications is that there is far less standardisation in the way in which records of educational qualifications are made: there are few obvious equivalents to the published 'occupational unit group' schemes around which the GEODE service is primarily oriented.

Furthermore, data on educational qualifications is conventionally considered much more dynamic than data on occupations. This is partly due to large scale transformations in educational sectors across countries, which lead to dramatic differences in the prevalence of qualifications between age cohorts and nations (e.g. Schofer & Meyer, 2005). Instability in educational qualification records is also pronounced due to programmes of institutional reform within countries and, in many instances, regional variations in educational institutions within nations (e.g. Gayle et al., 2009).

### ***Recording data on educational qualifications***

The first major contribution of the GEEDE project is intended to lie in providing resources to support the collection and storage of data on educational qualifications. Hitherto, whilst most data on occupations is recorded in such a way that it can ultimately be linked with an indexing scheme of ‘occupational unit groups’ (see above), the manner in which educational data is recorded is far less consistent.

Studies may record free text descriptions of qualification types. When textual data is collected, however, it is usually coded into a small range of qualification categories. Moreover, most social surveys have tended to collect data within their own set of prescribed categories (taxonomies) of educational qualifications. These taxonomies typically involve between 10 and 40 categories, and are usually distinctive from major categorical measures of educational attainment (although some studies, such as the European Social Survey, code their data directly into measures of educational attainment, see Kolsrud & Skjak, 2005). As a typical example of a UK social survey, the British Household Panel Survey asked questions from a taxonomy of around 20 different qualifications, and these categories can be reasonably, but not perfectly, aligned with the CASMIN measure of educational attainment (Brynin, 2003). After data collection, data on qualifications may be stored within a study dataset in various different manners. A common approach is a ‘multiple response’ format (listing, for each individual, whether or not the individual has each of the different qualifications identified). However many studies preserve data in a mutually exclusive manner (indicating a single qualification held by an individual, which will typically be the individuals ‘highest’ qualification, or their ‘most recent’). Such data collection strategies are typically parsimonious, but may ignore certain qualifications data which is potentially of value for certain analyses.

A major challenge for social scientists in this field is that there has been minimal consistency in the way in which qualification categorisations are stored and coordinated across studies. Within any nation, it is commonly the case that different major social surveys use different qualification taxonomies. Over time, taxonomies are also, typically, substantially updated (in response to institutional reforms). Lastly, no consistent numeric indexing of qualification taxonomies is used across studies. The retrieval of data on qualification taxonomies as recorded on major social surveys is therefore a major priority for the GEEDE resources. The assumed model is therefore that the GEEDE project will establish and develop databases of ‘educational unit groups’ (EUG’s) of a comparable nature to the Occupational Unit Groups described in section 2 above. The organisation and distribution of such standardised will in itself be of significant practical benefit to many social science researchers.

### ***Analysing data on educational qualifications***

The second major contribution of the GEEDE service is intended to lie with services and prescriptive advice concerning the analysis of data about educational qualifications such as in a social survey context. This most conventionally involves facilitating linking from data on educational qualifications (Educational Unit Groups, EUG’s) to meaningful measures of educational attainment (a model with obvious

similarities to the GEODE services for linking between occupational unit groups and occupation-based social classifications). However, there has been far less research on the properties of measures for the analysis of educational attainment than there has in the field of research on occupations, and, correspondingly, a significant contribution of GEEDE will be to explore and develop new analytical solutions for exploiting educational data.

Much progress in formalising data about educational qualifications has been driven by attempts at harmonisation of data on qualifications between countries and/or over time (e.g. Shavit & Blossfeld, 1993; Shavit & Muller, 1998; Hoffmeyer-Zlotnik, 2003a; OECD, 2004; Hoffmeyer-Zlotnick & Warner, 2005; Schneider, 2008). In particular four approaches to measuring concepts of educational attainment have been widely adopted in comparative research.

