NDS
THE NATIONAL
DATA STRATEGY

Administrative data as a research resource: a selected audit

Paul Jones and Peter Elias

December 2006

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

WARWICK INSTITUTE *for*
EMPLOYMENT RESEARCH
IER 1981- 2006
Celebrating 25 years

Table of Contents

Annexes

Tables

Figures

Acknowledgements

This report was undertaken on the recommendation of the UK Data Forum, a multi-agency body which monitors the work being undertaken to implement the National Strategy for Data Resources for Research in the Social Sciences (the 'National Data Strategy'). Within this strategy, administrative data resources are seen as important and potentially useful resources, yet there are obstacles to their use, relating particularly to the legislation through which they are obtained, ethical considerations and the need to preserve the confidentiality of individuals and organisations. To inform future activities to develop the potential of such resources for research purposes, a subgroup of the UK Data Forum was formed to establish a work programme and to monitor progress. The authors are particularly grateful to members of the subgroup; Professor John Fox (NHS Information Centre for Health and Social Care), Mr Mike Hughes (Office for National Statistics), Mr Ricky Taylor (Cabinet Office) and Ms Sharon Jones (Department for Work and Pensions) for their guidance and their help in facilitating the collection of information presented in this report.

The authors would also like to thanks all those who participated in the project. Particular thanks are given to those who took part in project-related interviews or meetings, and to those who supplied relevant information about the datasets covered in the report. A list of contributors is shown in the final annex of the report.

1. Introduction

The term 'administrative data' describes information which arises via the operation of a transaction, registration or as a record of service delivery. Such data relates specifically to the administration of a system or process and are not primarily generated as research resources.

While administrative data are not necessarily the preserve of government, most government departments keep records of the variety of services they deliver and the processes they register in considerable detail, often storing this information as electronic records that relate to individuals and/or organisations and summarising these data for statistical purposes.

Administrative data have significant potential as research resources, either in their own right or via linkage to other sources of information (*e.g.* censuses or surveys). This potential has been recognised in a series of recent reports which have also raised a number of important issues relating to the need to preserve their confidentiality, concerns about the privacy of individuals and the legality of sharing such information between government departments and agencies[1].

This audit has been undertaken to provide an up-to-date picture of the variety of information currently available for research purposes and the nature of the research that has been conducted using such resources. The aim is to guide and inform the National Data Strategy – a plan to ensure that the data resources required to inform future research issues are in place on time.

The scale of this review is potentially vast. For this reason, a selective approach has been taken, focussing upon a number of areas of potential interest for social scientific and related research interests and examining within these areas both the types of administrative data available and their potential as research resources. Five areas have been selected. These are:

- Education-related datasets;
- Labour market-related datasets;
- Health-related datasets;
- Business-related datasets;
- Demographic datasets.

The review has been undertaken within a relatively narrow timescale and with minimal resources. It has relied heavily upon information provided by a number of key

[1] See, for example:
Better use of personal information: opportunities and risks (Council for Science and Technology, November 2005)
(http://www.cst.gov.uk/cst/reports/files/personal-information/report.pdf)
Privacy and data-sharing: The way forward for public services (Performance and Innovation Unit, Cabinet Office, April 2002)
(http://www.strategy.gov.uk/downloads/su/privacy/downloads/piu-data.pdf)
Data Sharing for Statistical Purposes: A Practitioners' Guide to the Legal Framework (Office for National Statistics, Sep. 2005)
(http://www.statistics.gov.uk/downloads/theme_other/NSDataSharing.pdf)

informants, to whom we are most grateful for the information they have provided. It is undoubtedly incomplete, in that there exist new resources under development not included in this review, and many local and regional sources which remain unidentified in this report. Nonetheless, the intention is to demonstrate both the richness of these resources, their complexity and the variety of mechanisms that have been established to facilitate access to such data.

The report is structured in five main sections (Chapters 2-6), relating to the areas selected for study. Each chapter describes in detail the nature of some of the larger administrative data sources that are available for research purposes and lists selected research projects and papers which have made use of these data.

Chapter 7 presents an overview of issues in data access and sharing, covering the costs and benefits, the legal barriers to data sharing and issues relating to the risks associated with increased research usage of such data.

Chapter 8 attempts to summarise the 'state-of-the-art' in the areas selected for this audit and suggests how the social science community might move to build upon these developments in a manner consistent with the main aim of the National Data Strategy – to ensure that UK social science retains its leading edge position in the work by promoting access to such data resources in a safe, legal, efficient and timely manner.

2 Education related datasets

2.1 Introduction

Education microdata in England is both rich and well developed, facilitated by the collection of pupil level records from compulsory schooling and supplemented by data collected for students studying in further and higher education. This chapter examines the administrative datasets held by the Department for Education and Skills (DfES) and related datasets held by the Learning and Skills Council (LSC) and the Higher Education Statistics Agency (HESA). These datasets provide education related data information for pupils and students, past and present, in the education system primarily in England but also, where applicable, in the rest of the UK[2].

2.2 DfES data warehouse

Pupil level data for England is collected and provided by the Department for Education and Skills (DfES). The central structure and storage facility for data held by the department is contained within the so-called 'Data Warehouse'. The facility has been designed to provide a repository for data as well as facilitating easy access, in terms of defining each element of data together with its relationship to other elements within the database. The data warehouse provide facilities to store, retrieve and analyse raw data, as well as providing a gateway for access to relevant tabulated information and reports.

Information contained in the data warehouse relates to records on pupils characteristics and achievements with historical information archived. Basic information is also stored on schools such as size, funding type, staffing, finances, denomination, spatial coordinates, admission policy, etc. This data can then be link backed to pupil level data as required. Similarly, basic geographical information is also included and details relating to the Local Education Authority (LEA) are also incorporated. The linkages within the database are illustrated in Figure 1.

The database is updated on an annual cycle with information being loaded into the system as and when it becomes available throughout the year. Within this process the raw data is checked and transferred to table structures of the warehouse with items such as name and date of birth removed (or converted into alternative numerical forms including a unique pupil number (UPN)) in order to ensure pupil anonymity[3]. Although underlying tables are governed by a relational database structure, data can be retrieved from the data warehouse either in the form of a flat file (i.e. containing specified variables) or so called data cubes which can be created via on-line analytical processing in order to facilitate on line data analysis.

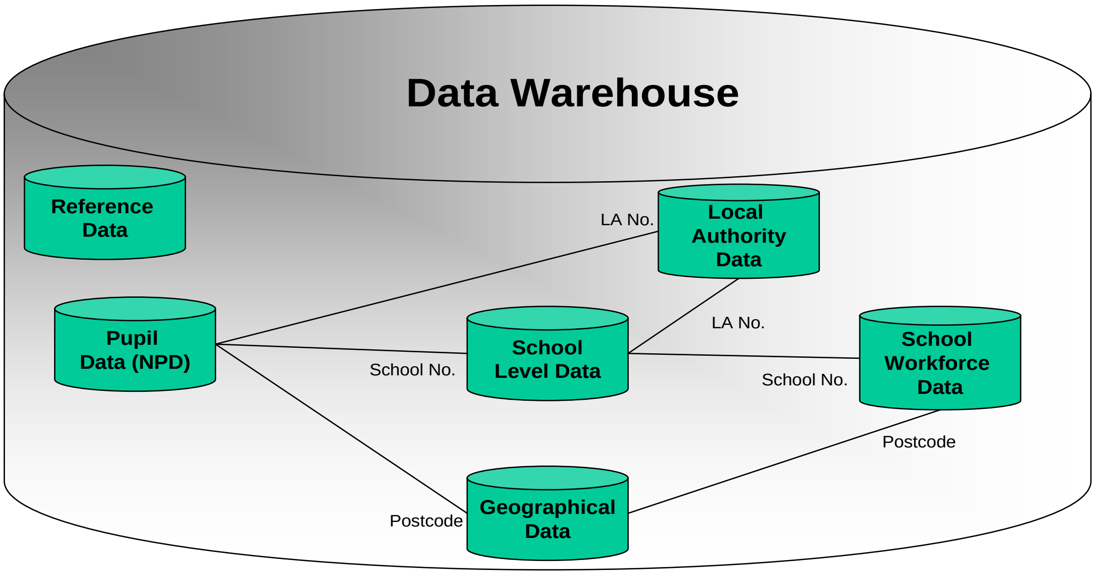
As well as holding data collected from within the department, the data warehouse also holds information collected by other agencies and key reference data. This includes data on individual learner records, collected by the Learning and Skills Council (LSC). There are also plans to integrate social care statistics, previously controlled by the Department

[2] Other parts of the UK outside England have different systems of data and effectively operate independently of the DfES. Data for Wales is collected by the Local Government Data Unit for Wales. Data for Scotland is provided by the Scottish Executive. Data for Northern Ireland is provided by the Department for Education and Learning for Northern Ireland.

[3] Note that all individual data items are protected by a security function such that access to the system is under restricted password access.

of Health (DOH), now held by the DfES. In addition there are current developments to incorporate and link external information on university applications from the Universities and Colleges Admissions Service (UCAS) and on higher education from the Higher Education Statistics Agency (HESA) into the database, linked via the UPN. This, in due course, will allow individual education records to be linked throughout the education process from earliest schooling through to university and higher education[4].

Figure 1: Illustrating the DfES data warehouse



Source: DfES

2.3 National Pupil Database (NPD)

The National Pupil Dataset (NPD) is a new dataset, first coming into being in being 2002, which contains linked individual pupil records for all children in the state school system. It is updated annually. In excess of 8 million individual pupil records are added to the database each year, including a range of variables such as pupil age, gender, ethnicity, special educational needs, free school meal entitlement, key stage assessments and public examination results, home postcode and school attended. The full census coverage of pupils, combined with a longitudinal aspect as records are linked each year, via a unique pupil number (UPN), allow longitudinal views as well as cross-sectional education and therefore make this potentially a powerful resource from which to conduct policy-related research relating to education.

The NPD is made up of several sub datasets which are held in separate, linkable, files using the relational database structure within the DfES data warehouse facility. The NPD dataset combines information on pupil and school characteristics data from the Pupil Level Annual School Census (PLASC) dataset with information on pupil attainment, as well as incorporating reference data on schools and local education authorities (LEAs). The main elements of the data are outlined below.

[4] These developments were ongoing at the time of writing.

2.3.1 Pupil Level Annual School Census (PLASC)

Since January 2002 the Department for Education and Skills (DfES) has carried out the Pupil Level Annual School Census (PLASC)[5]. This collects data on individual pupil characteristics (such as ethnicity and eligibility for free school meals) for all pupils in maintained schools in England. PLASC forms the foundation of the NPD database. At the beginning of each year, on an annual cycle, schools are required to return information to DfES on pupils within their schools in the form of individual pupil records. PLASC is statutory for all maintained primary, secondary and special schools. Maintained and direct grant nursery schools may also make an optional PLASC return. The submission of a PLASC return, including a set of named pupil records, is a statutory requirement on schools under section 537A of the Education Act 1996.

PLASC returns contain information on key pupil characteristics. These include variables such as ethnicity, a low-income marker, information on Special Education Needs (SEN). As well as information on pupils, information on teaching and support staff, classes as taught and (for some schools) admission appeals, is also included in the PLASC return. This provides a rich set of data on school characteristics. As it is a census, this includes details of the peer group (the school-cohort) for any particular child.

2.3.2 Attainment

In addition to the census of pupils in schools, facilitated by PLASC, the DfES also collects information each year on attainment for the approximately 2.5 million pupils sitting statutory Key Stages (KS) of the National Curriculum assessment at the following stages of educational achievement:

- Age 6/7 (KS1);
- Age 10/11 (KS2);
- Age 13/14 (KS3);
- GCSEs or equivalent (KS4);
- AS or Advanced level (KS5).[6]

In particular, the attainment data includes

- Levels attained in reading, writing and mathematics at Key Stage 1;
- Levels attained and the marks achieved in English, Mathematics and Science at Key Stage 2 and Key Stage 3;
- GCSE or equivalent results;
- AS and Advanced level results.

From 2003 onwards this information has been supplemented by a 10% sample of foundation stage profile data which records a range of teacher assessments at the end of the foundation stage (pupils aged 5).

[5] Note that from 2006 the 'School Census' replaces PLASC for maintained secondary schools, CTCs and Academies in England and, on a voluntary basis, Service Children's Education (Secondary Schools). The school census will be extended to earlier years from 2007 onwards.

[6] For information on the National Curriculum see: http://www.nc.uk.net/nc_resources/html/about_NC.shtml

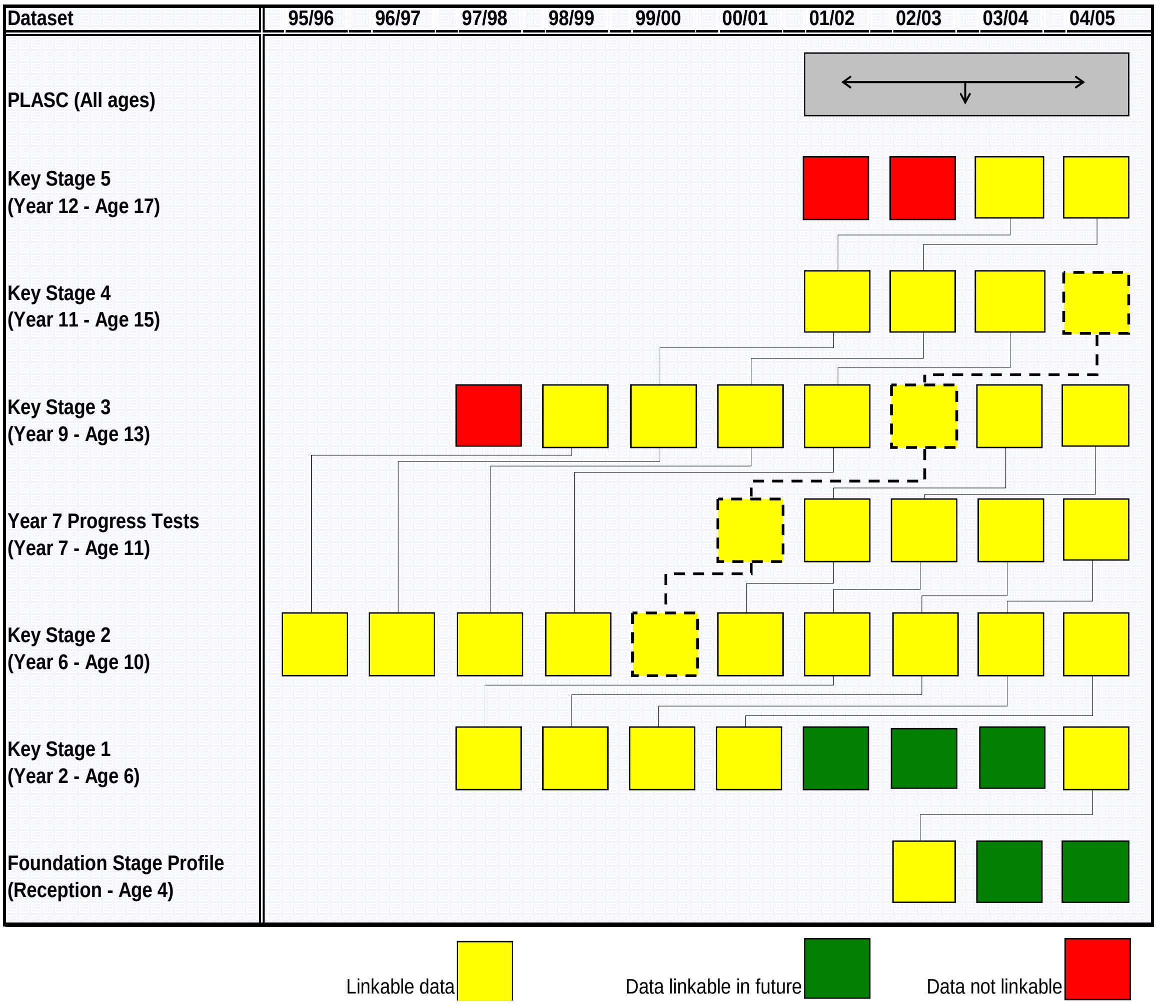
This information on pupil attainment described above is linked with the PLASC information using the data warehouse structure. The linked PLASC and attainment data form the National Pupil Database (NPD), to provide data source which links attainment to characteristics. The data which are now linkable at a micro level using the unique pupil number (UPN) is shown in Figure 2.

2.3.3 Longitudinal profile

The linking of pupil data on characteristics and attainment over time means that a longitudinal profile of pupils is now available, tracking children throughout their school careers. With repeated measurements of pupil attainment, at key points in their education rather than fixed time intervals, and a yearly record of their characteristics it is possible to identify and analyse the links between characteristics and pupil attainment and progress. Currently, it is possible to track children who were aged 6-7 in 1997/8 at Key stage 1 through to their current studies at Key stage 4 (see Figure 2). Over time this longitudinal aspect will increase as new annual datasets are added on an annual basis.

The advantage of longitudinal over cross sectional data is that one is able to track individual pupils over time and observe their transitions each year through the educational system. Statistically, such data allow for more precise formulation of models to consider the effect on educational outcomes of, say, pupil attributes or circumstances, policy interventions, and so on.