- 1) One involves attempts to directly measure an individual's total years of schooling. Such measures are widely used within Economics research, though with substantial differences in the criteria employed for calculating the number of years in education (cf. Brynin, 2003).
- 2) In sociological and educational research, most attempts have been to develop classifications of educational attainment which aim to locate different qualification titles into structurally equivalent categories. In this approach, a categorical typology of educational attainment is specified, then, typically, local expertise is used to specify rules over which educational qualifications fit within which categories of attainment. Of these classifications, two widely used and influential measures are the CASMIN scheme, which was first proposed by König et al. (1988) oriented to data from the 1970's, and has been updated with additional categories for the benefit of more contemporary analysis by Brauns and Steinmann (1999) and Kerchoff et al. (2002)); and the the ISCED schemes (see OECD, 2004; Schneider, 2008). Many Educational Information Resources document the coding of national data on qualifications into these schemes (e.g. Ganzeboom & Treiman, 1992; OECD, 2004; Korner & Meyer, 2005). In addition several major cross-national research projects have generated their own project specific measures of educational attainment, and have published educational information resources describing how they coded data from different nations to their bespoke schemes (Kolsrud & Skjak, 2005).
- 3) A less common analytical approach to educational qualifications has been to use summary metrics in order to seek 'functional equivalence' between qualifications over time or between countries (that is, comparability of meaning of the respective qualifications). Several projects, for example, have assigned scale scores to educational qualifications data in an attempt to develop an effective metric of relative educational advantage within the context of an age cohort or national location (e.g. Prandy et al., 2004).
- 4) A last approach to measuring educational attainment is to refer to 'educational achievement' or 'learning achievement' outcomes. This typically involves directly measuring outcomes for individuals (such as mathematics and literacy test scores) which themselves are seen as a function of educational experience

(e.g. Brown et al., 2005). A few individual level social surveys collect data of this nature, though most could only engage with such a concept by deriving a summarising function of educational qualifications in order to approximate probable learning achievement measures.

Alongside such comparative approaches, many national statistics agencies, and other interest groups and stakeholders, also publish recommended schemes for measuring educational qualifications for analytical purposes. Of these, by far the most common approach is similar to approach (2) above, namely, locating qualifications into a pre-specified scheme of attainment categories (e.g. ONS, 2007).

### *Envisaged GEEDE services*

The GEEDE services developed in DAMES will make contributions in the storage of data on educational qualifications, and in support of the analysis of data concerned with concepts of educational attainment. The precise format of the contributions is not yet established, but it is envisaged that a service of a similar nature to the GEODE portal will be used to collect and distribute Educational Information Resources, and corresponding metadata on those resources, and to offer services supporting non-specialists in accessing those resources.

Current prioritised activities are:

- i. Transparent conveying of information on how to operationalise major and influential classifications of qualifications such as ISCED
- ii. Listing of educational coding data collected on major UK social surveys with accompanying information on effective variable operationalisations for them
- iii. The development of an electronic databank of educational qualifications and their location with major measures of educational classifications
- iv. The promotion of existing, and the development of new, scaling measures of the relative advantage associated with educational qualifications. These will include new analyses of data on educational qualifications for the purpose of scoring educational qualification categories (such as, for example, the calculation of mean occupational advantage scores held by adults with the relevant educational qualifications)
- v. The promotion of guidance and resources to support longitudinal analyses of educational qualifications data in the context of significant cohort change in the prevalence of qualification types, such as through offering treatment of educational qualification data conditional upon the year of birth of the individuals concerned

#### **4. GEMDE: Establishing resources for data on ethnicity**

##### ***Requirements for specialist data on ethnicity***

The coordination of specialist data resources around the topic of ethnicity is significantly more complicated than in the case of data on occupational and educational qualifications. Despite its considerable interest to social scientists from a range of backgrounds, there has been relatively little systematic methodological research on the collection and analysis of data on ethnicity, and accordingly there are relatively few specialist information resources on ethnicity (at least if compared with the domains of occupational data and data on education). Moreover, the concept of ethnicity itself is highly contested within sociological research, and is inconsistently related to a range diverse 'referents' (that is, underlying concepts) according to different national traditions and political preferences (Hoffmeyer-Zlotnik, 2003b). Examples of ethnic referents commonly used in social science research include measures of subjective ethnic identity; nationality; country of birth; parents' nationality and/or country of birth; language used; religion; time since migration or parents' migration; and somantic differences. Data concerning any and all of these measures is considered to fall within the remit of the 'GEMDE' service for specialist data on ethnicity.

Nevertheless, in one interpretation, the use of data on ethnicity may appear straightforward. In many nations, national statistics agencies have taken steps to define standardised recommended measures of ethnic groups using one or more of the ethnic referents mentioned above. For example, in the UK, recommended classifications of subjective ethnic group are widely employed in major social surveys and other data collection instruments (ONS, 2007).

However, official measures prove limiting for numerous research projects. First, within particular nations, arguments highlighting the conceptual and analytical weaknesses of the official measures of ethnicity are easily identified (for UK examples, see Ballard, 1997; Ahmad, 1999; Smith, 2002). Second, for many social science datasets, the representation of subjects from many minority categories may be too sparse for convenience, and alternative bespoke categorisations may be more appropriate (for instance, Lambert, 2005 noted that although records from the European Social Survey contained detailed measures of ethnicity, the representation of minority categories was too sparse to support effective analysis in almost all countries). Third, within particular nations, the population of individuals from ethnic minority backgrounds is typically highly dynamic and highly concentrated. Minority populations are typically characterised by major cohorts of immigration in concentrated periods to concentrated regions. Subsequent to the immigration period, these concentrations disperse relatively rapidly, with particular measurement challenges raised by ethnic exogamy and 'mixed' ethnic groups. Such demographic patterns may mean that most suitable categorical distinctions can change significantly over time. A common response is to update official classifications periodically, though this brings continuity problems to the analysis of data from different time points (e.g. Platt et al., 2005; Bosveld et al., 2006). The dynamic nature of ethnic minority populations also raises more challenging analytical questions about the distinction between the effects of ethnicity, and of demographic differences concerned

with age, gender and regional concentration. Fourth and finally, comparative analysis across regions or countries becomes problematic when different measures are recorded in different regions (Hoffmeyer-Zlotnik, 2003b). Several common comparative strategies can be identified, typically either attempting to define equivalent categorisations across countries, and/or to construct scales indicating relative locations of minority categories according to other criteria, but all are to some degree sub-optimal (see Lambert, 2005 for a review).

Accordingly, contributions towards data management for specialist data on ethnicity may require quite different approaches to those applied towards data on occupations and ethnicity. The organisation of information around the plurality of different ethnic referents may require significant specialist data input. Given the relative absence of existing resources, greater manual efforts in the curation and/or construction of specialist data may be necessary to develop resources in this exciting but under-resourced field. Given the problems of comparative research on ethnicity, relatively more scientific effort may be required to develop plausible instruments for comparative research, such as data intensive research on the experiences of subjects across time or between countries. Lastly, given the sparsity of data on many minority groups, there may also be a need for services giving direct access to large scale data resources to supplement and enhance the sparse data that many researchers are likely to have initial access to.

### ***Envisaged GEMDE services***

Preliminary plans for the resources proposed for the GEMDE service are:

- i. Portal access to specialist data resources. *Following a comparable model to the portal facilities of the GEODE and GEDE services, facilities for storing specialist datasets on ethnic minority groups will be developed and made accessible to the research community. In particular, this effort will result in the generation of new metadata on the categories and measures associated with social science research on ethnic referents.*
- ii. Development of recommendations for tools for the analysis of ethnicity in comparative contexts. *In practical terms, this will involve research generating new measures such as categorisations and scales suitable for understanding ethnic differences in a cross-nationally and/or longitudinally comparative context.*
- iii. Development of services to directly interrogate large scale microdata. *The model here will involve resources to allow researchers to request results from a live analysis of major data resources with records on ethnicity.*
- iv. Integrated analysis with the context of age, gender and region. *An exciting scientific opportunity is presented in the potential development of data management resources which more clearly highlight the collinearity between many ethnic minority groups and other demographic and regional locations. In the UK and other social science research, few analysts offer satisfactory treatments for these collinearities, but relatively simple data manipulation activities (e.g. case selection by demographic characteristics) and analytical techniques (e.g. propensity score matching) could be deployed through GEMDE services to facilitate more appropriate analysis.*

## 5. Conclusions

In laying out our plans on the GE\*DE services associated with the DAMES Node as above, we conclude by highlighting three broad themes in the data management provisions proposed, and three significant challenges.

The first theme concerns the centrality of metadata to the services provided. In all services standardised metadata (using the DDI 3 format, see Vardigan et al., 2008) will be used to describe relevant specialist data resources and allow for linking between distributed resources. Tools and services for assigning suitable metadata are in turn a central requirement on the DAMES Node.