Figure 2: National pupil database (NPD) datasets and linkage

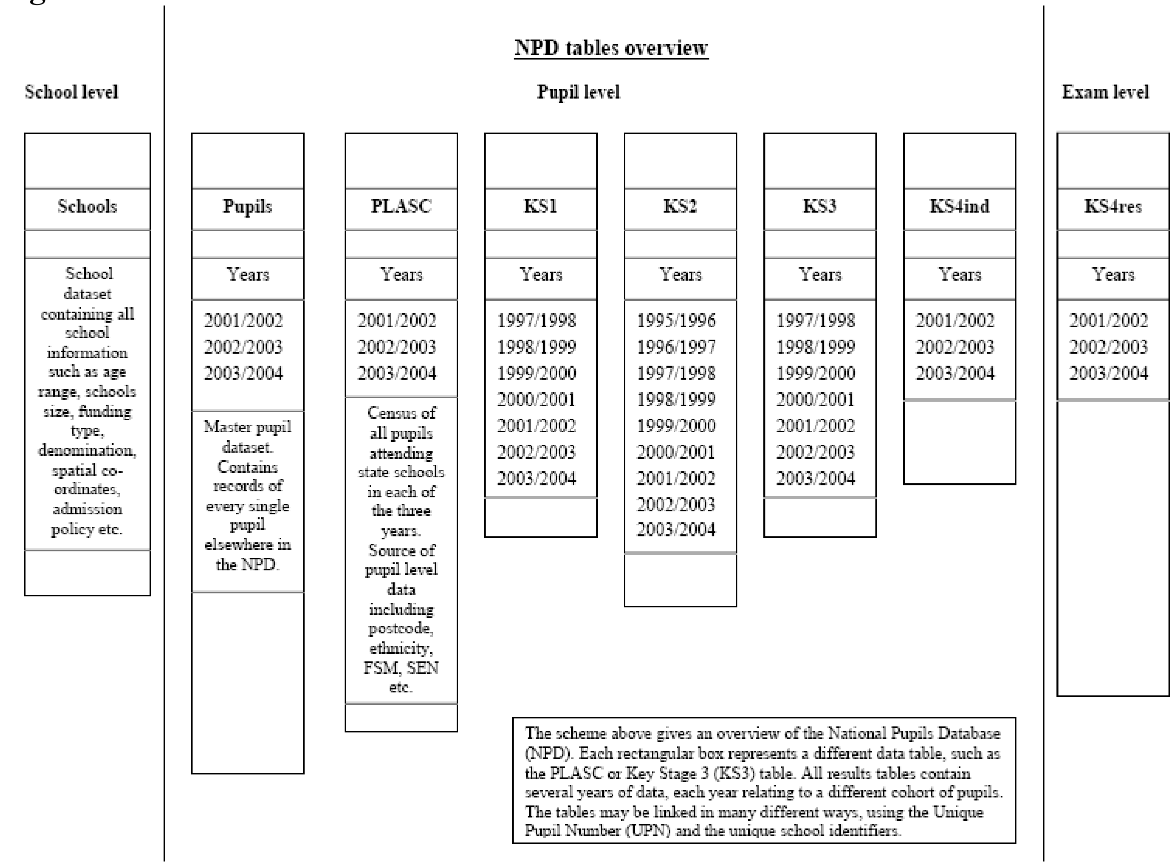


2.3.4 NPD key tables and variables

The National Pupil Database (NPD) contains millions of observations relating to, amongst other things, pupil characteristics and attainment over several years since 1996/7. This section summarises the key data tables within the NPD as well as an overview of variables.

The NPD is stored as a relational database with micro-level data generated by linking pupil level data using the unique pupil number (UPN) as an identifier. The key tables stored within the NPD as well as the timing of their availability is summarised in Figure 3. It is noted that as well as containing the PLASC and pupil attainment data described previously, the NPD also contains reference information relating to the school and LEA. In this way, as well as linking pupil data, it is possible to investigate the relationship between types of schools and pupil performance. Moreover, it provides scope for linking in geographical data from other related datasets - for example using information on neighbourhood socio-economic and demographic characteristics.

Figure 3: NPD tables overview



NPD tables overview

School level

Schools

School dataset containing all school information such as age range, schools size, funding type, denomination, spatial co-ordinates, admission policy etc.

Pupil level

Pupils

Years

2001/2002
2002/2003
2003/2004

Master pupil dataset. Contains records of every single pupil elsewhere in the NPD.

PLASC

Years

2001/2002
2002/2003
2003/2004

Census of all pupils attending state schools in each of the three years. Source of pupil level data including postcode, ethnicity, FSM, SEN etc.

KS1

Years

1997/1998
1998/1999
1999/2000
2000/2001
2001/2002
2002/2003
2003/2004

KS2

Years

1995/1996
1996/1997
1997/1998
1998/1999
1999/2000
2000/2001
2001/2002
2002/2003
2003/2004

KS3

Years

1997/1998
1998/1999
1999/2000
2000/2001
2001/2002
2002/2003
2003/2004

KS4ind

Years

2001/2002
2002/2003
2003/2004

Exam level

KS4res

Years

2001/2002
2002/2003
2003/2004

The scheme above gives an overview of the National Pupils Database (NPD). Each rectangular box represents a different data table, such as the PLASC or Key Stage 3 (KS3) table. All results tables contain several years of data, each year relating to a different cohort of pupils. The tables may be linked in many different ways, using the Unique Pupil Number (UPN) and the unique school identifiers.

Source: DfES

Detailed information regarding the variables contained in the NPD are provided in Annex 1. Annexes 1.1 and 1.2 describe variables contained in the LEA and schools lookup tables. Annex 1.3 describes variables contained in the Foundation Stage Profile Data for pre school children. Annex 1.4 describes variables contained in the Pupil Level Annual School Census (PLASC) data; and Annex 1.5 describes variables contained in the key stage attainment files.

2.3.5 Access to NPD data

Access to NPD data is now available through a central gateway via the PLASC/NPD User Group (PLUG). PLUG is a user group established to extend and facilitate access to the National Pupil Database (NPD) and Pupil Level Annual Schools Census (PLASC) datasets. The group is jointly funded by the Department for Education and Skills (DfES) who control and manage the data, and by the Economic and Social Research Council (ESRC)[7]. It is managed and resourced at the Centre for Market and Public Organisation (CMPO) at the University of Bristol[8].

Potential users must complete a request form, which contains fields relating to contact details and the nature of the data required and invites the researcher to outline the nature of the proposed research. This facility is open to all potential researchers, practitioners and students supported by their academic department. Subsequent to the form being submitted, it is processed by the DfES Analytical Services who act as a gatekeeper to the NPD. They respond accordingly to the request, although requests can take a period of time before they are processed. It is emphasised that requests are treated sympathetically and the department is keen to facilitate access to data as far as this is practicable. It is intended to make smaller "off the shelf" data extracts available quickly and easily through the PLUG website.

In the event of the request for access being successful the DfES provide the researcher with a flat data file (i.e. without a supporting database structure) containing all the relevant requested variables. Data extracts are anonymised and as part of the undertaking with DfES are required to be held securely, and cannot be passed to others. It is noted that once the data is released there is no further supervision provided by the department (where research is commissioned by the DfES itself) although support is available through the user group.

The size and complexity of the NPD data extracts can potentially cause problems relating to computing capacity and storage. The underlying NPD dataset is very large indeed (in any one year, approximately 8,000,000 pupil records are added to the dataset) so that anything other than a very well specified subset of the data, with minimum necessary variables specified, may be excessive in size, with respect to current limitations of personal computers. This should therefore be borne in mind when requesting data.

Finally, as well as providing a gateway for access to the data, the PLUG user group also aims to promote use of the data resource amongst the academic and wider research community. Whilst researchers have previously used the DfES data in a somewhat *ad hoc* manner, the PLUG aims to increase awareness of the data and facilitate the sharing of ideas and good practice. To this end, the user group is hosting a series of workshops through to 2008 aimed at attracting participants from academia and DfES as well as practitioners from LEAs and Learning and Skills Councils (LSCs). The workshops provide a forum for sharing best practice and the dissemination of research results. In

[7] The current funding arrangements run for three years to 2008.

[8] Access to DfES pupil level data is done by a request form via the PLUG website, which can be found at: http://www.bris.ac.uk/Depts/CMPO/PLUG/index.htm. NPD data for Wales can be found at http://www.lgdu-wales.gov.uk/. Provision of microdata is less well developed elsewhere. Education data for Scotland can be found at http://www.scotland.gov.uk/Topics/Statistics; and for Northern Ireland at www.deni.gov.uk

addition to the workshops, the user group actively updates current activities on the web page and has a publications mailing list. There is some data documentation on PLASC/NPD on the PLUG website now, and this will be significantly enhanced in the next few months.

2.3.6 Background reading and documentation

The underlying documentation on the NPD, including a summary of available data, lists of variable, variable coding, etc. are provided by the PLUG user group and can be accessed via the weblink. PLUG also post frequently asked questions (FAQs) in order to explain and simplify the use of the data. In addition to this, however, an introductory briefing for researchers and research users of the NPD data is provided by Ewens (2005a)[9].

2.4 Further Education data

The pupil level data documented in the previous section includes those in post compulsory education at school (for example studying for A level qualifications in schools) but not those studying outside the school system post-16 in further education establishments. Information on this group of students in England held by the Learning and Skills Council (LSC) in the form of Individual Learner Records (ILRs). The LSC is responsible, under legislation in the Learning and Skills Act 2000, for overseeing provision education and training for those above compulsory school age but not studying in higher education establishments.

2.4.1 Individual Learner Record (ILR)

The LSC collects information on students in further education in the form of Individualised Learner Records (ILRs). This information facilitates LSC strategic planning, including monitoring student retention, achievement, success rates and progress towards the LSC's targets. ILR records are supplied to the LSC by colleges and establishments providing education provision along with Individualised Staff Records (ISRs). Colleges of further education and other establishments providing education provision are obliged to supply information to the LSC if they receive one of the following types of funding:

- Further education (FE);
- Work based learning (WBL) including apprenticeships;
- Adult and community learning (ACL);
- European social funding (ESF);
- Train 2 Gain funding;
- Other LSC funding and have agreed to return ILR data.

The ILR is designed to capture the following information in respect of an individual learner:

- Personal details, including a unique learner number (ULN);
- Learning aim(s);

[9] Ewens (2005) also covers the more specialised London Pupil Dataset (which largely, but not entirely, a subset of the NPD). See: http://www.bris.ac.uk/Depts/CMPO/PLUG/publications/ewens.pdf

- Level of achievement;
- Financing information;
- Destination on completion / withdrawal of studying.

2.4.2 Uses of Individual Learner Record (ILR) data

The ILR data is used for internal purposes so that the LSC can monitor student retention and progress. In addition to this the LSC also uses ILR data to account for the funds allocated to it by central government. In addition to this, however, the data is used more generally. The annual ILR returns data is analysed and published in:

- DfES Statistical First Release (SFR) programme;
- LSC National Benchmarking data.

Details of the SFR, including relevant reports can be found later in this report in section 2.6.1. The national benchmarking data is published as part of the LSC's strategy to support institutions in raising the standard of their work. Benchmarking Data on student success, retention and achievement allows institutions to assess their performance, and assists their planning of action programmes to improve learning outcomes. The Benchmarking Data summarises data derived from institutions' Individualised Learner Record (ILR) and Individualised Student Record (ISR) returns and provides a range of national statistics[10].

2.4.3 Matching Individual Learner Record and the National Pupil Database

As a result of a recent project between DfES and the LSC, information from the ILRs has now been matched with information contained in the National Pupil Database, under the auspices of the DfES data warehouse, which currently includes records for all young people who were aged 16-20 attending a school in England at the end of the academic year 2004/05 (hence captured through PLASC). The dataset includes all individuals who have taken a qualification recognised by either DfES or LSC controlled awarding bodies and/or engaged in post compulsory schooling. Data on ethnicity, gender, and learning difficulties/disability are included in the matched dataset, as well as data on entitlement to Education Maintenance Allowance (EMA), where applicable, and geographical data for each learner (at local LSC and local authority level). The dataset covers between 98 and 99 percent of all people in this age group, therefore establishing a micro level dataset which offers the ability to track pupils beyond 16 and into further education. The matched dataset is available via the DfES data warehouse and, in anonymised form, to the LSC's data partners as long as they have completed the appropriate data sharing protocol forms.

More general access to Individual Learner Record (ILR) data is restricted by the LSC data sharing protocols. Whilst the LSC are bound by the Data Protection Act 1998 to regulate the processing of information relating to individuals, the Learning and Skills Act 2000 empowers the LSC to process and share learner with other designated bodies providing that the LSC controls aspects of data sharing and protects confidentiality. This is done by the establishment of a set of protocols which have been developed in order to

[10] A summary report for 2005/6 can be found at the following web link: http://www.lsc.gov.uk/National/Partners/Data/Statistics/LearnerStatistics/LearningAimOutcomes/FESuccessRates/fe-benchmarking-data.htm

form an agreement between the LSC and third parties that request access to anonymised or aggregated data. The agreement allows the organisation signing the protocol to request information either as one-off datasets or regular data feeds, and provides a framework under which the processing of the data provided by the LSC should be performed. The agreement lists the requirements that the LSC places on organisations wanting access to such data, and highlights at a high level the process of how the LSC will consider such a request[11].

2.5 Higher Education Statistics

2.5.1 Higher Education Statistics Agency (HESA) Data

The Higher Education Statistics Agency (HESA) is the central source for higher education statistics and is responsible for the collection and provision of data on students in higher education the UK. HESA collects data via an online network from universities and colleges of higher education numbers of students and staff in employment, as well as on destinations of higher education graduates following the completion of their courses. HESA collects the following data sets on an annual basis, relating to the respective academic year.

- Census of students;
- Census of staff;
- Destinations of Leavers from HE;
- Financial aspects of HE institutions.

The HESA data is rich in detail and contains many variables collected at the individual level, including those in the student dataset**,** which covers a census snapshot of the current student body; and the student first destination dataset, which covers students who completed their courses in the previous academic year. Details regarding major variables covered in these datasets are shown in Annex 1.7.

HESA provides access to anonymised microdata relating to individual students or staff on 'request-by-request' basis. Individual researchers/ organisations can request a specific set of data (with specified variables) by completing the necessary data request form. HESA operates as a commercial organisation and sells the data to interested parties and researchers. The requested data is supplied in the required data format on a CD-ROM. A major limitation, however, is cost as the data can be prohibitively expensive for some researchers. As well provides access microdata, HESA also produces a range of annual publications which summarise these datasets, taking various cross sectional cuts of the data[12].

[11] Parties interested in requesting data from the LSC in the first instance should look at the following weblink, selecting 'Data Sharing Protocols': http://www.lsc.gov.uk/National/Partners/Data/default.htm

[12] Online summaries of the student data are available at: http://www.hesa.ac.uk/holisdocs/home.htm

2.5.2 Matching Higher Education Statistics Agency data and the National Pupil Database

Efforts to match HESA data to the NPD at the individual level are currently being made at the time of writing. It is emphasised that this is a large undertaking, in practical terms and in terms of resources, and whilst DfES Analytical Services are supportive of the idea, such data linked at pupil level do not currently exist within the data warehouse. (Note that the HE data will be fuzzy – linked rather than based on the UPN).

The initiative to drive this forward is primarily the results of effort by interested researchers at the Institute of Education, London School of Economics and Institute of Fiscal Studies. In particular, the research team of Anna Vignoles, Alissa Goodman, Stephen Machin and Sandra McNally, funded via ESRC Teaching and Learning Research Programme, are attempting to create a longitudinal data built upon PLASC which traces the lifecycle of pupils/ student out of school and into university. The collected data is intended to provide data for the Project: 'Widening Participation in Higher Education: A Quantitative Analysis' which aims to identify the determinants of participation in higher education as well as barriers to progression. In the simplest terms this is done by tracking and link the following cohort:

- Year 11 pupils in state schools in England 2001/02;
- Completion of A level / F.E. / Year 13 studies in 2003/04;
- Those who enter HE at age 18 or age 19 will appear in HESA data for 2004/05.

As well as tracking students' transitions, the project also aims to match in pupils GCSE (2001/2) and Year 13 (2003/04) attainment data as well as application to university collected by the Universities and Colleges Admissions Service (UCAS) who for this cohort will apply primarily in 2002/03. Details of the project are available in Annex 1.7[13].

2.6 Relevant research

The use and analysis of the pupil level data resources is undertaken both internally and externally. Internal use of the data is performed primarily by DfES Analytical Services in order to provide the analysis of data in the Statistical First Release (SFR) series, the basis of the DfES research output. In addition, the DfES utilises researchers contracted to DfES to undertake policy relevant research, usually involved in analysis of specific educational programs or initiatives as part of a general drive to provide better targeting of funding, and the monitoring and development of policy. External use of the data, *i.e.* outside the control and remit of the DfES is chiefly undertaken by interested academics, but also more broadly by other parties such as local authorities or the Learning and Skills Councils. The research output, undertaken by both internal and external researchers, is considered below.

[13] Note that Mark Corver of HESA research is also linking PLASC and HESA databases.

2.6.1 Statistical First Releases

The DfES uses much of collected information on pupils and attainment, supplemented by information from Individual Learner Records (ILRs) and the Higher Education Statistics Agencies (HESA), in order to produce statistical summaries. These summaries are typically based on annual cross sectional data and are updated accordingly. The data we summarised and released via the DfES Statistical First Release (SFR) Series (SFRs are updated into final products some time after release) and published on the DfES Research & Statistics Gateway[14].

Table 1 summarises the main DfES statistical release products relating to pupils and students in England. Note, however, that this list is indicative rather than exhaustive.

Table 1: Selected DfES statistical releases

SCHOOLS AND PUPILS
• **Schools and Pupils in England – s**ummary of the Annual Schools' Census, reported on various dimensions; • **Pupil Absence in Schools in England –** absence in schools in England based on data collected from the Absence in Schools' returns; • **Foundation Stage Profile –** national results for the Foundation Stage Profile assessments.
ATTAINMENT
• **National Curriculum Assessment, GCSE and Equivalent Attainment and Post-16 Attainment by Pupil Characteristics –** summary of achievements and post-16 attainment of young people in England by different pupil characteristics; • **National Curriculum Assessments of 14 year olds in England –** Key Stage 3 National Curriculum assessment results for all pupils in all schools in England; • **National Curriculum Assessments of 7 and 11 year olds in England –** Key Stage 1 and Key Stage 2 data; • **National Curriculum Assessments at Key Stage 3 and Key Stage 2 to Key Stage 3 Value Added for Young People in England–** providing information on value added measures; • **National Curriculum Assessments and Key Stage 1 to Key Stage 2 Value Added Measures of 11 year olds in England –** providing information on value added measures.
FURTHER EDUCATION
• **GCE/VCE A/AS Examination Results for Young People in England –** annual summary of results by school type, gender and subject; • **Further Education and Work Based Learning for Young People - Learner Outcomes in England –** using data from LSC Individual learner records; • **Student progress between GCSE/GNVQ and GCE/VCE A/AS Levels: Schools and FE Colleges in England –** analysed by school type, gender and subject.
HIGHER EDUCATION
• **Higher Education Statistics for the United Kingdom –** annual summary based on HESA records; • **Destinations of Leavers from Higher Education in the United Kingdom –s**ummary of first destinations of HE students based on HESA records.

[14] The SFR series can found via either of the following weblinks: http://www.dfes.gov.uk/rsgateway/index.shtml; http://www.statistics.gov.uk/

2.6.2 Policy related research

Much of the recent research undertaken using the NPD and PLASC data sources has been undertaken on behalf of the DfES in order to investigate particular policy relevant issues. Table 2 summarises some of the relevant, published, research in this area, based on DfES program evaluations and available in the department's research reports series. This list is organised around some of the recent research themes, these being: (1) the 'Aim Higher' programme, a recent initiative aimed at increasing participation in higher education amongst young people; (2) the 'Excellence in Cities' programme, a recent initiative targeted programme of support for schools in deprived areas of the country; (3) the 'Playing for Success' program, an initiative establishing out of school hours study support centres at football clubs and other sports' grounds; as well as (4) the wider analysis of school resources and pupil attainment undertaken on behalf of the DfES. Note that this list is indicative rather than exhaustive.

Similarly, Table 3 lists other related research using DfES pupil level data, undertaken primarily by academics with interests in education related fields. These are organised under selected research themes; *i.e.* (1) Ethnic Minority Achievement, (2) Pupil Mobility and (3) Value Added from Education. Whilst this list is also indicative, it is noted that academic research on these micro datasets, conducted independent of DfES contractual stipulations, is relatively new and has primarily been facilitated via the PLUG.

Table 2: Selected DfES programme evaluations utilising NPD/PLASC

'AIM HIGHER' PROGRAMME
• Excellence Challenge Pupil Outcomes One Year on. See: Morris *et al.* (2005); • Excellence Challenge. The Early Impact of Aim Higher: Excellence Challenge on Pre-16 Outcomes: An Economic Evaluation. See: Emmerson *et al.* (2005).
'EXCELLENCE IN CITIES' PROGRAMME
• An Analysis of Pupil Attendance Data in Excellence in Cities (EIC) Areas. See Morris and Rutt (2005); • Evaluation of Excellence in Cities Primary Pilot 2001-2003. See Ridley and Kendall (2005); • Economic Evaluation of Excellence in Primary Schools. See Emmerson *et al.* (2003). • Improving Pupil Performance in English Secondary Schools: Excellence in Cities. See Machin et al (2004)
'PLAYING FOR SUCCESS' PROGRAMME
• Playing for Success: the Longer Term Impact: A Multilevel Analysis. See Sharp *et al.* (2004).
SCHOOL RESOURCES AND PUPIL ATTAINMENT PROGRAMME
• Estimating the Relationship between School Resources and Pupil Attainment at GCSE. See Jenkins *et al.* (2006a); • Estimating the Relationship between School Resources and Pupil Attainment at Key Stage 3. See Jenkins *et al.* (2006a).

Table 3: Other selected research utilising NPD/PLASC

DIVERSITY AND ETHNIC MINORITY ACHIEVEMENT
• Ethnic segregation and educational performance at secondary school in Bradford and Leicester. See Johnston, Wilson and Burgess (2006); • Special educational needs and ethnicity: Issues of over- and under-representation. See: Lindsay, Pather, and Strand (2006); • The Dynamics of School Attainment of England's Ethnic Minorities. See Burgess *et al.* (2005); • Combining multilevel analysis with national value-added datasets: a case study to explore the effects of School diversity. See Schagen and Schagen (2005); • Ethnicity, educational attainment and the transition from school. See Bradley and Taylor (2004).
PUPIL MOBILITY
• The Mobility of English School Children. See Machin, Telhaj and Wilson. (2006); • Moving Home and Changing School: Widening the analysis of pupil mobility' Greater London Authority. See Ewens (2005).
VALUE ADDED
• An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value. See Taylor and Nguyen (2006).

2.7 Summary

The National Pupil Database (NPD), facilitated by the DfES data warehouse infrastructure is a new and exciting development in terms of providing a platform for conducting evidence-based policy research in education. The sheer scope of the data, both in terms of its longitudinal nature and census coverage, mean that it provides a resource for researchers that is unmatched in many other areas of social research, and is a valuable research resource in its own right and with the potential to link to other survey – based research resources.

The forward looking nature of DfES Analytical Services, in creating a permissive and supportive environment for access to data, has helped raise awareness of the data and broaden use of the data resource. Not least of all this has been facilitated through the creation in 2005 of the Pupil Level Data User Group (PLUG) who provide a gateway for access to the data as well as support for users. Before this time access pupil microdata was limited to limited circle of researchers and was not well documented.

In addition to the development of the NPD over recent years there are now active moves to extend the linking of pupil information from the NPD to include students in further and higher education. The data link with students in further education outside school (*i.e.* in colleges, work or learning providers) has now been established within the data warehouse, based on information from the Individual Learning Records (ILR) provided by the Learning and Skills Council. Similar efforts are now being made to establish a link between pupil records in the NPD (identified by a unique pupil number) with students in the Higher Education Statistics Agency (HESA) data, at university or studying in a college of higher education.

3 Labour market related datasets

3.1 Introduction

A great deal of information is routinely collected at the level of the individual relating to their contact with the labour market. Details collected through the 'Pay as You Earn' (PAYE) taxation and National Insurance (NI) payment system contain much information on employment and earnings. Similarly information on receipt of benefits, whether through unemployment or inactivity, helps to complete the picture with respect to an individual's economic activity. Efforts to consolidate this information have recently been undertaken by the Department for Work and Pensions. The resulting longitudinal study achieves almost census coverage of working age individuals over many years. This resource, still in its infancy in terms of utilisation by government and researchers, offers unique opportunities to study labour market transitions and outcomes[15].

3.2 Department for Work and Pensions Longitudinal Study (WPLS)

The recently created Work and Pensions Longitudinal Study (WPLS) is a longitudinal, spells-based administrative database that allows analysis of individual DWP client history. It brings together and stores a wealth of administrative data relating to the past labour market status of each individual in the UK from 1998 onwards. All individuals who have since this time had contact with the Department for Work and Pension (DWP) are included in the dataset. The dataset has existed in its current form since October 2005. The WPLS enhances the Lifetime Labour Market Database (LLMDB2) which performs broader functions around National Insurance Contributions, but only covers a 1 percent sample of the population.

The 100 percent coverage means that the WPLS is unmatched in the UK as a longitudinal data resource. As well as providing a very much enhanced facility for welfare-to-work and pension planning related policy analysis within the department, it also provides an important resource for future research work on individual labour market experiences and transitions[16].

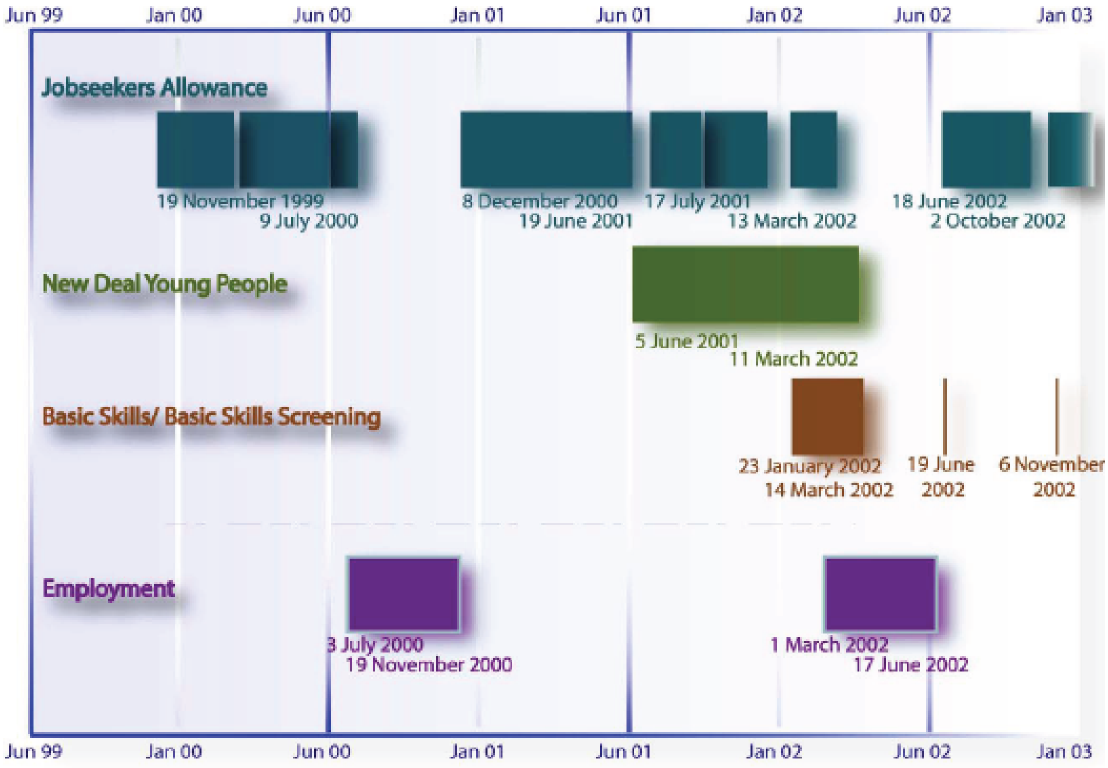
The WPLS is a large relational database which brings together benefit and programme held by the DWP on its customers with information collected from Her Majesty's Revenue and Customs (HMRC). The WPLS acts as a national benefits database. It also stores information for all individuals on employment and earnings based on information collected via the income tax system from P45 and P46 tax return records. The linking of data within the WPLS database is done using National Insurance Numbers (NINOs) and a range of other variables relating to individual characteristics using fuzzy matching algorithms.

[15] This chapter focuses on employment related administrative data held by the Department for Works and Pensions (DWP). In addition to this departmental source, administrative based labour market data - on claimant unemployment and registered vacancy stocks and flows – is available through NOMIS, who provide open access for researchers. See: www.nomisweb.co.uk

[16] The data-sharing facility between DWP and HMRC was facilitated by the Employment Act 2002. The data sharing was initiated by the DWP in the first instance in order to track individual activities in relation to benefit fraud.

In essence the longitudinal data in the WPLS stores information which allows one to construct an entire labour market history for each individual in the UK back to 1975, including spells of work, unemployment or worklessness, and providing information relating to earnings during these periods. The information, in the most basic terms, is retrievable from the database as a series of spells of economic activity. Employment spells are recorded as are period of non-employment characterised by type of benefit being claimed during the interim period. Likewise, periods of worklessness where no benefits are being claimed, *i.e.* inactive unpaid (housewives, full time students, *etc.*), can be inferred from the gaps between employment and benefit spells[17]. The spell information records the date an activity/program commenced and was completed as well as the interim duration. Income information (earnings or benefits) is recoverable by spell period. An illustration of a hypothetical event history which could typically be constructed from WPLS is shown in Figure 4 for illustration.

Figure 4: Hypothetical event histories from WPLS



Source: DWP

The WPLS covers 100 percent of benefit and pensions related information based on historical information stored by the respective departments. It also covers all employment spells for individuals who have had contact with DWP at some point in the past as a benefit claimant. In terms of size, when the data was released in 2005, it stored details on 27 million individuals who had had contact with DWP at some time as a customer. This is combined with a total of 135 million employment spell records. The data is stored in a relational database which contains a whole array of variables. The main categories of variables in WPLS are summarised in Annex 2.1.

[17] Note, however, that these gaps could also feasibly refer to periods when the individual was out of the country or in an institution (prison, hospital, etc).

3.3 Lifetime Labour Market Database (LLMDB2)

The Lifetime Labour Market Database (LLMDB2) enhances the WPLS in terms of key data resources for monitoring longitudinal patterns of economic activity (work and worklessness). The impetus behind LLMDB2 is to track a large number of individuals through their working lives, to obtain information on the patterns of transition between the different work types and benefits, and to follow their progress into retirement. The LLMDB2 provides a resource of information about National Insurance Contributions (NICs) and second tier pension provision and has attracted the attention of a number of academic researchers. The research output from this source is documented in section 3.6.

The Lifetime Labour Market Database (LLMDB2) is a 1 percent sample (based on National Insurance Numbers ending in the digits 14) of all people in the UK and abroad (paying voluntary NICs) appearing on the National Insurance Recording System (NIRS) data between 1977/8 and 1995/6 and its replacement NIRS2 system from 1996/97 to the present.. NIRS2 contains details of National Insurance records for over 60 million individuals that are required to calculate entitlement to benefits and State Pension. The database provides an historical view of the labour market and is a well established longitudinal resource for studying labour market transitions over this period with little or no sample attrition rate and without many of the problems of bias encountered in other longitudinal studies. Very much similar to its larger WPLS counterpart, the LLMDB2 contains a complete employment, reckonable earnings and contributions record of sampled individuals. The variables covered in the LLMDB2 relate to annual responses and cover the following areas.

- Annual earnings[18];
- Tax contributions;
- Pension contributions collected under employer schemes;
- National Insurance Contributions;
- Number of weeks of employment / self employment / unemployment;
- Number of weeks of receipt of sickness, disability or carers benefits;
- Industry of employment;
- Benefit receipt.

In principle the database allows a complete (synthetic) life history to be built up over time for each individual, allowing micro-based modelling of labour market activities and transitions over the lifecycle. It is also possible from the database to infer various pieces of lifecycle information such as individual mortality, retirement and partnership formation and separation. Work to this end has been undertaken within DWP using the LLMDB2 in work on the pensions simulation exercise using the pension simulation model described in section 3.6.

An additional strength of the LLMDB2, along with it larger and more unwieldy WPLS counterpart is that it can be linked to other datasets via the NINO identifier, including those held elsewhere in Her Majesty's Revenue and Customs or DWP.

[18] It is noted that one unfortunate feature of these data is that they record annual gross earnings without adjustment for weeks or hours worked. This creates potential difficulties in analysis/comparison.

3.4 Other Department for Work and Pensions datasets

Datasets that form part of WPLS exist in their own right as microdata and are held as part of the DWP's data warehousing facilities which support the longitudinal study. The benefits data listed in Annex 2.1 are available by type of benefit and are held as separate data files as a 100 percent and also using 5 percent coverage, available annually and quarterly. In addition the datasets are linked and stored as micro data in the **100% Benefits Dataset** (or what was previously known as the Working Age Statistical Databases – WASD). More details regarding this data are shown in Annex 2.2.

In addition to the benefits data which inform the WPLS, the DWP also holds a number of other micro-datasets. These datasets hold information on contact with individuals through job centres and on employment programmes and initiatives. They are used internally to monitor take up of programmes but can also be used to monitor success, performance, etc. These cover the large DWP employment initiatives such as the *New Deal*, *Basic Skills*, and so on, but also track routine contact with clients. Table 4 summarises the main datasets used within the Information Directorate of the DWP. However, it is noted that the list is not exhaustive.

The datasets listed in Table 4 are stored as microdata as an individual level spells database. More detailed information describing the data is provided in Annex 2.3. In terms of structure, each row in the datasets records a discrete period of time on a spell or benefit. The data contains a unique identifier for each client and individual details, including; date of birth, gender, address, ethnicity and national insurance number. In addition to the individual datasets there is also a **Master Index** file which brings together the historical program evaluation databases from the Employment Service and WASD (Working Age Statistical Database). The Master Index is the feed into the WPLS and aims to provide, for the first time, easy cross benefit and program analysis. The Master Index covers the main New Deal (ND) programs and further programs, pilots and pathfinders, including the various sub-programmes: NDYP, NDLTU, NDLP, ND50, NDDP, Basic Skills, Joint Claims, WBLA, EZONES, JRRP and EMO (see Table 4 for more details). The dataset also contains ten benefits taken from WASD and JOT pilot data. The main variables are a unique identifier, start and end dates of spells and the benefit/ program type for that period. Further details are contained in Annex 2.3.

In addition to the benefit and programme datasets listed above, the DWP also holds data in relation to 'Programme Protection' and tracking of fraudulent claims. There are also more derived datasets from master index files or produced on an ad-hoc basis, which are not listed here.

3.5 Access

Access to LLMDB2 is now well established and, as documented in section 3.6, has been used variously both internally and by external researchers. At the time of writing general access to the WPLS is not well established. Means of access to the DWP data resources is via written application to the department. There is currently no portal for open research access to microdata as has been established in other departments such as DfES and ONS.

Table 4: DWP Programmes and related datasets

Access to Work
Basic Skills Programme
Childcare Barriers to work
Employment Retention & Advancement Scheme (ERA)
Employment Zones
Ethnic Minority Outreach (EMO)
European Social Fund (ESF)
Extended Schools Childcare
Incapacity Benefit Reforms
Job Retention and Rehabilitation Pilots (JRRP)
Jobcentre Plus pathfinder
Jobcentre Plus vacancies and Jobcentre Plus employer
Joint Claims
New Deal Programmes:
- New Deal 50+ (ND50)
- New Deal for Disabled People (NDDP)
- New Deal For Lone Parents (NDLP)
- New Deal for Long Term Unemployed (NDLTU)
- New Deal for Partners of the Unemployed (NDU)
- New Deal for Young People (NDYP)
Non Jobcentre Job extract
Progress to Work & Progress to work LinkUP
Social Fund
Step up
Sustainability extract
Work Based Learning for Adults (WBLA)
Work Focused Interviews for Partners (WFIP)
Workstep Programme

Source: DWP
Note: See Annex 2 for a description of each dataset.

3.6 Uses of the DWP data

The DWP data resources are used on the main part for internal purposes, in terms of providing statistics and up to date management information both on the targeting of clients and on the impact of various programmes designed to help people on benefits get back to work. In addition, the department uses the datasets (and the WPLS in particular) for cross referencing information about individuals to investigate fraud. There has been some use of the DWP resources by academic researchers. In the main part this has involved use of the LLMDB2 data, as the longitudinal study is still relatively speaking a new resource; although a range of academics researchers have used WPLS whilst specifically contracted to DWP.

3.6.1 Government research

The DWP has modernised its use of statistics though the use of the Work and Pensions Longitudinal Study (WPLS) database. The WPLS has been able to facilitate significant improvements to the analytical evidence base and operational effectiveness of DWP statistics. The list of uses of WPLS, with respect to benefits and back to work programmes, includes:

- providing statistics, management information and research on the success of Jobcentre Plus in helping people into work and keeping them in work;
- helping to evaluate individual policies and their impact in the short, medium and long-term;
- aiding in the investigation of fraud.

With respect to pension provision, the WPLS enables the department to:

- understand working patterns in retirement and the impact on the claiming of benefits;
- understand the links between savings held and the benefits system in retirement, and how people are using or accumulating savings in retirement.

To this end the routine statistical analysis such as the provision of 'benefit to work' measures, monitoring job centre plus, and publishing National Statistics on benefits and New Deals is now done automatically through the new database. In addition to the automatic reporting of statistics the DWP also routinely performs or commissions research to look at the effectiveness of its programs. This is often undertaken by academic researchers and/or consultant researchers.

An example of work done in the past using micro data from the LLMDB2 is the work on the Pensions Simulation Model (PENSIM). PENSIM is a model built by DWP to estimate the future distribution of pensioner incomes and the impact of various policy changes on cost of pension provision. PENSIM worked by the construction of a dynamic micro simulation model using individual level data on economic activity, labour market transitions and income from various sources including the LLMDB2. In simple terms, the model establishes relationships between variables such as pensions, savings, labour market status, earnings, etc. This is done by constructing synthetic life histories for each individual using samples of individuals and households so that individuals within the model database become representative of the population. Outcomes with respect to pensions and pension entitlement are simulated forward into the future. PENSIM also allows the user to investigate various 'what if' policy scenarios based around pensions provision. A new version of the model (Pensim3) which incorporates new information from WPLS is due to be released during the next few years.

In addition to the above departmental uses of microdata the WPLS also has been used operationally to exclude those ineligible for Pension Credit from Pension Credit take up marketing campaigns.