These comments lead directly to a second significant theme within the provisions described above, namely the characterisation of tasks in variable operationalisations and data linking as ‘workflows’ of data management activities. Indeed, Scott Long (2008) conflates the term workflow with the syntactical documentation of data manipulation and analysis commands in his review of the Stata package. However, the term workflow has a wider meaning than simply documentation, and procedures for modelling workflows are currently the subject of several e-Science research programmes alongside the DAMES Node. Within the GE\*DE programme of research, there are clear potential benefits to documenting workflows in variable operationalisations and linking data associated with the specialist data topics outlined above.

The third theme concerns the challenges of service delivery. The community of social science researchers is diverse. Many researchers have high levels of fluency in various software environments and in undertaking data manipulation tasks (cf. Levesque & SPSS Inc, 2008; Scott Long, 2008). Equally however, many researchers have much less confidence in using such packages and manipulating datasets. In the GE\*DE we will seek to provide resources of use to researchers across the range of relevant expertise, which will require considerable efforts in communication of usage instructions at different levels of details.

The fourth theme concerns attention to the ‘functional form’ of analytical measures in these specialist areas. The research programmes described above are unified by offering provisions for plurality of alternative measures for related concepts (such as several different alternative occupation-based social classifications). This plurality, which is hitherto uncommon in all three of the specialist areas, opens doors to alternative arithmetic treatments (known as using different ‘functional forms’) of comparable measures. These different assumptions may in turn have significant implications for the analytical methods used and the complexity of subsequent analyses.

The first challenge concerns the need for services to maintain high standards of documentation and replicability of data management activities supported. In social science survey data analysis, documentation is traditionally achieved through syntactical records of data management and data analysis tasks in the form of the command languages of popular data analysis packages such as SPSS, Stata, SAS and

R. In the development of the DAMES services, we treat as a high priority an equivalent or comparable level of syntactical record of activities for the benefit of future replicability. Dealing with online resources raises challenges, since many externally designed services exploit 'point-and-click' interfaces which lack comparable documentation services. A major challenge for the DAMES Node's services is therefore to maintain a human readable documented path describing activities undertaken.

A second challenge concerns coordination of services developed within DAMES with other initiatives in the field. These include both social science data projects, and e-Science projects developing comparable services in other fields. Clearly, different projects work on different timescales and a top-down model of coordination is unlikely ever to be completely feasible (although in a few significant instances, particularly concerning the harmonisation of GE\*DE services with the CESSDA-PPP initiative, timetables do allow some degree of inter-dependent activities). Instead, coordination is typically planned by using common standards and approaches whenever feasible, and maximising communication with other programmes and the documentation of the development of facilities.

A third challenge concerns incentivising good practice in data management activities amongst the social researchers who may benefit from the GE\*DE provisions. There have hitherto been few pay-offs to social researchers to deposit and share their specialist data (though reputational benefits associated with being the author of widely used resources may be significant in some instances). In current plans, our approach will involve efforts to obtain (and subsequently promote) citable references for all deposited resources; and efforts to make the entry of suitable metadata on resources as easy as is feasible. There have also hitherto been few motivations amongst users of data to exploit anything other than the most convenient resources for subsequent analyses (for instance, a sociologist interested in using an occupation-based social classification would most commonly employ the first classification they both easily have access to, and think that they understand). Yet, during the process of low-level research with the data resources provided in each field, methodological insights will be gained such as concerning the relative benefits of different measures of educational attainment. Such insights allow the GE\*DE projects to be in a position to offer prescriptive advice on data management practices which might well contradict the original plans of users of the data. Encouraging social scientists to follow recommended advice is not easy, but may be facilitated in GE\*DE services by foregrounding recommended methods, and by attempting to demonstrate the scientific benefits embodied by them.