3.6.2 Academic research

Academic research using DWP resources mainly centres on the use of the LLMDB2 to study issues relating to work, labour market transitions and earnings. In the main part the use of this dataset is due to longitudinal aspect which allowed individuals to be tracked through time. The WPLS is sufficiently recent in its development that at the time of writing little published research had emerged from this source. Examples of research using either the LLMDB2 or WPLS are listed in Table 5.

Table 5: Examples of research using the LLMD

- Employment, Welfare and Exclusion. See Burchardt and McKnight (2006);
- Labour Market Mobility in the UK. See Dickens (2006);
- Evaluation of Basic Skills Mandatory Training Pilots. See Joyce et al (2006);
- Evaluation of Skills Coaching trials and Skills Passports. See Hasluck et al (2006);
- Evaluating England's 'New Deal for Communities' Programme using the difference in difference Method. See Covizzi et al (2006);
- Gateway to Work New Deal 25 Plus pilots evaluation. See Page et al (2006);
- The profile of exits from incapacity related benefits over time. See Bertoud (2004);
- The employment effects of full participation in ONE. See Kirby and Riley (2003);
- New Labour and the labour market. See Dickens, Gregg, and Wadsworth (1999) ;
- Patterns of Work, Low Pay and Poverty - Evidence from the Lifetime Labour Market Database and the British Household Panel Study. See Endean R (1999);
- Male earnings mobility in the lifetime labour market database. See Dickens, Gregg, and Wadsworth (1997);
- The Department of Social Security's Lifetime Labour Market Database. See Nicholls, Marland and Ball (1997);
- Male Earnings Mobility in the Lifetime Labour Market Database. Ball and Marland (1996).

3.7 Summary

The Department for Work and Pensions (DWP) now collects and stores a great deal of high quality micro data not only on an individual's contact with the department itself but, using data obtained from Her Majesty's Revenue and Customs (HMRC), on the labour market experience of individuals in terms of employment, unemployment and spells of economic inactivity using data stored in the newly constructed Work and Pensions Longitudinal Study (WPLS). The WPLS stores millions of observations relating to the experiences of all individuals in the UK who have had contact with the department since 1999. Using this data facility, it is now possible to construct labour market 'event histories' for individuals over their life course. By achieving 100 percent rather than 1 percent coverage this resource improves upon and enhances the existing resources available under the Lifetime Labour Market Database.

4. Health related datasets

4.1 Introduction

A wide range of health and social care related statistics are now produced from various sources, broadly under NHS governance. The recently established Information Centre for Health and Social Care (Information Centre or IC for short) produces a range of health statistics. The Department of Health (DoH) also publishes a restricted range of statistics itself[19]. In addition to this the Office for National Statistics (ONS) publishes cancer statistics. These sources of health related data are discussed separately in this section of the report.

The process of collecting data is done by the central statistical function of the NHS which periodically collects information from data providers, primarily NHS trusts. Dta collection is governed by the Review of Central Returns (ROCR) which ensures that central data collections are done efficiently and appropriately so that resources are not duplicated[20]. The information required under ROCR is collected via the NHS-Wide Clearing Service (NWCS), although there is currently an ongoing programme to modernise and replace this system, as described in Section 4.6. ROCR operates by setting submission deadlines for data providers, with a rolling structure throughout the year so that data is processed by batching requests. All approved collections are allocated a ROCR Reference Number. Once collected, data is added to the DoH data warehouse.

Much of the focus of this section relates to the Hospital Episode Statistics (HES). This data relates to episodes of patient contact with the NHS hospital trusts. Information contained in HES originates from the Patient Administration Systems (PAS) of individual hospital trusts. The data is then sent at set times during the year to the NHS wide clearing service (NWCS) who store the information to a centrally held database. The data are then cleaned, anonymised, derived variables added, and subjected to quality checks before becoming the final HES dataset which is 'fixed' on a quarterly basis.

4.2 NHS Information Centre for Health and Social Care (IC)

The ROCR service is provided by the Information Centre for Health and Social Care (IC). The IC was created in April 2005 out of the former NHS Information Authority and the Department of Health Statistics Unit. Its remit is to reduce bureaucracy in the NHS by collecting information in the most efficient manner and making information more accessible. The IC is responsible for co-ordinating data collections and analysing information requirements in health and social care. As well as being responsible for data collection, the IC is also charged with responsibility for data quality and dissemination, developing analytical solutions and sharing 'best practice' within the user community. In addition the in-depth knowledge and expertise of the IC will be used to support users and to respond to emerging NHS policy requirements[21].

[19] Lists are available at: www.dh.gov.uk/PublicationsAndStatistics/Statistics/CodeOfPractice/fs/en

[20] The Review of Central Returns Steering Committee reviews NHS information requirements on an ongoing basis.

[21] Details regarding the operations of the IC, including links to the centre's data catalogue, can be found at the following website: http://www.ic.nhs.uk/ .The data catalogue can be found at: www.ic.nhs.uk/infocat .

4.3 Hospital Episode Statistics (HES)

The Hospital Episode Statistics (HES) is a centrally held data store which details all patient contact (*i.e.* patients, outpatients and via Accident and Emergency) with NHS establishments in England. This includes contacts with hospitals as well as mental health, primary care trusts and mental health trusts.[22] The database contains personal, medical and administrative details of all patients admitted to hospitals in England. The data are 'episode' based, in that each record in the dataset contains information relating to one episode of patient care, including its duration. The database contains approximately 13 million episode records each data year from 1989-90 onwards.

The HES data contains a wealth of information of potential use for research purposes. It contains clinical information relating to treatments received as well as demographic and geographical information relating to patients. In this respect HES offers a unique resource from which to conduct policy related research into health. Amongst the multitude of research questions which the data can potentially address are: trends in health; waiting times and length of stay; patient demographics; factors affecting health ill health; geographic variation in care; and so on.

4.3.1 HES data[23]

The Hospital Episode Statistics (HES) is an episode level dataset which records all patient contacts with NHS in England. Contacts are recorded through the Patient Administration System (PAS). An episode is recorded and processed once it is complete and each episode is assigned as HES identification number (HESID).

HES data contain information on each hospital episode. Each HES record holds approximately 100 items of information (including derived variables) relating to the episode of care. This includes clinical information on diagnoses and operations, dates of admission and treatment, administrative information regarding the timing and duration of the car episode, details relating to the individual and the NHS organisation, as well geographical information about the location of treatment and where the patient lived. It is noted that the dataset is anonymised so that personal identifiers such as the patients name and date of birth are excluded from the HES dataset. The main headings under which variables are categorised within the dataset are as listed below:

- Patient;
- Period of care;
- Detail of admission, discharge and episode spell;
- Healthcare resource resources used;
- Geographical details;
- Clinical information, including:
- Maternity (if applicable);
- Critical care (if applicable);
- Psychiatric (if applicable);
- NHS organisation information;

[22] Data for NHS hospitals in Northern Ireland, Scotland and Wales are collected separately by the appropriate sections within the Northern Ireland, Scottish and Welsh Offices.

[23] An extended summary of HES is available in the 'HES Book' (see Liffen, Maslen, and Price, 1988).

- General practitioner information;
- Other system variables.

For an extensive list of the variables contained in each of these categories, see Annex 3.1. Note that most of the variables with HES are collected at point of contact from the Patient Administration System (PAS). However, there are also a set of 'derived' variables which are imputed from other information contained within the HES system. Similarly, some variables are added to the HES system ex-post such as information on the date of death (when applicable)[24]. Other variables are not entered as raw data but are coded under HES for administrative purposes.

4.3.2 Access to HES microdata

Access to HES microdata (*i.e.* at episode level) is closely controlled. Despite the fact that the HES data are anonymised and do not hold individual names and addresses, the nature of the dataset is such that there is deemed to be a potentially high risk of disclosure (*i.e.* the dataset contains information that could, in combination or with prior knowledge, potentially identify patients). This is combined with the fact that the data contains facts about individuals which are highly confidential. Consequently, HES records themselves cannot normally be released, even in abbreviated form.

These considerations, however, are balanced against the acknowledged need of the research community who may need access to the best possible data (*i.e.* HES) in order to address important medical or health care policy related issues. In order to accommodate the research community's needs, access to an extract of the HES data (*i.e.* containing relevant fields and time period) may be given only with the permission of the Security and Confidentiality Advisory Group, an independent adjudicating body, following a stringent application procedure. Requests are made via the Information Centre. The right of access to the HES data and to data held more generally by the Information Centre is facilitated by the Freedom of Information Act 2000. This piece of legislation gives a general right of access to information not otherwise publicly available when such access is deemed to be in the greater public interest[25]. The stated policy on microdata release is as follows:

> "In line with the principles set out in the Caldicott Review of Patient-Identifiable Information, the Security and Confidentiality Advisory Group seeks to limit data access to what is absolutely necessary for the purpose" *(IC Information Governance[26])*

Each request is considered on a case by case basis in view of data security and patient confidentiality. It is noted that when providing an application, researchers must justify the need for HES records and explain clearly why aggregated data will not suffice for the purposes of the research.

In instances where access to the data is granted, individuals are given the relevant extract of the data. An extract contains record-level data for selected cases, showing selected information for each case. Access to the data is permitted only on signing strict

[24] This information is taken from the Office of National Statistics (ONS) mortality statistics.

[25] Individuals may also make a request for their own personal data under the Data Protection Act 1998.

[26] See http://nww.nhsia.nhs.uk/infogov/igt

undertakings of confidentiality (*i.e.* not to further disclose the data to a third party) and to abide by the HES data protocol (*i.e.* to use the data only for the specified purpose, to keep the data only for the specified length of time)[27].

4.3.3 Aggregated HES data

In most cases access to microdata is not appropriate or necessary. This is the case for example when *ad hoc* pieces of information, tabulations or summaries of data are required from HES. Instances of this kind may involve health professionals or managers, researchers or members of the public wanting access information on particular health issues, conditions or diseases.

In order to meet this demand the HES data is available in **aggregated** form where summary tables are produced based on the sum (or mean) of data over many individuals, NHS trusts, or possibly over various time periods. In practice, these tables are comprehensive enough to answer most general queries. For example, users can access information on the number of operations of a particular kind that were carried out in each region over a particular period of time[28].

Tabulations of the HES data, available from the website link, are available based on the following search categories:

- Patient diagnosis / condition treated;
- Operations undertaken ;
- Treatment type (using Healthcare Resource Groups categories);
- Hospital provision;
- Residence-based provision.

Follow the link on the HES website (footnote 25) to "Accessing Data", the website provides an extensive tabulation facility which allows cross tabulations of the HES dataset. The tabulations allow the user to design their own tailor-made table based on either a single total or a multi-dimensional query. This data is available interactively and on request ands is delivered via live online interrogation of the database using browsing software. The data is also downloadable in PDF or Microsoft Excel (XLS) format. In addition to the tabulation facility available online, summary reports are regularly published from HES (details in the next section). Furthermore it is also possible to request a tailor-made report on a particular aspect of HES, facilitated by the Information Centre, subject to a fee.

HES related statistical studies are also provided by the Department of Health (DoH) who publish regular statistical summary reports[29]. This includes statistical summaries of:

[27] HES protocol contains guidance on rules for disclosure when publishing small numbers of episodes in data tables. As a general rule, use of small numbers in tables is suppressed where suppressed where it is possible to infer information about an individual.

[28] This facility is made available through the HES website which can be found at: http://www.hesonline.nhs.uk

[29] Links to this data, along with a whole host of publications provided by the DoH statistical function can be found at: http://www.dh.gov.uk/PublicationsAndStatistics/Statistics/fs/en

- Hospital Activity statistics (inluding bed occupancy; outpatient numbers; A & E attendances; NHS day care provision; monitoring complaints; imaging and radiodiagnostics; and adult critical care);
- NHS waiting lists and waiting times;
- NHS perfomance against benchmark management indicators.

4.3.4 HES related research

Research involving HES is plentiful. A search of articles containing the term "Hospital Episode Statistics" through the *Web of Science* citations catalogue produced 84 articles published in academic science or related journals. The papers relating to results of studied based on HES, were published in leading scientific journals including the *British Medical Journal, Journal of Public Health* and the *British Journal of Clinical Pharmacology*, indicating the value of this data to the research community. Table 6 summarises a subset of these articles, i.e. those published in 2005/6.

4.4 Cancer registration data

The system of cancer registration in England and Wales offers a second potential source of microdata, although different channels of access exist for these data. The cancer registration system in the UK currently operates outside the HES framework and the data is controlled (via a complex chain of data collection) by the Office for National Statistics (ONS).

Table 6: Recent journal articles utilising HES

- The effects of surgical volumes and training centre status on outcomes following total joint replacement: analysis of the Hospital Episode Statistics for England. See Judge, Chard, Learmonth *et al.* (2006);
- Incidence of primary and recurrent acute urinary retention between 1998 and 2003 in England. See Cathcart, van der Meulen, Armitage *et al.* (2006);
- Mortality associated with delay in operation after hip fracture: observational study. See Bottle and Aylin (2006);
- Charlson scores based on ICD-10 administrative data were valid in assessing comorbidity in patients undergoing urological cancer surgery. See Nuttall; van der Meulen; Emberton (2006);
- Comparison of hospital episodes with 'drug-induced' disorders and spontaneously reported adverse drug reactions. See Barow, Waller, Wise (2006);
- Analysis of regional variation in hip and knee joint replacement rates in England using Hospital Episodes Statistics. See Dixon, Shaw and Dieppe (2006);
- Excess winter morbidity among older people at risk of cold homes: a population-based study in a London borough. See Rudge and Gilchrist (2005);
- Cardiovascular admissions and mortality in an inception cohort of patients with rheumatoid arthritis with onset. See Goodson, Marks, Lunt *et al.* (2005);
- Defining the minimum hospital case-load to achieve optimum outcomes in radical cystectomy. See McCabe, Jibawi, Javle (2005);
- Investigating population level trends in head injuries amongst child cyclists in the UK. See Hewson (2005);
- Comparison of Hospital Episode Statistics with the Association of Coloproctology of Great Britain and Ireland colorectal cancer database. See Garout, Tekkis, Darzi *et al.* (2005);
- A description of radical nephrectomy practice and outcomes in England: 1995-2002. See Nuttall, Cathcart, Van der Meulen *et al.* (2005);
- Was Rodney Ledward a statistical outlier? Retrospective analysis using routine hospital data to identify gynaecologists' performance. See Harley, Mohammed, Hussain *et al.* (2005);
- Follow up of people aged 65 and over with a history of emergency admissions: analysis of routine admission data. See Roland, Dusheiko, Gravelle *et al.* (2005).

4.4.1 Collection and storage of cancer data

The collection of cancer data in England and Wales is conducted by ten independent regional registries[30]. The regional registries collect data on cancers for individuals in their area based on information returns from hospitals[31]. The regional observatories each submit a standard dataset on these registrations to the Office for National Statistics (ONS)[32]. The coordinated collection of data by the ONS provides of national cancer registration scheme so that the data can exist in one place in the **Cancer Registrations**

[30] The national assembly for Wales is now responsible for cancer registration in Wales. Data for Scotland is collected separately under the auspices of the Scottish Executive.

[31] The regional cancer registries are also responsible for the analysis and data dissemination within their region, with their remit based on of public health surveillance and health protection.

[32] This system of data collection has been in place for over 30 years. However, the supply of cancer information has been compulsory since 1993.

dataset. The cancer data is individual patient record based and therefore creates a large micro dataset on cancer occurrence.

Within ONS the cancer data serves two functions. Firstly the cancer data is linked onto the Longitudinal Study (LS) census data which tracks 1 percent of the UK population over time through linked census data from 1971 to 2001. (More details regarding the LS are given in section 6 of the report). In simple terms, a cancer occurrence 'flag' is attached to the LS dataset to identify individuals who have had instances of cancer care[33]. This provides an excellent statistical resource for tracking cancer patients (and cancer survival) over time. An extended discussion of cancer data in the LS is presented in Section 6.2.5 of the report.

Secondly, the cancer registration data records are held as a stand alone dataset, i.e. National Cancer Registrations Database (or 'minimum data set'), which now contains almost 10 million records dated from 1971 onwards. A detailed list of the variables contained within the dataset is shown in Annex 3.2. Information is published in annual volumes, monitors, and on CD-ROM. In addition, within the ONS the National Cancer Intelligence Centre (NCIC) is responsible for data dissemination and secondary analysis of the data.

In addition to the data supplied to the national ONS cancer registration system there exist a number of parallel datasets relating to specific cancer conditions. These data originates in the first instance from regional registry data but, for historic reasons, are stored by various interested bodies. A list is provided in Annex 3.3.

4.4.2 Access to UK cancer data

Microdata on cancer registrations, containing other relevant personal and demographic details, are potentially available from two different sources. These are:

- Regional Cancer Registries;
- The ONS Longitudinal Study.

The regional registries are able, and have on previous occasions, supplied data to academics and medical researchers. Registries can supply *bona fide* researchers with patient level data for approved projects. Moreover, registries may be able to facilitate contact with cancer sufferers to provide further relevant information in order to support the relevant study (*e.g.* regarding occupation or lifestyle). Such contact would take place via the patient's general practitioner or hospital consultant, with the patient's full consent. Researchers pursuing this route will need to submit detailed project proposals to their regional registries (details of procedures differ from region to region). In addition, before commencing, the project must also be approved the relevant (regional) research ethics committee. Similar rules of confidentiality and disclosure in published statistics apply as with the HES data.

The alternative source of microdata on cancer is from the ONS LS. Whilst this source of data does not allow the benefit of flexibility potentially offered by the registry data, the main advantages of the dataset are (i) its large sample sizes, covering 1 percent of the UK

[33] The linking of the cancer data 'flag' to the LS is done via an ONS third part using the National Health Service Central Register (NHSCR).

population; and (ii) the fact that individual level data is linked over time from the 1971 through to the 2001 census. This allows a great deal of scope in terms of breadth of project design and also a high degree of statistical confidence.

Access to the LS data is granted via application to the LS Research Board (LSRB). The research proposal, which must provide an outline of the study and the requirement for LS data, must be approved by the LSRB, and undertakings signed. The data are delivered as a flat file containing relevant LS variables that are drawn from a relational database of all LS records held on a secure server at ONS. LS microdata can only be accessed in a secure environment at ONS[34]. A guide to cancer research using this data source can be found in Donkin and Hattersley (2001).

It is noted that in addition to the microdata, summary data of patterns and trends in cancer are available from the NCIC. Statistical summary reports can be accessed via the ONS *Statbase* webpage[35]. Additional summary data, including cancer data from HES, is also available through the National Cancer Services Analysis Team is an in-house provider of medical information services to the NHS[36].

4.4.3 Relevant cancer related research

Studies relating to incidence and causes of cancer using localised UK sample groups are too numerous to list. Articles relating to UK based empirical studies of cancer are regularly published in journals such as the *British Journal of Cancer, British Journal of Radiology* and the *British Medical Journal.*

With particular reference to the LS data, this has also has also been used variously over the past two decades to answer questions relating to incidence and causes of cancer. The dataset is particularly suited to studying both: (a) survival rates from various types of cancer, as well as (b) the link between incidence of cancer and socio-economic factors. It is to this end that most of the LS based studies have been undertaken. Table 7 provides a list of relevant recent cancer research using the LS.

[34] Details regarding the LS data and the application procedure can be obtained from the following web link: http://www.celsius.lshtm.ac.uk

[35] See: www.statistics.gov.uk

[36] http://www.canceruk.net/

Table 7: Recent cancer research undertaken using the ONS LS

- Inequalities in lung cancer mortality by the educational level in 10 European populations. See Mackenbach, Huisman and Andersen (2004);
- Living arrangements and place of death of older people with cancer in England and Wales: a record linkage study. See Grundy *et al.* (2004);
- Socio-economic and socio-demographic inequalities in cancer incidence and survival in the older population of England and Wales. See Sloggett (2004);
- A population-based case-control study for examining early life influences on geographical variation in adult mortality in England and Wales using stomach cancer and stroke as examples. See Maheswaran *et al.* (2002);
- Breast Cancer Survival in England and Wales: the Influence of Socio-Economic Status, Social Support and Parity. See Young (2002);
- Longitudinal study of socio-economic differences in the incidence of stomach, colorectal and pancreatic cancers. See Brown *et al.* (1998);
- The incidence of cancers among second generation Irish living in England and Wales. See Harding (1998);
- Cancer incidence among first generation Scottish, Irish, West Indian and South Asian migrants living in England and Wales. See Harding and Rosato (1999);
- Incidence of Health of the Nation cancers by social class. See Brown *et al.* (1997);
- Sources and uses of data on cancer among ethnic groups. See Harding (1996).

4.5 Other NHS data sources

Other datasets, in addition to those listed above, exist in within the NHS / DoH and generated by NHS Wide Clearing Service (NWCS). These are either smaller specialist datasets, for example relating to particular medical conditions, or are less well developed in terms of offering researchers a well trodden path to health microdata (as in the instances detailed so far). Additional relevant datasets, although not exhaustive, are listed below[37]:

- General Pharmaceutical Services dataset[38];
- Mental health minimum dataset[39];
- DOH Immunisation statistics[40];
- Diabetes Audit dataset and Diabetes Paediatric Audit dataset[41].

In addition several datasets exists tracking patients with cardiac and related illnesses. These are:

- UK Cardiac Surgery register;
- UK Heart Valve register;

[37] This list of other data sources is not exhaustive. Other datasets of a more minor nature exist within the NHS and are published by the IC, DH and other NHS and health organisations.

[38] See: http://www.ic.nhs.uk/psu

[39] Summary tables are available at: http://www.icservices.nhs.uk/mentalhealth/dataset

[40] For a summary of published statistics see also: http://www.ic.nhs.uk/pubs/

[41] For data sources see: http://www.library.nhs.uk/diabetes/. For a recent IC report on diabetes see http://www.ic.nhs.uk/news/press/pr190906b.

- Cardiac Ablation procedures;
- National Adult Cardiac Surgical database;
- National Audit Programme of the UK Association of Cardiothoracic Anaesthetists;
- National Implantable Cardioverter Defibrillator database;
- National Pacemaker database;
- British Heart Foundation Cardiac Rehabilitation database;
- National Stroke Audit dataset.

4.6 Modernising NHS information management

Important and ambitious initiatives are currently taking place in the collection, consolidation and provision of health data in the England, based on information available in National Health Service (NHS) care records. Related initiatives: the NHS 'National Programme for IT' and the establishment of the 'Secondary Uses Service (SUS)', both ongoing at the time of writing, provide the potential to revolutionise the way in which health data in England is collected and distributed. The programme is intended to bring together what are currently eclectic datasets, held locally, and of variable standard and quality. The data will be held centrally with consistent standards applied to data collection and provision.

4.6.1 National Programme for IT in the NHS

The NHS National Programme for IT (NPfIT) is being delivered under the auspices of the recently created NHS Connecting for Health organisation. This major project aims to introduce modern IT systems to link information on patient care in the NHS, replacing existing but somewhat antequated systems. By 2010, the NPfIT will provide an integrated data warehouse for holding patient records, connecting all 30,000 General Practitioner surgeries in England to almost 300 hospitals, transforming the way the NHS works and giving patients access to their personal health and care information.

4.6.2 Secondary Uses Service (SUS)

The Secondary Uses Service (SUS) is currently being established as part of the National Programme for IT (NPfIT). This replaces the existing NHS-Wide Clearing Service (NWCS) data exchange system which allows NHS organisations to exchange information with each other and the Department of Health[42]. The Secondary Uses Service (SUS) will protect patient confidentiality and will provide timely, anonymous patient data and other information for purposes other than direct clinical care.

The SUS is a core data warehousing infrastructure which, once completed, will establish a single, secure data environment for storage of pseudonymised, patient-based data with comprehensive coverage of records for the whole NHS. The resulting patient-based database will facilitate the following functions:

- Collection of incoming information from NHS Care Records (NHSCR);
- Facility to incorporate management and clinical information;

[42] At the time of writing the decommissioning of the NHS wide clearing service (NWCS) was not complete.

- Data management function – deriving and validating data and linking patient records on an ongoing basis;
- Quality, validation and consistency checks for data output;
- Anonymisation of data, where necessary;
- Data extraction, including tools and facilities to generate specific extracts of data;
- Data analysis and reporting.

Under SUS secondary users will therefore be able to carry out comprehensive analysis and reporting in a consistent and effective way. There will be increased ability for sharing information, particularly of aggregated data for comparative purposes. The potentially wide range of secondary uses will include both supporting internal functions such as health care planning, audit, performance improvement and benchmarking as well as providing a point of access to data for health related researchers.

The main responsibility for the supporting and implementing the SUS is the NHS Information Centre for Health and Social Care (IC). The IC also charged with responsibility for data quality and dissemination, including promoting SUS within the NHS and sharing 'best practice' or analytical solutions within the SUS user community.

4.6.3 National Health Service Care Records Service (NHSCRS)

A major change being implemented as part of the National Programme for IT (NPfIT) is the creation of the NHS Care Records Service (NHSCRS) which aims to replace the current paper based system and bring together information relating to individuals currently held at different locations (whether different buildings or different IT systems) into a central database for all England. The reasoning behind this is to ensure that patients' health records can be accessed instantly by NHS staff wherever and whenever they are needed, irrespective of location. Under the current timetable, the new system should be completed by 2015 at the latest.

Under the new system a national database will store a summary care record containing personal details (name, address, NHS number and date of birth) along with key health information such as allergies, major or continuing treatments, visits to A&E, and so on. In addition the central record will flag links to other more detailed personal health information stored and available at the local level(s). The key nationally available information, together with the more detailed local information, combine to produce the patient's complete Care Record[43]. This structure, known as the **NHS Spine**, will replace and add detail to the existing NHS Central Register. In addition, under the current plan, the spine will contain supporting information on patient prescriptions from the Electronic Transmission of Prescriptions (ETP) which will allow prescriptions generated by GPs to be transferred electronically from their surgeries to pharmacies.

4.6.4 New datasets

Part of the remit of the Information Centre is to provide accessible, standardised data. To this end the NHS Datasets Service has been commissioned, alongside the NfIT and SUS initiatives in order to drive forward this agenda. The service will, once fully established, provide national datasets to allow high quality standardised information to

[43] Access to this database will be confidential and restricted only to designated NHS staff. Individuals will have the right of access to their own records.

be transferred across the NHS, independent of source (in terms of the organisation or system that captures the base data).

Although currently at an embryonic stage, the new developments in the Secondary Users Service (SUS) and linking of patient level data, ongoing in the NHS, will give rise in due course to much more extensive and detailed micro health data. The data will be collected via a single source and collated using SUS but will be stored under separate (linkable) databases under a common data warehouse structure. Under the new system data which currently exists only at the local level will be integrated into the HES records to build up a much more complete picture of patient health. This development will give rise to new and exciting datasets which are currently in the process of being constructed. These datasets will replace and enhance existing data provision. Once completed these datasets will offer a portfolio of health based micro datasets comparable to that already available under HES. Once implemented the datasets will be available under the SUS, and will be available for researcher access. Table 8 provides a list of datasets which will be available in due course, with a brief description of the datasets.

4.7 Summary

Information collected on health, via the NHS central data collection service, provides an excellent source of microdata for researchers, whether from the science, medical or social science community. Use of this data is already well developed in some areas of research and the Hospital Episode Statistics (HES) and the Cancer Registration Data both provide good examples of well developed and documented datasets which have provided a source of much valuable research. However, in other areas, data are less well developed and data availability and access is either limited, non existent or done only through *ad hoc* arrangements or priveleged access to localised datasets. The reason for this is that much of the NHS data has, for historical reasons, been stored locally (in local NHS establishments) with data storage, standards, and IT systems variable. The General Practice records data provide an excellent example of this.

However, the NHS is in the process of embarking on a data revolution which, once completed, will provide high quality standardised microdata, generated by from a central source know as the DoH Spine. The spine will collect and link patient level records from birth, documenting through an extensive data warehouse facility, all patient contact with the NHS. The current Connecting for Health, NHS National Strategy for IT, and the Secondary Users Service are all initiatives working to this end. In this respect the next decade promises a new era in the provision and availability of health data for the research community.

Table 8: New SUS datasets under construction

CANCER DATASET
The datasets cover head and neck, lung, colorectal, breast, urology, gynae and upper gi, brain, sarcoma and skin cancer. The datasets will provide standardised and consistent data collection in all relevant NHS organisations and are intended to provide essential amount of information in order to follow individual patient journey through cancer treatment.

CHILD HEALTH DATASET
The national dataset for child health will support the National Service Framework for Children, Young People, and Maternity Services and Children's & Maternity Service Information Strategy, 2004. The purpose of the dataset is to provide an NHS standard for data collection in order to support child health services.

CHILD AND ADOLESCENT MENTAL HEALTH SERVICES DATASET
Complementing the Child (CAMHS) dataset and developed within the Framework for Children, Young People, and Maternity Services and Children's & Maternity Service Information Strategy, 2004. [44]

CORONARY HEART DISEASE (CHD) DATASET(S)
The CHD Datasets project was initiated in July 2001 to support the information requirements of the National Service Framework for coronary heart disease. The dataset is intended to support four areas of CHD, namely: Acute Myocardial Infarction (AMI); Paediatric Cardiac Surgery; Angioplasty; and Adult Cardiac Surgery.

DIABETES DATASET(S)
This dataset has been initiated to develop a health management dataset suitable for use by people with diabetes and their care providers. A core dataset for Diabetes (the Diabetes Continuing Care Reference dataset, DCCR), has been developed and is currently being pilot tested.

MATERNITY DATASET
Complementing the Child (CAMHS) dataset and developed within the Framework for Children, Young People, and Maternity Services and Children's & Maternity Service Information Strategy, 2004.

MENTAL HEALTH DATASET
This dataset is intended to upgrade and enhance the Mental Health Minimum Dataset (MHMDS) which is already in existence. The dataset is a nationally defined framework of data on adult mental health patients, including older people. All providers of specialist mental health services for adults and older adults are mandated to collect the MHMDS.

OLDER PEOPLE DATASET(S)
Under development, these datasets will include and monitor incidence of: stroke, falls, dementia, continence. It will also link in due course to the cancer dataset.

RENAL DATASET
A project is now underway to develop a dataset to support the National Service Framework for Renal Services. Following testing and design iterations of the dataset it is hoped to have a finalised version during 2007.

[44] See http://www.camhoutcomeresearch.org.uk/NationalCAMHSDataset.asp

5. Business related microdata

5.1. Introduction

The Office for National Statistics (ONS) collects a large amount of microdata relating to business entities. Such data are collected routinely as part of the ONS monitoring of business related activity in the UK and is routinely used to produce detailed industry and small area statistics. Similarly collection of business related data also facilitates regular publication of various measures of business related activity, such as those relating to production of good and services, investment in capital, foreign direct investment, and so on. Data are linked through time to create a panel of establishment level data, provides a valuable source for academic researchers.

Above and beyond collecting these data for national statistical purposes, during the past few years the ONS has made great strides towards sharing this data with the academic and wider government community. The Business Data Linking Project was started in 2001 with funding under the Treasury's Evidence Based Policy Fund. Early successes were studies on manufacturing productivity and unpicking the multinational effects in comparative studies. However, while the value of providing access to this restricted micro data was evident, the nature of the funding and projects was necessarily short-term. At the end of 2002, ONS began a long-term strategy to develop a permanent accessible resource for the research community as a whole. The Virtual Micro data Laboratory (VML) was launched at ONS in January 2004 to provide a secure facility for access to confidential business micro data across all ONS sites. These moves have given a significant boost to microeconomic research in the UK[45].

5.2 Inter-Departmental Business Register (IDBR)

The starting point and building block for the provision of ONS business related data is the Inter-Departmental Business Register (IDBR). The IDBR is a comprehensive list of UK businesses that is used by government for statistical purposes. Essentially, this includes all UK businesses except those run by sole traders and falling below VAT threshold and some non-profit making organisations. Currently the IDBR covers approximately 2.1 million businesses, covering approximately two thirds of businesses and 99 percent coverage of UK's official business activity.

The IDBR's primary function is to provide a sampling frame for surveys of businesses carried out by the ONS or other government departments. In addition, however, since each business has its own unique reference number in the IDBR, establishment level data can be linked either over time, as via the Business Structure Database (BSD), or linked to other relevant business datasets. The IDBR is constructed by identifying business entities based on administrative data collected from three separate sources:

- Her Majesty's Revenue and Customs (HMRC) data on traders registered for Value Added Tax (VAT) purposes;
- Her Majesty's Revenue and Customs (HMRC) data on employers operating a Pay As You Earn (PAYE) scheme for employees;
- Companies House data on registered businesses.

[45] For further background information on the Business Data Linking (BDL) project see: http://www.statistics.gov.uk/about/bdl/

The data collected from these sources is reconciled with and added to the ONS Business Register Survey on an annual basis and incorporated into the IDBR database. The variables contained in the IDBR relate to basic information about the organisation (including name, address and other contact details), details regarding its legal status and ownership, as well as basic information regarding employment and turnover.

The IDBR assigns a unique ID reference number relating to establishments at various levels of business activity. Identified firms are sent an IDBR reporting form. The address to which the form is sent (generally the head office of the firm) is referred to as the **reporting unit**. The form covers the enterprise as a whole and also parts of the enterprise identified by lists of local units. Beyond this the IDBR also identifies the **enterprise unit** and **local unit**. A group of legal units under common ownership is called an enterprise group. The enterprise code indicator assigns local units (and establishments) to a common owner. An individual site (plant or office operating at a single location) in an enterprise is called a local unit[46].

5.3 Linked business datasets

5.3.1 An overview of relevant datasets

The availability of business microdata comes from various ONS administrative data sources, namely:

- Inter-Departmental Business Register (IDBR);
- Business Structure Database (BSD);
- Annual Business Inquiry (ABI);
- Annual Respondents Database (ARD).

All of these data sources provide data, either directly or indirectly, which once linked over time provide information that allows firms' existence, growth, survival/ demise to be tracked, providing a potentially invaluable source of information for economic researchers.

The inter-relationship between these datasets is somewhat complex, but illustrated in Figure 5. As described above, the Inter-Departmental Business Register (IDBR) provides the starting point for ONS data collection. As an exhaustive list of businesses it yields a sampling frame for ONS survey of businesses in the UK. The IDBR is a strictly controlled and confidential dataset containing names and addresses of businesses along with some basic details relating to the characteristics of the firm. However, IDBR is available to researchers in its own right, suitably anonymised in the form of the Business Structure Database (BSD). The BSD is, in essence, is a reduced longitudinal form of the IBDR which links establishments over time and provides basic demographic details relating to the firm. This dataset is summarised in section 5.3.2 of the report.

Running in parallel to these datasets is the Annual Business Inquiry (ABI). The ABI is a compulsory survey undertaken by the ONS which records information from UK firms. The survey exists in two parts (later integrated), recording (a) employment details of firms and (b) financial and accounting information from businesses. The amount of

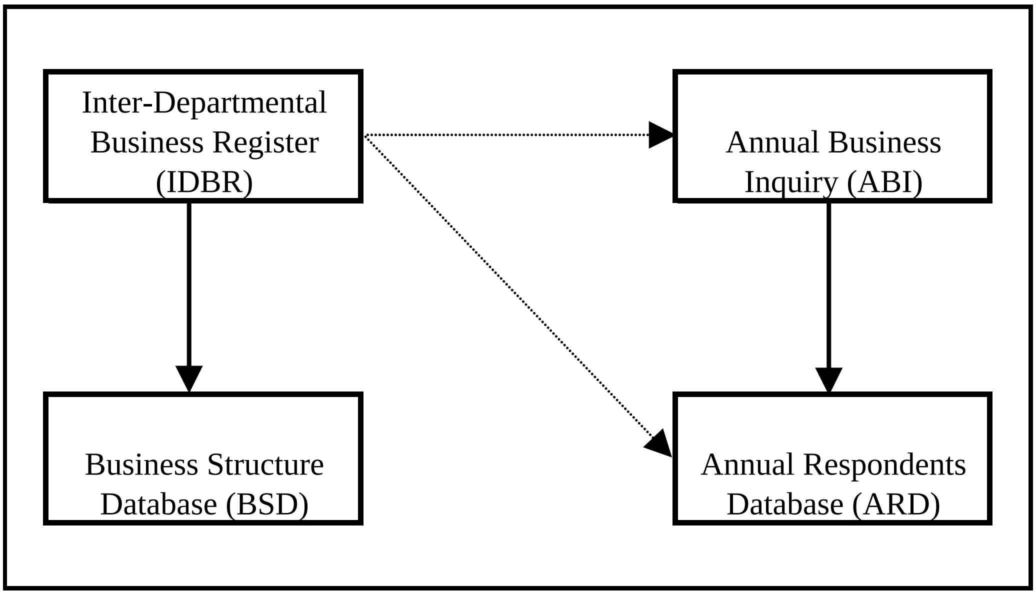
[46] A detailed discussion of how ONS maintains information on the structure of enterprises and local units is provided in Criscuolo et al (2003).

information and level of detail provided by the survey goes beyond the basic information provided by the IDBR. However, the ABI uses the IDBR as a sampling frame for conducting the survey. The survey covers most industry sectors of the economy (the ABI incorporates distribution and service sectors after 1997) with all large firms surveyed annually and smaller firms less frequently[47].