## References

- Ahmad, W. (1999). Ethnic Statistics : Better than nothing or worse than nothing? In D. Dorling & S. Simpson (Eds.), *Statistics in Society* (pp. 124-131). London: Arnold.
- Ballard, R. (1997). The construction of a conceptual vision : Ethnic groups and the 1991 UK census. *Ethnic and Racial Studies*, 20(1), 182.
- Bihagen, E., & Ohls, M. (2007). Are Women Over-represented in Dead-End Jobs: A Swedish study using empirically derived measures of dead-end jobs. *Social Indicators Research*, 84(2), 159-177.
- Bosveld, K., Connolly, H., Rendall, M. S., & (2006). *A guide to comparing 1991 and 2001 Census ethnic group data*. London: Office for National Statistics.
- Brauns, H., Scherer, S., & Steinmann, S. (2003). The CASMIN Educational Classification in International Comparative Research. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 221-244). New York: Kluwer Academic.
- Brauns, H., & Steinmann, S. (1999). The CASMIN Educational Classification in International Comparative Research. *ZUMA-Nachrichten*, 44(23), 7-44.
- Brown, G., Micklewright, J., Schnepf, S., & Waldmann, R. (2005). *Cross-National Surveys of Learning Achievement: How Robust are the Findings?* Southampton: S3RI Applications and Policy Working Paper, Southampton Statistical Sciences Research Institute, University of Southampton, and <http://eprints.soton.ac.uk/16250/01/s3ri-workingpaper-a05-05.pdf>
- Brynin, M. (2003). Using CASMIN: The Effect of Education on Wages in Britain and Germany. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 327-344). New York: Kluwer Academic.
- CESSDA-PPP. (2008). Preparatory Phase Project for a Major Upgrade of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure. Retrieved 1 December 2008, from <http://www.cessda.org/project/>
- Chan, T. W., & Goldthorpe, J. H. (2004). Is There a Status Order in Contemporary British Society. *European Sociological Review*, 20(5), 383-401.
- Chorley, A., Edwards, P., Hielkema, F., Phillip, L., & Farrington, J. (2008). *Developing Ontologies to Support eSocial Science: The PolicyGrid Experience*. Paper presented at the 4th International Conference on e-Social Science.
- Committee on Techniques for the Enhancement of Human Performance: Occupational Analysis. (1999). *National Research Council: The Changing Nature of Work: Implications for Occupational Analysis*. Washington: National Academy Press.
- CQeSS. (2008). Collaboratory for Quantitative e-Social Science. Retrieved 1 October, 2008, from <http://e-science.lancs.ac.uk/cqess/>
- Dale, A. (2006). Quality Issues with Survey Research. *International Journal of Social Research Methodology*, 9(2), 143-158.
- Davies, R., Elias, P., & Ellison, R. (2003). Standard Occupational Classification for the Destinations of Leavers from Higher Education Institutions: SOC(DLHE). Retrieved 1 December 2008, from [http://www.hesa.ac.uk/dox/informationProvision/5\\_digit\\_dhle\\_cats.pdf](http://www.hesa.ac.uk/dox/informationProvision/5_digit_dhle_cats.pdf)
- Elias, P. (2000). Status in Employment: A World Survey of Practices and Problems. *Bulletin of Labour Statistics*, 1-19.
- Eurooccupations. (2008). Eurooccupations Research Lab. Retrieved 1 December 2008, from <http://www.eurooccupations.org/main/researchlab>

- EUROSTAT. (2008a). NACE Rev. 2: Statistical Classification of Economic Activities in the European Community, Rev. 2. Retrieved 1 December 2008, from [http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP\\_PUB\\_WELC](http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC)
- EUROSTAT. (2008b). RAMON - Correspondence Table List. Retrieved 1 December 2008, from [http://ec.europa.eu/eurostat/ramon/rerelations/index.cfm?TargetUrl=LST\\_REL](http://ec.europa.eu/eurostat/ramon/rerelations/index.cfm?TargetUrl=LST_REL)
- Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods and Research*, 36(2), 2007.
- Ganzeboom, H. B. G. (2008). Tools for deriving status measures from ISKO-88 and ISCO-68. Retrieved 1 March, 2008, from <http://home.fsw.vu.nl/~ganzeboom/PISA/>
- Ganzeboom, H. B. G., & Treiman, D. J. (1992). International Stratification and Mobility File: Conversion Tools (Version 92-08-25). Utrecht: Department of Sociologie.
- Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research*, 25(3), 201-235.
- Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised 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.
- Gayle, V., Lambert, P. S., & Murray, S. (2009). School-to-Work in the 1990s: Modelling transitions with large-scale datasets. In R. Brook (Ed.), *Transitions from Education to Work: New Perspectives from Europe and Beyond*. Basingstoke: Palgrave MacMillan.
- GeoVUE. (2008). GeoVUE: Geographic Virtual Urban Environments. Retrieved 1 October 2008, from <http://www.casa.ucl.ac.uk/projects/projectDetail.asp?ID=57>
- Guveli, A. (2006). *New Social Classes within the Service Class in the Netherlands and Britain: Adjusting the EGP class schema for the technocrats and the social and cultural specialists*. Nijmegen: Radboud University Nijmegen.
- Guveli, A. (2008). Conversion tools for the adjusted EGP class schema. Retrieved 1 December 2008, from <http://www.ayseguveli.nl/research.aspx>
- Hagenaars, J. A. (2008). COMPSOC: Exploiting, documenting and enriching comparative data from large-scale surveys in the social sciences. Tilburg University: Investeringen NWO-middelgroot.
- 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.
- Hoffmeyer-Zlotnick, J. H. P., & Warner, U. (2005). How to Measure Education in Cross-National Comparison: Hoffmeyer-Zlotnick/Warner -Matrix of Education as New Instrument. In J. H. P. Hoffmeyer-Zlotnick & J. Harkness (Eds.), *Methodological Aspects in Cross-National Research*. Mannheim: GESIS - ZUMA Zentrum für Umfragen, Methoden und Analysen (ZUMA Nachrichten Spezial, Vol 11).
- Hoffmeyer-Zlotnik, J. H. P. (2003a). The Classification of Education as a Sociological Background Characteristic. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 245-256). New York: Kluwer Academic.
- Hoffmeyer-Zlotnik, J. H. P. (2003b). How to Measure Race and Ethnicity. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 267-277). New York: Kluwer Academic.
- IER. (2008). CASCOT: Computer Assisted Structured Coding Tool. Retrieved 1 December 2008, from <http://www2.warwick.ac.uk/fac/soc/ier/publications/software/cascot/>
- IISH. (2008). History of Work Information System. Retrieved 1 December 2008, from <http://historyofwork.iisg.nl/>