Finally, the Annual Respondents Database (ARD) provides a reduced form version of the ABI which links business data over time, annually, providing much valuable additional information relating to business establishments beyond that available in the BSD. It is noted, however, that for smaller establishments information from the ABI is generally not available on an annual basis. In this case the ARD is supplemented and made exhaustive (mapping back to the universe of firms as defined by the IDBR) using supplementary, but more basic, information from the IDBR.

Note that whereas access to the ABI is restricted to internal ONS use, the ARD is made available for use to researcher via the Virtual Microdata Laboratory (VML). Details information regarding the ARD is provided in section 5.3.3 of this section of the report.

Figure 5: Linked business datasets


Inter-Departmental Business Register (IDBR)
Annual Business Inquiry (ABI)
Business Structure Database (BSD)
Annual Respondents Database (ARD)

5.3.2 Business Structure Database (BSD)

The fact that the IDBR provides an annual snapshot of business activity with businesses identified by unique reference numbers means that observations can be linked over time, where firms continue to exist. Alternatively, where firms cease to exist or change form, for example via merger or acquisition, these events can be recorded appropriately. Following recent developments this information is now held in the Business Structure Database (BSD) which is a derived dataset based on annual observations of the IDBR. A prototype version of the data has recently been released at the time of writing.

BSD is currently available to researchers in a flat file format from 1997 – 2005, with datasets available (separately) at enterprise and local unit level data. The dataset

[47] For further details regarding the ABI see: Jones (2000).

summarises demographic information regarding the firm including information on the date of 'birth' and 'death' (where applicable) of the firm, as well as identifying restructuring event by which there are changes in the nature of business enterprises, including local unit transfer. This information provides an exhaustive and annually updated dataset for studying the universe of business enterprises. This provides a resource for studying firm demographics, such as firm birth, death and survival, and takeover and merger activity. The main variables contained in the BSD are listed below[48]:

- Unique IDBR Reference number;
- Relationship to enterprise / local unit;
- Employment and employees;
- Turnover;
- Legal status;
- Country of ownership;
- Standard industrial classification (derived);
- Demographic event identifiers;
- Date of birth of establishment;
- Date of death of establishment (where applicable).

5.3.3 Annual Respondents Database (ARD)

The Annual Respondents Database (ARD) was formed from a number of ONS business surveys to provide a longitudinal database for micro-data research. The Annual Respondents Database (ARD) is richer in detail than the BSD and contains linked business data over the longest available period (since 1970 for the productive sector, and since 1997 for service sector establishments). This, and the fact that it predates the BSD in terms of development, make it the currently the primary source of longitudinal studies relating to UK firms. Much of the existing research using establishment level microdata (reviewed later) has been undertaken using the ARD.

The ARD contains establishment level micro-data collected by the Office for National Statistics (ONS) in the Annual Business Inquiry (ABI)[49]. The ABI is collected in two parts which are integrated and included in the ARD, i.e. ABI1: an employment record, and ABI2: relating to financial information. Each year a stratified sample is drawn for the ABI and stored in the ARD, with establishment data converted into a single consistent format linked by the IDBR references over time. The linked dataset covers all sectors of the economy since 1997, with a subset of the dataset covering only the production and construction sectors of the economy from 1970 – 2005.[50] The dataset rescales, recombines and manipulates variables so that each variable is consistently defined (as far as possible) each year.

Currently the ARD contains over 80,000 business entities. Data are not available in the ABI on an annual basis for smaller firms (with fewer than 20 employees) and in most cases smaller firms receive a 'short form' of the ABI. Small firms (*i.e.* those employing

[48] For a detailed guide to BSD see Davies (2006a).

[49] Prior to 1998 information collected via the ABI was collected from the Annual Census of Production (ACOP) and the Annual Census of Construction (ACOC). These surveys have now been incorporated into the ABI.

[50] Work is currently under way to include past service sector data into the ARD from 1994 to 1997.

less than 20 people) complete a survey every 3 years rather than each year. Note that prior to 1995, the very smallest businesses were not sampled at all. In these cases the ARD is directly supplemented wherever possible with information from the IDBR. In addition for some variables the ONS uses imputation from similar donors to generate figures for companies who only receive 'short forms'.

Although in principle linking individual businesses through time based on IDBR reference numbers is possible, in practical terms many inconsistencies arise in the data. For example, longitudinal linking may return missing values for firms since firms selected for ABI survey change year to year. Similarly, there are problems of reclassifications over time, for example based on changes in size bands and industrial classification (from SIC 80 to SIC 92). Consequently, since late 2005 Business Data Linking (BDL) branch at ONS has been undertaking a restructuring of the dataset to form a cleaner, simpler dataset.

Currently the ARD is stored in the VML in three forms:

- **ARD master file** – the comprehensive dataset covering: 1973-2003 (production); 1994-2003 (construction); 1997-2003 (services);
- **ARD standard variables file** – covers a subset of variables standardised over all years; 1980-2004;
- **ARD industry panel** - an ARD derived panel dataset with consistent industrial classifications over time, 1980-2004.

The data now exists in full form and as a slimmed down a subset of the variables with a standard set of names and variables for as many years as possible. The list of variables contained in these datasets is similar to those in BSD listed above, but also includes various measures of inputs, outputs, costs, investment expenditure and value added[51]. The inclusion of these variables facilitates research on productivity, as detailed in section 5.7.

5.3.4 Practical considerations in using Annual Respondents Database (ARD)[52]

With respect to practicalities, the ARD is not the most user-friendly datasets. Although a cleaned, standardised version of the dataset now exists, there are several issues which should be borne in mind when using the ARD data. These are now well documented as a result, by and large, from the feedback received from early users of the dataset. Particular issues of concern revolve around (i) consistent treatment of business units; and (ii) measuring capital stock. These issues are considered in turn below.

(i) Treatment of business units

In the ARD establishments are linked based on enterprises (or 'reporting units'). Potential problems arise from this, in using the data, since reporting units are not synonymous with 'plants' or local operating units (such as offices, outlets or shops). Whilst in the vast majority of cases for small and medium sized enterprises, enterprises consistent of single units (plants), this is not the case for large firms. For large firms a

[51] More details regarding the ARD dataset, including new data developments, can be found in Robjohns (2006).

[52] This sub-section draws upon the following papers: Barnes and Martin (2002) and Harris R (2002).

reporting unit will generally contain more than one local unit (plant). Consequently researchers, whose main focus is often at plant level, will encounter the following problems:

- Changes in the composition of the enterprise through local units closing or changing ownership may distort the data from enterprises;
- Local unit cessation (death) and disappearance due business restructuring cannot be distinguished at plant level[53];
- Financial data is not available at plant level, only at enterprise level.

(ii) Measuring the capital stock

Much of the applied work using the ARD relates to productivity (see section 5). Whilst measures of labour productivity can be readily constructed from observations of output and employment, measuring total factor productivity of firms (TFP) is more problematic since the ARD has no capital stock information. Ways around this have been found based on inferring the value of capital stock, although this is less than satisfactory. This is generally done using ARD on firms' capital expenditure, for example using the Perpetual Inventory Method of calculation, following Martin (2002). To this end the ARD contains imputed measures of capital stock for each reporting unit in each year 1980-2004.

5.4 Other business microdata

An important dimension of the business data sets held within the VML is the availability of unique IDBR reference numbers that allow for the linking of records across surveys. As the largest business survey, the ABI generally forms the 'spine' of data linking, with other smaller surveys being matched to it. As well as the linked data from ABI / IDBR the ONS Virtual Micro Laboratory (VML) also stores numerous other micro-datasets relating to information collected from businesses. In particular, there are dataset relating to earnings; workplace industrial relations; production and capital expenditure; along with a somewhat eclectic collection of business level survey constructed from samples based on the IDBR sampling frame. These datasets are described in detail in this section of the report.

5.4.1 Annual Survey of Hours and Earnings (ASHE) data

The ONS's Virtual Micro Laboratory (VML) stores earnings micro data collected from UK businesses. The relevant earnings datasets are the Annual Survey of Hours and Earnings (ASHE) available from 2004-05. This replaces the New Earnings Survey (NES) data which exists from 1986 – 2003. Both datasets also exist in panel data form (*i.e.* tracking individuals over time) for a limited set of variables.

Annual Survey of Hours and Earnings (ASHE) is an employer survey which collects individual level information on hours and earnings of employees in the UK. These data exist from 2004 onwards and replace the New Earnings Survey (NES) as ONS' main

[53] Note that such distinctions can be made using BSD.

source of information on earnings[54]. ASHE collects information on earnings based on a 1 percent sample of Pay as You Earn (PAYE) pay records from the Inland Revenue. Samples are obtained based on all jobs in which an employee's National Insurance number (NINO) ends with a specified pair of digits. The sample is taken in April of the relevant year and relates to employment over the previous year. The survey provides a large amount of information on earnings and hours including the composition of pay, bonuses, overtime, etc. Variables contained in the ASHE/NES datasets include:

- Earnings, including bonuses, overtime, etc;
- Occupation;
- Industry (derived from the IDBR);
- Hours worked;
- Sex;
- Age;
- Place of work;
- Job tenure.

One feature of ASHE is that individual observations are now assigned a weight, so that pay information can be grossed up to the national picture for purposes of reporting. The individual weights are calculated based on a calibration with the Labour Force Survey (LFS) on a grid defined by a cross-classification of occupation, sex, age and workplace region. The NES datasets does not contain any weights. One potential problem with ASHE (similarly NES) is that due to the way in which the data are collected it does not capture very low earners, i.e. not registered in a PAYE scheme because they are earning below the tax threshold[55].

The ASHE data is also stored in panel data form, in the Annual Survey of Hours and Earnings Panel Dataset (ASHEPD). This dataset links records of earnings from ASHE based on linked NINOs so that individuals can be tracked over time. The data is made up of an anonymised and reduced form of the full set of variables. The ASHEPD contains annual observations from 1975 - 2004 and contains information relating to approximately 170,000 individuals. This longitudinal (panel) data is particularly valuable for analysis of earnings data, for example in implementing so called fixed or random effects regression models.

In addition to data stored in ASHE / ASHEPD the VML retains data from the New Earnings Survey (NES) which preceded it. This survey similarly collected information on earnings in the same manner based on a 1% sample of all employees based on National Insurance numbers. These data are available annually from 1986-2003. The panel dataset, New Earnings Survey Panel Dataset (NESPD) is also available, derived from NES based on linked NES records from 1975-2003.

[54] Whilst the core data collected by the new questionnaire is essentially the same, a four page ASHE questionnaire replaces the older NES two side questionnaire. According to the ONS "The new questionnaire improves the layout, routing, wording and definitions used and will lead to more consistent responses improving the quality of the data collected. The questionnaire is also easier to understand and to navigate and so reduces the time taken by users to complete."

[55] For more details regarding the ASHE data, including details of weighting see: Bird (2004).

5.4.2 Workplace Employment Relations Survey (WERS 2004)

The Workplace Employment Relations Survey, 2004 (WERS2004) is a national survey of people at work in order to provide large-scale evidence about a broad range of industrial relations in Great Britain[56]. The sample for the WERS 2004 Cross-Section Survey was taken from the IDBR. The scope of the WERS 2004 was to cover all workplaces with 5 or more employees, located in Great Britain (England, Scotland and Wales) and engaged in activities within Sections D (Manufacturing) to O (Other Community, Social and Personal Services) of the Standard Industrial Classification (2003). The survey covers both private and public sectors. The Cross-Section Survey covered 30 per cent of all establishments in Britain (a total of 697,000 establishments).

WERS collected information from managers, trade union or employee representatives, and employees themselves (with responses available as separate cross sectional surveys: *i.e.* the cross-section survey of managers, cross-section survey of employee representatives, and cross-section survey of employees. Questions in the survey vary across the three surveys and are summarised by broad topic in Annex 4.1. In addition to these questions, WERS 2004 also collected financial data via the 'Financial Performance Questionnaire'. Whilst the rest of the WERS data set can be accessed via the UK Data Archive, this part of WERS can currently only be accessed via ONS VML.[57]

In addition, the version of the WERS data held by ONS has also been linked to the ABI. This serves two purposes. Firstly, by linking WERS to ABI data, the accuracy of financial information supplied by WERS respondents can be validated. Secondly, not all respondents to WERS completed an FPQ. By linking to the ABI, it has been possible to 'fill in the gaps' for those organisations who did not respond to the FPQ. Unlike the FPQ which will subsequently be made available through the UKDA, the matched WERS/ABI data will only be available from the VML.

WERS 2004 includes both a cross-section and panel element. The panel element for 2004 forms Wave 2 of the 1998-2004 panel survey, with wave 1 comprised the cross-sectional managers' survey conducted for WERS 98. A total of 956 establishments first contacted in 1998 participated in the 2004 panel.

5.4.3 Surveys of business production and capital expenditure

The ONS Virtual Microdata Laboratory (VML) also stores microdata on production and business expenditure on capital. These datasets are:

(i) Monthly Production Inquiry (MPI);
(ii) Business Spending on Capital Items (BSCI) survey;
(iii) Quarterly Capital Expenditure Inquiry (QCES).

A short description of each dataset is provided below.

[56] WERS 2004 replaces the original Workplace Industrial Relations Survey. For more details regarding WERS see: http://www.wers2004.info/

[57] FPQ will be deposited with the UKDA in Spring 2007.

(i) Monthly Production Inquiry (MPI)

Monthly Production Inquiry (MPI) is available monthly from January 2000 - December 2004. The MPI is a statutory survey with a sample of 9,000 businesses in production industries, drawn at the 4 digit level of the Standard Industrial Classification 2003 (SIC). The MPI collects data on total turnover, export turnover, employees, sold goods, plus, for the engineering industries, orders-on-hand, export orders-on-hand, new orders and new export orders. The Monthly Production Inquiry (MPI) is a primary input to the Index of Production (IoP) and the Workforce Job estimates. It is a main indicator of short-term changes in economic activity and forms a significant component of the output measure of GDP.

(ii) Business Spending on Capital Items (BSCI) survey

The BSCI survey is available annually from 1998-2004. The BSCI is a survey of large businesses (employing 100 or more employees) and monitors total capital expenditure (acquisitions and disposals) for: (a) Vehicles and other transport equipment; (b) Land and buildings and other construction work; (c) Purchases of services associated with capital goods; (d) Intangible produced assets; and (e) Other capital expenditure (equipment). The survey uses the IDBR as a sampling frame.

(iii) Quarterly Capital Expenditure Survey (QCES)

The QCES is available quarterly from 2001Q1-2005Q4. The survey collected information on capital expenditure for various industry groups by asset type, based on a statutory survey with a total sample size of 32,000 covering the private sector of part of A and most of sections C to O of the Standard Industrial Classification of Economic (SIC). Data are collected on capital expenditure in the following categories: land and buildings, vehicles, computer hardware and software, other capital equipment, exploration expenditure. Results from the survey provide essential information for the National Accounts and feed into the compilation of gross capital formation.

5.4.4 Other business surveys

The ONS Virtual Microdata Laboratory (VML) also stores microdata from a mix of other surveys. These datasets are detailed below:

(iv) E-Commerce Survey;
(v) Community Innovation Survey (CIS);
(vi) Annual Inquiry into Foreign Direct Investment (AFDI);
(vii) Survey into Business Enterprise Research and Development (BERD)

A short description of each dataset follows:

(iv) E- Commerce Survey

The E-Commerce Inquiry is a survey of access to information and communication technologies (ICT) conducted annually from 2000. The latest available data relates to 2005. The early surveys were experimental in their design and are regarded as being of relatively poor quality. The survey has also been subject to significant changes year on year. Surveys from 2002 onwards exhibit the highest levels of continuity. The survey covers: businesses' use, and plans for use, of ICT; use of internet to place and receive orders; use of electronic networks other than the internet; use of any electronic network to make or receive payments; and the integration of business processes.

(v) Community Innovation Survey (CIS)[58]

Data from the Community Innovation Survey is available for 1994-1996 (wave 2) 1998-2000 (wave 3) and 2002-2004 (wave 4). The CIS is a voluntary survey looking into a number of aspects of innovation within UK firms, based on a Eurostat core questionnaire. The CIS 2 and CIS 3 cover three-year periods each, covering a stratified sample of firms with more than 10 employees, drawn from the IDBR. The areas it covers are: all production and construction, wholesale trade (exc. Motor vehicles), transport, storage, communication, financial intermediation, and business services. The surveys cover approximately 2,300 (wave 2), 8,100 (wave 3) and 16,400 (wave 4) establishments, respectively.

(vi) Annual Inquiry into Foreign Direct Investment (AFDI)

AFDI data is available from 1996-2004. The AFDI collects information on the ownership of UK firms and of the subsidiaries abroad owned by UK firms. The register from which the firms are sampled comes from sources including HM Customs & Excise, the Inland Revenue, Dun & Bradstreet's "Worldbase" system, and ONS inquiries on acquisitions and mergers. Information is collected on financial flows for the firms on the register, in the form of foreign direct investment done abroad by UK firms and that done in the UK by foreign firms. A panel of the AFDI files has been created containing information covering foreign ownership and affiliates. This could be used to answer basic questions about multinationals in the UK.

(vii) Survey into Business Enterprise Research and Development (BERD)

The BERD survey is available annually from 1996-2003. BERD is an annual survey designed to measure Research and Development (R&D) expenditure in the UK. The survey identifies firms engaged in R&D using information from the ABI and other DTI and Scottish executive information on firms engaged in R&D. Firms are survey based on a sampling frame covering industry and firm size. The largest firms receive a 'long form' survey questionnaire whilst smaller firms receive a 'short form'. The survey gathers information on R & D expenditure broken down into areas of basic research, applied research and experimental research. Capital expenditure related to R&D is also recorded, broken down into land and buildings and plant and machinery. The survey also seeks information on the sources of funds for expenditure on R&D. Finally, the survey also seeks information on the breakdown of average employment of scientists and engineers, technicians and related staff engaged in R&D related work.

[58] For a summary of the results of the 2001 survey see: Stockdale (2003).

5.5 Linking business microdata

5.5.1 Linking IDBR based datasets

The survey microdata contained in the ONS Virtual Micro Laboratory (VML) which uses an IDBR sampling frame is, in principle, linkable with any other IDBR based dataset. This applies to the vast majority of the datasets listed in the previous section, for which a unique enterprise IDBR number is present in the dataset. Further, even if an IDBR reference number is not available in the survey, the data can still be matched in principle to other datasets by tracing back establishments to the original IDBR sampling frame using name of establishment, as address, postcode, etc and/or applying methods of 'fuzzy matching'.

Initiatives of linking of data have already taken place or are presently ongoing at ONS. These include efforts to link the Annual Respondent's Database (ARD) to:

- WERS 2004;[59]
- ASHE/NESS earnings data;
- ESS;
- Community Innovation Survey;
- Inquiry into Foreign Direct Investment;
- E-commerce survey.

In this respect the whole suite of datasets contained in the VML create an opportunity for researchers to build new datasets, for example linking employer with employee data. Moreover, efforts have been made to link the ARD with sources held elsewhere in the ONS and externally. (Examples of such initiatives are listed in the next section of the report). The ability to do this kind of matching opens up new avenues of research relating for example to consequences of policy interventions (see reference to the DTI programs in section 5.5.2. of the report).

Finally, however, it is important to sound a note of caution about linking datasets via IDBR or, as is most common, onto the ARD. Such linking can be problematic in two respects. Firstly the matching may not be exact and many observations may be lost since in practical terms unique IDBR flags may not exist in both dataset. This can happen for a number of reasons and may be due to companies ceasing to exist or changing form, or may be due to administrative-related issues such as missing or wrongly specified information in one or both of the datasets. These issues are evidence, for example, by Davies (2006) when matching WERS to ARD and Harris (2001) when matching DTI datasets to the ARD.

Secondly, once a linked dataset has been created there are general statistical concerns relating to the merged dataset[60]. Such concerns may arise due to the nature of the underlying sampling frame used by ONS in the first instance to construct the datasets. Equally statistical issues may arise due to biases introduced by the loss of observations when linking datasets.

[59] For details of this see: Davies (2006b).

[60] For a summary of statistical issues see: Chesher and Neisham (2004).

In terms of practicalities when merging ONS microdata, linked datasets will tend to be dominated by large firms (because the stratified sampling design means that most information comes from large firms the linking process tends to increase the joint probability of selection). Further biases might be introducing in terms of industry selection, etc. since ONS datasets tend to use complex sampling schemes which once linked might create unrepresentative samples. It is not altogether clear at this point how to deal with these issues (in terms of application of appropriate weights, etc.) and this is currently receiving attention at ONS.

5.5.2 Examples of cross linking of business datasets

Table 9 lists examples of work undertaken linking of datasets using firm-level data. Although not exhaustive, the list contains references to recent reports and working papers. Note that much of this work is still ongoing and final published work is not yet available.

Table 9: Linking of business datasets

- ARD and ESS – Galindo-Rueda and Haskel (2005);
- ARD and WERS – see Haskel (2006);
- ARD and ASHE – see Lam *et al.* (2006);
- ARD and DTI datasets (under privileged access) – see Harris and Robinson (2001);
- ARD and ESS – see Harris *et al.* (2005);
- ARD and NESPD – see Hijzen *et al.* (2005);
- ARD and UK Quarterly Fuels Inquiry (QFI) – see Martin (2005).

5.6 Access arrangements to ONS microdata

5.6.1 Access

The ONS is charged with a responsibility of keeping business data confidential under the 1947 Statistics of Trade Act. Under this legislation, business are legally obliged to complete statutory business surveys conducted by ONS. However, these businesses are given assurances this that this information will not be released to third parties outside of ONS or Other Government Departments who have the authority to access ONS data. Balanced against this is the need to share data through best practice and provide appropriate resources for the academic and wider government research community.

Access to the datasets is therefore restricted. The data is not made freely available and physical access to the datasets is provided only through the Virtual Micro Laboratory (VML) which currently operates at ONS sites at Newport in Wales, Drummond Gate in London, Titchfield on the south coast and Southport in the North West Access to the data for research purposes is permitted through a policy of 'safe setting - safe people'.

'Safe setting' refers to the VML research environment where the unauthorised disclosure of information is prevented, as far as possible, by the implementation of appropriate restrictions and procedures. In this respect access to the data is only available via the secure VML, where academic researchers can carry out statistical analyses. This data is confidential, therefore access is tightly restricted.

'Safe people' refers to individuals approved by ONS and who are deemed to have a valid reason to use the data. ONS policy states that: "only researchers fully employed at bona fide academic or charitable research institutes, or civil servants, may have access."

Researchers wishing to access the data must submit an application to the ONS to use the relevant dataset(s). The applicants must convince the ONS Micro-Data Release Panel (MRP) that their reason for wanting access to the data is related to a viable and relevant research topic (often undertaken on behalf of another government department) and that they have an ability to undertake the research, where work undertaken will be of a high quality and will be deemed to add significant value to existing research in that area. Researchers must submit a research proposal to BDL, along with a timetable for research, and will normally receive a reply within four weeks[61].

Once a request for access has been processed and approved the individual must sign a contract which undertakes to ensure confidentiality and ensure that data is not disclosed to third parties. A further agreement must also be signed by the researchers employing organisation confirming that the applicant is a trusted employee. The researcher must also undergo an initial one day BDL training session at the ONS which covers aspects of confidentiality and disclosure[62].

5.6.2 Disclosure of information[63]

A major consideration for the ONS, in its operation of the VML, is to minimise risk of disclosure of confidential information[64]. In order to achieve this, the ONS imposes a set of rules and limitations on research outputs from BDL datasets. This is done by imposing a set of **disclosure tests** on output which are designed to protect the identity of individual firm by controlling the level of aggregation at which data output can be shown and reducing the risk of disclosure by deduction from other sources to nil. ONS applies two rules to output in the form of data tabulations to avoid *primary disclosure* (i.e. the identification of individual data from the contents of a cell). These rules are:

- There must be at least three enterprise groups in a cell;
- The sum of the data for of all reporting units excluding the largest two must be at least ten per cent of the value of the largest one.

In addition to these considerations there is risk of s*econdary disclosure* by which data values might be inferred from a table (or series of tables) by simple ex-post arithmetic procedures. Possible examples of this include disclosure by differencing (inferring values through subtracting elements of he data across rows or columns) or disclosure by regression (possible under certain regression designs, for example including dummy variables)[65].

[61] It is noted that there is a cost of using the lab, payable by the researcher on a daily basis.

[62] Details of the procedures in applying for data can be found via the following weblink: http://www.statistics.gov.uk/about/bdl/mag_map.asp

[63] For a fuller discussion of the issues relating to disclosure see: Ritchie (2005).

[64] ONS is, naturally, very risk adverse with respect to potential disclosure of confidential information as unwarranted disclosure of such information could severely damage ONS' credibility and capability for future information gathering.

[65] Note regression results which do not report all parameters are non-disclosive for all practical purposes.

In order to control for these disclosure risks the BDL team at the ONS check all research output derived from the datasets. Due to the complex nature of much analysis undertaken within the VML, automated disclosure checking would not be feasible. Outputs from the lab are limited to data tabulations and regression results. Researchers may only have access to **approved output**, once it has been checked and cleared. Note that no transfer of data out of the server is allowed or possible, with printers, disk drives, email and internet access all disabled. Any output which the researcher wishes to remove (for discussion with colleagues or for publication) is controlled by ONS staffs, who check the output for disclosure before it can be released. Under the agreement signed between the researcher, their employer and the ONS, any researcher found breaching disclosure rules will be subject to disciplinary procedures, with potential strict sanctions[66].

5.7 Relevant research

Interest in the BDL datasets and awareness of their existence is currently growing as the VML lab itself expands and more and better data becomes available. The BDL datasets have been used during the past five years or so by researchers who have managed to publish high quality academic work based aided by analysis of the datasets. Much of the work done in this respect has been undertaken by a small community of interested parties. Notably these include prominent individuals of groups of researchers, including those at the Centre for Research into Business Activity (CERIBA)[67] led by Jonathan Haskel, or by individual researchers such as Richard Harris at the University of Glasgow[68].

Most of the research up to this point has focussed on utilising the **ARD** dataset to provide answers to interesting research topics in the area of applied industrial and labour economics. A list of prominent areas of research, although not exhaustive, which utilise the linked business data include:

- Understanding entry, survival and exit;
- Understanding job creation and job destruction;
- Understanding merger and takeover activity;
- Understanding productivity differentials between firms;
- Understanding productivity and skills ;
- Understanding productivity and growth ;
- Understanding the effects of foreign direct investment;
- Testing policy interventions[69].

Table 10 lists examples of research using the ARD, organised by broad topic areas. In addition to this there have been fruitful research utilising earnings microdata held by

[66] All outputs submitted for removal are recorded and archived, whether they are released or not. Approved outputs are recorded in a separate database. There is thus a complete audit trail of approved outputs

[67] For a link to research see: http://www.ceriba.org.uk/

[68] For a link to research see: http://www.gla.ac.uk/departments/economics/staff/harris_richard.html

[69] As an example of work in this area, the DTI have sponsored a number of projects that have focussed upon programme evaluation. Schemes evaluated have included Regional Selective Assistance (RSA) given to small firms to provide funds for capital investment. By combining administrative information about who has received these policies with ONS data (particularly panel data), researchers have tried to identify the effects of these schemes on financial performance and employment measures. See Harris and Robinson (2001).

BDL as well research using some of the lesser known datasets. Table 11 lists examples of research using earnings data, whilst Table 12 lists examples research using other datasets.

Table 10: Examples of research utilising ARD

ENTRY, EXIT AND SURVIVAL

- Entry, exit and establishment survival in UK manufacturing. See Disney, Haskel and Heden (2003);
- Job Creation, Job Destruction and the Contribution of Small Businesses: Evidence for UK Manufacturing. See Barnes and Haskel (2002).

PRODUCTIVITY[70]

- The U-shaped relationship between vertical integration and competition: theory and evidence. See Aghion, Griffith and Howitt (2006);
- Plant-level analysis using the ARD: Another look at Gibrat's law. See Harris and Trainor (2005);
- Restructuring and productivity growth in UK manufacturing. See Disney, Haskel and Heden (2003) ;
- Import competition, productivity, and restructuring in UK manufacturing. See Criscuolo, Haskel and Martin (2004).

FOREIGN OWNERSHIP

- Foreign ownership and productivity in the United Kingdom - Some issues when using the ARD establishment level data. See Harris (2002);
- Blessing or Curse? Domestic Plants' Survival and Employment Prospects After Foreign Acquisition. See Girma and Görg (2005);
- Using the ARD establishment level data to look at foreign ownership and productivity in the United Kingdom. See Griffith (1999).

[70] For an overview of the role of the ARD in productivity research see: Barnes, *et al.* (2001), Understanding productivity: new insights from the ONS business data bank. Paper presented to the Royal Economic Society Conference, April 2001.

Table 11: Examples of research utilising NES/ASHE

- Worker-job matches, job mobility and real wage cyclicality. See Hart (2006) ;
- A panel data analysis of the effects of wages, standard hours and unionization on paid overtime work in Britain. See Kalwij and Gregory (2005);
- Nominal wage rigidity and the rate of inflation. See Nickell and Quintini (2003);
- The National Minimum Wage and hours of work: Implications for low paid women. See Connolly and Gregory (2002) .

Table 12: Examples of research utilising other BDL datasets[71]

INOVATION (COMMUNITY INNOVATION SURVEY)

- Productivity, Exporting and the Learning-by-Exporting Hypothesis: Direct Evidence from UK Firms. See Crespi, Criscuolo and Haskel (2006).

E-COMMERCE (E-COMMERCE SURVEY)

- The Productivity impact of E-Commerce in the UK, 2001: Evidence from microdata. See Rincon, Robinson and Vecchi (2001).

R & D (BUSINESS ENTERPRISE RESEARCH AND DEVELOPMENT, BERD)

- R&D and Productivity in the UK: evidence from firm-level data in the 1990s. See Rogers (2005).

INVESTMENT (BUSINESS SPENDING ON CAPITAL ITEMS BSCI)

- Information Technology and Productivity: It ain't what you do it's the way that you do I.T. See Sadun and Van Reenen (2005).

5.8 Summary

ONS collects large amounts of administrative microdata in the course of its business.

These data and the potential they offer have grown over recent years due to the efforts of the Office for National Statistics Business Data Linking (BDL) unit, and in part due to the evolution of a small academic user community who have developed hands on experience in using the data. Consequently, what are now tried and tested micro datasets now exist relating to business performance. Data are linked over time to create rich longitudinal data sources. Moreover, the datasets can be cross linked with other relevant business datasets held by BDL.

Thanks to ongoing developments at ONS, data are becoming increasingly user friendly and better documented. They offer a potentially excellent resource to the research community, which can access the data under restricted condition in the Virtual Microdata Lab (VML). Awareness and use of the data are growing year on year and this facility offers exciting possibilities for researchers in the future.

[71] For a bibliography of work undertaken based on the WERS dataset see: www.wers2004.info.

6 Demographic datasets

6.1 Introduction

The ONS Longitudinal Study (LS)[72] is a large scale data resource, available in England in Wales and currently being extended to Scotland and Northern Ireland[73]. The dataset draws upon the Census of Population, collected under statutory authority and governed by the Census Act (1920). The data, described in some detail in the next section, provide a dynamic account of demographic, health, geography, household and occupational information linked from census information, collected each decade from 1971 to 2001.

The ONS Longitudinal Study (LS) is a unique source of information with great potential for research. The nature of the database, which tracks individuals from decade to decade via the census, enables the study of small groups, change over successive time points. Whilst lacking the detail and regularity of follow-up of other longitudinal studies, such as the birth cohort studies or British Household Panel Survey, the LS has the major advantage of very large sample coverage, with approximately 1 percent of the population of England and Wales captured in the study. The other advantages of the LS are the fact that it is created with no further respondent burden (*i.e.* the data is collected for census purposes and required by law), its quality is high and degree of attrition over time limited.

The ONS has invested significant resources over a number of years to maintain the LS facility. By precise matching of data through the NHS population register, with high matching rates and rigorous data checking, the LS data is of especially high quality compared with that in many administrative datasets. The facility is similar to census longitudinal studies in Denmark, Finland, France, Israel, the Netherlands, Norway, Sweden and the USA.

The dataset is an exceptional resource in its own right. It is by far the largest longitudinal resources available in the UK, in terms of sample coverage[74]. Potentially it allows researchers to address many relevant social policy related questions. Topics of interest, detailed later in this section, include: longevity, health, geographic and social mobility, housing, education, and so on. Above and beyond this, however, the LS has potential as a vehicle for facilitating linking to other administrative datasets. The LS has well-developed mechanisms for tracing individuals and matching information about them over time. Given that it already exists and is tried and tested, the LS provides a potential platform for matching of administrative data from other sources, such as those relating to health, education and social security records.

[72] It should be noted that the ONS Longitudinal Study (LS) is separate and distinct form the DWP LS (covered in chapter 3) and the two should not be confused.

[73] The launch of the Northern Ireland Longitudinal Study has recently taken place, in December 2006.

[74] The disadvantage with the LS, compared to other longitudinal studies such as the British Household Panel Survey (BPHS), is the fact that there is a long period (i.e. a 10 year gap) between censuses.

6.2 ONS Longitudinal Study (LS)

6.2.1 Description[75],[76]

The ONS Longitudinal Study (LS) is a study which links census records and vital event information for a one percent sample of the population of England and Wales, initially defined in 1971 and linked through successive census data[77]. The study is designed to link together census and life event information from data routinely collected by the Office for National Statistics (ONS). The longitudinal aspect of the study allows a large sample of individuals to be tracked over four successive decades.

The LS database now includes information from census samples in 1971, 1981, 1991 and 2001. The LS contains linked information on a sample of individuals born on four dates across the calendar year, updating records from the original 1971 sample and replenishing the dataset with all newborns born on one of the relevant dates of birth each year. In the same way, members are also added to the study by linking information on immigrants using the same selection criteria. Members of the study are lost each year due to mortality, *i.e.* In a similar fashion, no more information is linked where study members have records indicating that they have left England and Wales (though linking recommences if they return to the UK).

As well as including information from the census, the LS also links other major events from other ONS sources. These are: births registered to women in the study (including birth-weight), cancer registrations, widowhoods/widowerhoods and deaths. As well as recording census records for the individual, the LS also records census information for all people living in the same household as the LS member at the time of the census. It is important to note, however, that the LS does not follow up household members in the same way, with household composition fluid over time.

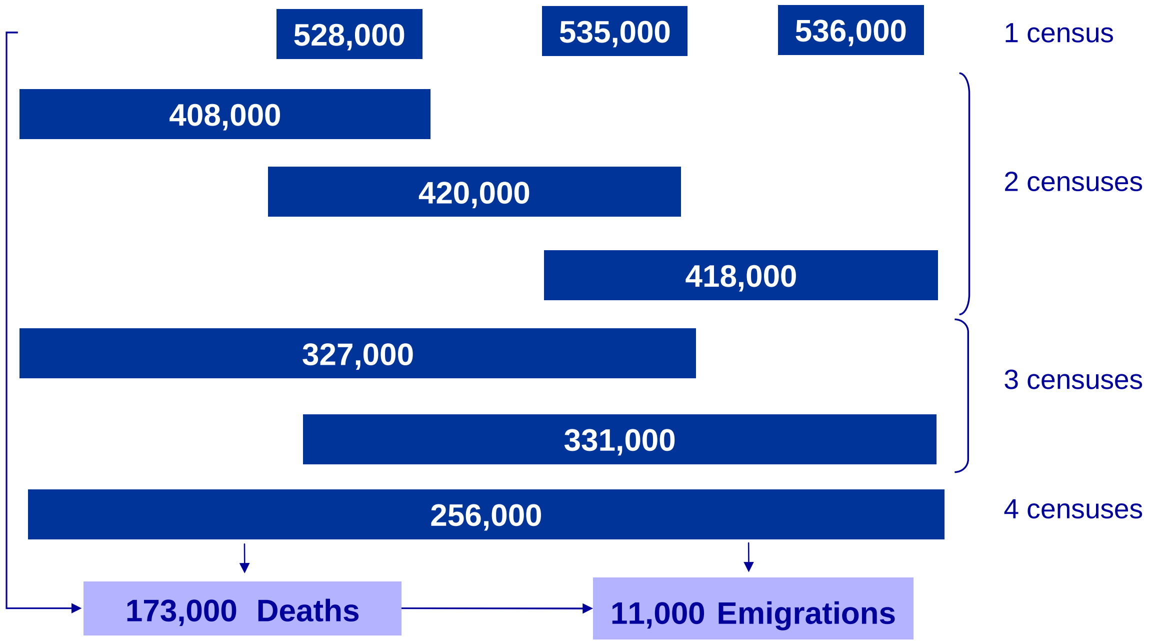
In terms of sample numbers, the LS covers over 500,000 people at any one time and linked in adjacent censuses. Figure 6 illustrates the sample coverage. It is noted that because the sample is replenished over time through entry (birth and immigration) and exit (death and embarkation), since the study began approximately 256,000 new LS members have been included in total.

[75] For a detailed guide to the LS, see Hattersley and Creeser (1995).

[76] Note that cross-sectional census microdata, not linked through time are available fro 1991 and 2001 via the Samples of Anonymised Records (SARs).

[77] Strictly the sample covers 4/365 (approximately 1.1 percent of the population) based on 4 selected birth dates.