- ILO. (1969). *International Standard Classification of Occupations : Revised Edition 1968*. New York: International Labour Office.
- ILO. (1990). *ISCO-88 : International Standard Classification of Occupations*. New York: International Labour Office.
- ILO. (2008a). ISCO: International Standard Classification of Occupations. Retrieved 1 December 2008, from <http://www.ilo.org/public/english/bureau/stat/isco/>
- ILO. (2008b). LABORSTA Internet: An International Labour Office database on labour statistics. Retrieved 1 December 2008, from <http://laborsta.ilo.org/>
- Kerckhoff, A. C., Ezell, E. D., & Brown, J. S. (2002). Toward an Improved Measure of Educational Attainment in Social Stratification Research. *Social Science Research*, 31(1), 99-123.
- Kolsrud, K., & Skjak, K. K. (2005). Harmonising Background Variables in the European Social Survey. In J. H. P. Hoffmeyer-Zlotnick & J. Harkness (Eds.), *Methodological Aspects in Cross-National Research*. Mannheim: GESIS - ZUMA Zentrum für Umfragen, Methoden und Analysen (ZUMA Nachrichten Spezial, Vol 11).
- König, W., Luttinger, P., & Müller, W. (1988). *A Comparative Analysis of the Development and Structure of Educational Systems. Methodological Foundations and the Construction of a Comparative Educational Scale*. Mannheim: University of Mannheim, CASMIN Working Paper no. 12.
- Korner, T., & Meyer, I. (2005). Harmonising Socio-Demographic Information in Household Surveys of Official Statistics: Experiences of the Federal Statistical Office Germany. In J. H. P. Hoffmeyer-Zlotnick & J. Harkness (Eds.), *Methodological Aspects in Cross-National Research*. Mannheim: GESIS - ZUMA Zentrum für Umfragen, Methoden und Analysen (ZUMA Nachrichten Spezial, Vol 11).
- Lambert, P. S. (2005). Ethnicity and the Comparative Analysis of Contemporary Survey Data. In J. H. P. Hoffmeyer-Zlotnick & J. Harkness (Eds.), *Methodological Aspects in Cross-National Research* (pp. 259-277). Mannheim: ZUMA-Nachrichten Spezial 11.
- Lambert, P. S. (2008). CAMSIS project: Files for distribution covering occupational unit codes and translations. Retrieved 22 December 2008, from <http://www.camsis.stir.ac.uk/occunits/distribution.html>
- Lambert, P. S., & Tan, K. L. L. (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. L., Gayle, V., Prandy, K., & Bergman, M. M. (2008). The importance of specificity in occupation-based social classifications. *International Journal of Sociology and Social Policy*, 28(5/6), 179-192.
- Lambert, P. S., Tan, K. L. L., Turner, K. J., Gayle, V., Prandy, K., & Sinnott, R. O. (2007). Data Curation Standards and Social Science Occupational Information Resources. *International Journal of Digital Curation*, 2(1), 73-91.
- Leiulfsrud, H., Bison, I., & Jensberg, H. (2005). *Social Class in Europe: European Social Survey 2002/3*. Trondheim: NTNU Samfunnsforskning/NTNU Social Research Ltd.
- Levesque, R., & SPSS Inc. (2008). *Programming and Data Management for SPSS Statistics 17.0*. Chicago, IL: SPSS Inc. .
- McGovern, P., Hill, S., Mills, C., & White, M. (2007). *Market, Class and Employment*. Oxford: Oxford University Press.
- NCeSS. (2008). e-Infrastructure for the Social Sciences. Retrieved 1 December 2008, from <http://www.ncess.ac.uk/research/einfrastructure/>
- O\*NET. (2008). O\*NET Online: Occupational Information Network. Retrieved 1 December 2008, from <http://online.onetcenter.org/>
- OECD. (2004). *OECD Handbook for Internationally Comparative Educational Statistics*. Paris: Organisation for Economic Cooperation and Development.
- OECD. (2008). SourceOECD Education Statistics. Retrieved 1 October 2008, from <http://lysander.sourceoecd.org/>
- Oesch, D. (2006). *Redrawing the Class Map: Stratification and Institutions in Britain, German, Sweden and Switzerland*. Basingstoke: Palgrave.

- ONS. (2005). *Harmonised Concepts and Questions for Social Data Sources: Primary Standards: Other Primary Standards, Version 3.0*. London: Office for National Statistics.
- ONS. (2007). National Statistics Harmonisation. Retrieved 1 June, 2007, from <http://www.statistics.gov.uk/about/data/harmonisation/>
- Platt, L., Simpson, L., & Akinwale, B. (2005). Stability and change in ethnic groups in England and Wales. *Population Trends*, 121, 35-46.
- Prandy, K., Unt, M., & Lambert, P. S. (2004). *Not by degrees: Education and social reproduction in twentieth-century Britain*. Paper presented at the ISA RC28 Research Committee on Social Stratification and Mobility.
- ReDReSS. (2008). Resource Discovery for Researchers in e-Social Science. Retrieved 1 October 2008, from <http://redress.lancs.ac.uk/>
- RELU. (2006). *Guidance on Data Management*. Colchester: Economic and Social Data Service, Rural Economy and Land Use Programme Data Support Service.
- Rose, D., Harrison, E., & Pevalin, D. (2007). ESEC : A European socio-economic classification. Retrieved 1 June, 2007, from <http://www.iser.essex.ac.uk/esec/>
- Schneider, S. L. (2008). *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*. Mannheim: MZES.
- Schofer, E., & Meyer, J. W. (2005). The World-Wide Expansion of Higher Education in the Twentieth Century. *American Sociological Review*, 70(6), 898-920.
- Scott Long, J. (2008). *The Workflow of Data Analysis Using Stata*. Boca Raton: CRC Press.
- Shavit, Y., & Blossfeld, H. P. (Eds.). (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Boulder, CO.: Westview.
- Shavit, Y., & Muller, W. (Eds.). (1998). *From School to Work*. Oxford: Clarendon Press.
- Sinnott, R. O., Watt, J., Ajayi, O., & Jiang, J. (2006). *Shibboleth-based Access to and Usage of Grid Resources* Paper presented at the IEEE International Conference on Grid Computing, Barcelona, Spain, 28-29 September 2006.
- Smith, K. (2002). Some critical observations on the use of the concept of 'ethnicity' in Modood et al., *Ethnic Minorities in Britain*. *Sociology*, 36(2), 399-417.
- StataCorp. (2008). *Stata Statistical Software, Release 10*. College Station, TX: StataCorp LP.
- UK Data Forum. (2007). *The National Strategy for Data Resources for Research in the Social Sciences*. Warwick: University of Warwick, <http://www2.warwick.ac.uk/fac/soc/nds/> (Accessed 18 June 2007).
- Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Towards a Standard for the Social Sciences. *International Journal of Digital Curation*, 3(1), 107-113.
- Weeden, K. A., & Grusky, D. B. (2005). The Case for a New Class Map. *American Journal of Sociology*, 111(1), 141-212.

---

<sup>1</sup> The data management tasks to which the DAMES Node is oriented may be distinguished from some other uses of the term 'data management', which are concerned with archiving, distributing and monitoring research materials. See for example the ESDS guidelines RELU (2006)