Figure 6: ONS Longitudinal Study (LS) sample coverage



Source: Office for National Statistics

6.2.2 Variables

The data in the LS come from two sources. Firstly the complete census returns from 1971 onwards (linked at the individual level). The entire census record for the LS member and all members of that person's household are entered into the LS. This source provides information on key characteristics of individuals and their households. Secondly information from vital event data routinely collected by the ONS is linked onto the LS. This provides key demographic information. Annex 5.1 summarises the variables covered by each of these sources and which are contained in the LS. It is noted that the census data variables listed in the table are available for both LS members and all other members of the household in which they lived at the time of the census. It is also noted that census questions have varied slightly over time so that some variables are only available for selected years.

6.2.3 The LS data linking mechanism

The identification of individuals in the LS, the linking of individual census records over adjacent decades, and the linking of vital life events data is all undertaken via the National Health Service Central Register (NHSCR). The NHSCR is the central database (or spine) used to record information on all individuals who are know to be alive and assumed to be still living in the England or Wales[78].

In short, data linkage of individuals across censuses or onto the census takes place by tracing LS members' records on the National Health Service Central Register (NHSCR). This ensures that the same individuals are linked over time and is necessary because the LS itself does not carry identifying information such as name and address. The NHSCR

[78] The NHSCR records surname, forename, initial of second forename (if applicable), date of birth and gender.

is part of the Office for National Statistics but is operated separately from the ONS census collection and so acts, in essence, as a trusted third party. It was originally set up for the payment of GPs. It is the central database used to record information on all patients who are or who have been registered with a GP in England and Wales. In addition, the cross-referencing of census information with NHSCR ensures high quality matching for the inclusion of data for life events. It is noted that Governance of NHSCR and its relationship with ONS are under review as part of ONS Independence.

The NHSCR includes a 'flag' if an individual is present in the LS (*i.e.* with a prescribed date of birth). Tracing of individuals who appear in the census is done using details relating to surname, forename and any two parts of the date of birth and postcode of residence. This is done using a fairly complex auto match algorithm. Records which cannot be matched in this manner are dealt with separately where only 'exact' matches for individuals are tolerated. Having linked the census records, the vital life events data (collected continuously) are matched onto the LS separately and on a continuous basis. Birth, death and cancer registrations are traced at NHSCR while migration and enlistment records are supplied to the LS from NHSCR.

The use of the NHSCR linking facility results in a high rate of tracing of individuals completing census records. Tracing rates, amongst other things, are an important indicator of the quality of the underlying dataset. In the 2001 census 99.3 percent of individuals were successfully traced via NHSCR. This figure was an improvement on already high (96-98 percent) rates of tracing in previous censuses.

As a result of tracing individuals at NHSCR at the time of Census, LS data across Census dates (i.e. from 1971 forward to 1981, 1981-1991 and 1991 to 2001) can be linked. In this process there is again a degree of attrition and some individuals are termed to be 'lost to follow-up'. Reasons for this are not generally known. A major factor in this respect is due to individuals not living in the UK at the time of census, whether due to permanent or temporary emigration, without this being notified to NHSCR. Other reasons include non-enumeration in the latter of the linked censuses (or administrative error in recording) or information on the individual's death not being recorded in the NHSCR. The forward linkages rates for the LS are approximately 90 percent.

The very high level of tracing and relatively high forward linkage rates (compared to many longitudinal studies) should give reassurance to researchers concerned with issues of bias introduced into longitudinal data as a result of systematic attrition of individuals. This said, it is noted that in the small minority of cases where tracing of individuals in the LS was found not to be possible, further analysis revealed some systematic patterns. Factors associated with non-traceability from census and/or subsequent linkage failure to earlier census included[79]:

- being a young adult;
- being divorced;
- being born outside the UK;
- being a full-time student ;
- being in a minority ethnic group;
- being long-term unemployed or inactive;

[79] See Blackwell *et al.* (2005) for details.

- in the armed forces;
- living in London.

6.2.4 Accessing the LS

Academic users working in the UK higher or further education sectors have an established right of access to the LS data. This is facilitated by the Population Statistics Act 1938, providing that necessary measures are taken to avoid disclosure of identifiable personal information. Access to the data is provided and assisted by the Centre for LS Information and User Support (CELSIUS)[80].

The CELSIUS user support group provides the following facilities:

- A gateway to access the LS data;
- Documentation relating to the LS;
- Web access to the ONS LS data dictionary [81];
- A range of training courses, including online modules;
- Research dissemination through a mailing list;
- A list of current and completed LS research projects.

Access to the LS data is done, in the first instance by completing an application to the LS to have access to the data. A description of the application procedure as well as an application form is available on the LS website. Once received the application is processed and must be given approval from the LS Research Board (LSRB)[82]. The request for data must include specified variables and relevant geographical coverage and/or time period, which may be modified by ONS. All non-LS staff using the data directly at individual level are required to complete stringent undertakings before using the data and a confidentiality agreement must be signed by the researcher(s) involved.

Access policy for the LS operates under the same 'safe people-safe setting' guidelines as operated by the Business Data Linking (BDL) described in the previous section of the report. The 1998 LS Review states:

> "All non-LS staff using the data directly at individual level are required to complete stringent undertakings and a high level of trust is accorded to them. The Data Custodian and all staff working in the unit reserve the right to be confident of a user's skills and commitment to confidentiality when allowing access to the data." (ONS, 1998)

Access to data is allowed using the LS open access area (LSOAA) at the ONS in London and, shortly, Titchfield. ONS is also testing use of LS data via the VML in Newport. This restriction on access is imposed by the ONS in order to ensure confidentiality, since

[80] Although working closely with the ONS LS census team, for historical reasons CELSIUS are based at the London School of Hygiene and Tropical Medicine. Details of CELSIUS can be found at the following website: http://www.celsius.lshtm.ac.uk

[81] The CELSIUS website provide access to the LS data dictionary, which describes the 4,500 variables currently held in the LS. The dictionary contains a short description for each variable, with variable coding/labelling and their labels.

[82] As part of the approval process issues of potential confidentiality / risk of disclosure are carefully explored.

the dataset will contain potentially sensitive information relating to the individual (although information, such as name and address are not held in the LS).

Once permission is given for a project to proceed, the relevant subset of the data is created containing only relevant fields and cases. This dataset is made available to the individual as a flat microdata file (note that the LS is stored within a relational database structure) which can be manipulated using a suitable statistical package such as *stata* or SPSS. It is noted that this data, including outputs from it, remain the property of ONS at all times. The data may only be accessed by the researcher under supervision. Finally, confidentially is also protected through monitoring of output, and (as was the case with BDL datasets) only outputs which are checked by the LS team and cleared for disclosure risk will be made available to the researcher[83].

6.2.5 Research using the LS[84]

The LS provides a unique and invaluable resource for academic and policy related analysis. The data are relevant to research in many areas including: demographics, geography, health, socio-economics and many other areas of social policy. The study is particularly powerful, in terms of its design, in that it facilitates **follow-up** type studies so that 'events' (life event and health) can be related to initial circumstances. Thus one can study associations, for example, between socio-economic or employment status and future health, mortality or fertility. Alternatively, the linked data over time may be used to study patterns of change, for example relating to migration, housing and social mobility[85].

The LS has been used widely in a variety of research area, has most notably including demographics or health based research to which the study lends itself most readily. A list of the main research topic areas is given below. This list is neither exhaustive nor are the topic areas mutually exclusive.

(i) Demographics (including household formation and change);
(ii) Health (or health inequality);
(iii) Local geography and migration;
(iv) Socio-economic change (including social mobility);
(v) Housing and households;
(vi) Equal opportunities (ethnicity and gender studies).

The relevance of the LS to each of these areas and examples of relevant research are described in more detail below.

(i) Demographics

The LS was established, in the first instance, as a means of studying patterns of mortality and fertility in the UK. To this end, the LS provides a unique data resource for the UK. The database provides a means of studying various demographic phenomena in relation

[83] With respect to disclosure, policy states that "only personal information on an individual or individuals should not be disclosed or presented in a form which may lead to its disclosure."

[84] This section draws, in part, on chapter 3 of ONS (1998).

[85] The LS can also facilitate international comparisons of cross sectional data. Countries such as France, Denmark, Finland and the USA have similar programmes to the LS.

to other factors such as social class. Topics of interest include mortality, including cause of death, longevity and ageing on the one hand, and on the other fertility and births. In addition it is possible to study events around the household and its composition, family formation and/or divorce. Recent examples of research in this area are listed in Table 13.

Table 13: Demographic research using the LS

- Person, place or time? The effect of individual circumstances, area and changes over time on mortality in men, 1995-2001.See White *et al.* (2005);
- Fertility, timing of births and socio-economic status in France and Britain: social policies and occupational polarisation. See Ekert-Jaffe *et al.* (2002);
- Children's changing families and family resources. See Clarke *et al.* (2001);
- Comparing the childrearing lifetimes of Britain's 'divorce-revolution' men and women. See Rendall *et al.* (2001);
- Trends in births outside marriage. See Babb and Bethune (1995).

(ii) Health

In addition to using the LS to measure mortality (and/or social differentials in mortality) the LS contains a range of other health indicators. The LS contains information on cancer registration and birth-weight as well as on infant mortality. In addition there are questions in the census on self-rated health and long-standing illness. The strength of the LS for facilitating health related research is the ability to link health to socio-demographic information about the individual. This presently cannot be done using any other source with the same degree of statistical power. Thus the LS has spawned much research into differential mortality and health inequality. Other interesting topic areas relate to geographical differences in health, occupational health, infant mortality and cancer incidence and survival (the latter topics were covered in more detail in the earlier chapter on medical research using the LS). Recent examples of research in this area using the LS are listed in Table 14.

Table 14: Health research using the LS

- Inequalities in lung cancer mortality by the educational level in 10 European populations. Mackenbach *et al.* (2004);
- Accumulated labour market disadvantage and limiting long term illness: data from the 1971-1991. See Bartley and Plewis (2002);
- Do attitude and area influence health? See Mitchell (2000);
- Occupational mortality of women aged 15-59 years at death in England and Wales. See Moser and Goldblatt (1991).

(iii) Local geography and migration

The LS is ideally suited to study changing patterns of settlement and local geography as well as factors affected long term migration. The LS has been used variously in this respect. Work in this area includes the study of internal migration. For example, the link between inter-regional and social mobility to reveal the role of the South East region as an ‘escalator’ has been studied. Similarly the LS has been used to the tendency towards ‘counter-urbanization’, documented by the LS elsewhere in the country also has implications for housing demand and planning. Several studies use the LS to track spatial distribution of urban population (*e.g.* in London) or patterns of out-migration (e.g. in Cornwall or Liverpool). Relevant research in this area is listed in Table 15.

Table 15: Local geography and migration research using the LS

- The leaving of Liverpool: an examination into the migratory characteristics of Liverpool. See Davies, Williamson and Holdsworth (2006);
- Migration and social change in Cornwall 1971-91. See Williams (2000);
- Counter urbanisation and social class. In: Boyle, P J, Halfacre, K H, editors. Migration in Rural Areas: Theories and Issues. See Fielding (1998);
- Migration and social mobility - South East England as an escalator region. See Fielding (1991).

(iv) Socio-economic change

In addition to the research on geographical mobility, the LS is able to facilitate research on changing socio-economic patterns and social mobility. Topics of interest in this respect include research on labour market change, occupational mobility, as well as social mobility measured by transitions across generations (children compared to their parents). Relevant research in this area is listed in Table 16.

Table 16: Socio-economic research using the LS

- The long shadow of childhood: associations between parental social class and own social class, educational attainment and timing of first birth. See Buxton (2005);
- Life after mining: hidden unemployment and changing patterns of economic activity amongst miners in England and Wales, 1981- 1991. See Fieldhouse and Hollywood (1999);
- Car availability change in England and Wales 1971-1981. See Axhausen (1995);
- A century of change in occupational segregation 1891-1991. See Hakim (1995).

(v) Housing and households

The LS has been used in research on housing. Topics of interest in this area include: trends in home ownership and private renting, housing deprivation (measured by overcrowding and lack of amenities) and living arrangements and household composition. Again, the strength of the LS is that it allows these topics to be analysed in conjunction with information on socio- economic and labour market factors. Relevant research, including a general guide to LS based research in this area (by Brassett and Grundy, 2003), are listed in Table 17.

Table 17: Housing and household research using the LS

- Researching Households and Families Using the ONS Longitudinal Study. See Brassett-Grundy (2003);
- Co-residence of mid-life children with their elderly parents in England and Wales: changes between 1981 and 1991. See Grundy (2000);
- Divorce, Remarriage and Housing: the Effects of Divorce, Remarriage, Separation and the Formation of the New Couple Households on the number of Separate Households and Housing Demand and Conditions. See Holmans (2000);
- Labour and housing market change in London: a longitudinal analysis. See Hamnett (1987).

(vi) Equal opportunities - ethnicity / gender

Analysis of the LS underlies a considerable body of research on equal opportunities with respect to both ethnicity and gender. This research ranges from studying patterns of segregation at work (for example with respect to occupational segregation) to more general issues of social status and housing. Recent examples of research in this area using the LS are listed in Table 18.

Table 18: Equal opportunities research using the LS

- Stability and change in ethnic groups in England and Wales. See Platt *et al.* (2005);
- Occupational sex segregation and part-time work in modern Britain. Gender, Work and Organisation. See Blackwell et al (2001);
- Women and Science Teaching: the Demographic Squeeze. See Blackwell (2001);
- Immigrant workers and the class structure of England and Wales: a longitudinal analysis of the social mobility of Britain's black and Asian populations. See Fielding (1998);
- A century of change in occupational segregation 1891-1991. See Hakim (1994);
- Why don't minority ethnic women in Britain work part-time? See Dale and Holdsworth (1994);
- Caribbean tenants in council housing: race, class and gender. New Community. See Peach and Byron (1993).

6.3 Scottish Longitudinal Study (SLS)

The Scottish Longitudinal Study (SLS) is a close replica of the England and Wales Longitudinal Study (LS)[86]. The study is based on a 5.3 percent sample of Scottish census data, *i.e.* based on 20 birth dates in the year and hence achieving 5 times more coverage than the corresponding survey in England and Wales in percentage terms. In total, the study will cover approximately 274,000 individuals (SLS members) identified from the 1991 census and linked forward to 2001. It also includes information on the other household members in the households containing the SLS member at each census, although these people are not linked through time. The SLS links census data for 1991 and 2001 (earlier linked data is not available) in the same manner as described previously

[86] In parallel to the ONS Longitudinal Study for England and Wales and Scotland a similar resource for Northern Ireland was launched in December 2006, though linkage back to 1991 is not currently available.

and tracing is done through the NHSCR, which then allows linkages to be made to other data sources, such as vital events. At the time of writing the study is still undergoing its construction phase. However, it will be launched for access to researchers in March 2007.

The Scottish LS is smaller in absolute size than its counterpart study in England and Wales and does not include such a long timespan. However, it does have the advantage that it is particularly valuable for health based research. As is the case south of the border, the study will link census data to vital events data (births, deaths, and marriages), information about migration (from NHSCR) and cancer registrations. In addition, however, the Scottish LS will include information for sample members on hospital episodes (hospital admissions and discharges). This is available from the NHS hospital episode statistics (HES) in Scotland[87]. It is noted that the health datasets will not be held as part of the SLS database. The linking will be done remotely via an NHSCR flag and will only be provided on an 'as and when needed' basis. The data will provided only for approved research studies, validated by the SLS team and a health research ethics committee. The inclusion of health related information will greatly enhance the value of the LS as a research resource and represents a great step forward in terms of data linking. The same facility is not yet available in England and Wales.

Access to the SLS will be available (provisionally) from late 2006 onwards. Access to the database will operate in a similar way to that used for the England and Wales LS. The data will be made available to academics upon the successful processing of an application to be submitted to the SLS Research Board. The application form includes a clear outline of the study, its purpose and detailed data requirement. The researcher is required to sign a confidentiality declaration included in the application form. Access to the data will be possible either in a safe setting (in Edinburgh), or remotely. However, microdata will not be released to remote users – rather they will submit syntax code to be run on the file that has been prepared for them and results from these analyses will be returned to users. Support will be provided by SLS staff and supporting documentation, currently being written, will also be available via the website[88],[89].

6.4 Matching administrative data using the Longitudinal Study

A great strength of the longitudinal studies for the future is the fact that they provide latent potential as a platform and vehicle for the matching on of other administrative data collected at individual level. Conceptually, one can imagine a scenario whereby routinely collected administrative data can be matched onto the LS data, for example by identification and flagging of individuals via the NHSCR. This would enhance the LS without the extra burden of data collection.

'Proof of concept' has already been established. Instances of this include: the now routine linking of ONS vital events data and cancer registration data to the LS, as well as the linking of health records onto the LS for Scotland. In addition to this, a recent initiative to link individual unemployment records data onto the LS is now coming to fruition. This is described in the next subsection of the report. Other ideas for linking

[87] See chapter 4 for a description of HES.

[88] Details of training in longitudinal data handling, analysis and modelling, along with on-line training modules and step-by-step guides to the SLS data will be posted on the website in due course.

[89] Details regarding the SLS, including how to access the data, can be found at the following link: http://www.lscs.ac.uk/sls/data.htm

have been explored in some detail previously in the 1998 LS review in England and Wales. Section 6.4.2.of the report provides a summary of this potential by citing possible examples of future linkage.

The potential stated in this section is for illustration only at this stage. It is accepted that there are potentially major barriers to making these linkages operational. These relate to issues of cross- departmental sharing of data within government as well as legal implications. There are also the broader issues of whether or not the LS is the best vehicle to facilitate such linkage as well as resource implications. These macro issues are discussed in some detail in section 7 of the report.

6.4.1 Linking the LS with unemployment records

The ONS has recently been successful in linking individual unemployment claimant records to the ONS Longitudinal Study[90]. This was done using Joint Unemployment Vacancies On-Line System (JUVOS), originally collected by the Department for Work and Pensions (DWP). JUVOS is a database with 100 per cent coverage of claimants, and contains personal and occupational information as well as claimant dates. The inclusion JUVOS data provides a continuous record of LS members' unemployment events. This is available for the period 1994 to 2005. The inclusion of these data facilities research relating the experience of unemployment spells to social, economic and geographic factors at the individual level.

6.4.2 Other potential linkages to the LS

The linkage of the LS to JUVOS as described above is one of many potential linkages suggested in the *1998 LS Review* document which has come to fruition. However, the review also highlighted many other such linkages of the LS to administrative data sources are feasible in theory. These suggestions are described briefly below. These include an LS link to the various data sources describes elsewhere in the report[91]:

(i) Department for Works and Pensions Longitudinal Study (DWPLS)

A link to the DWPLS would allow the incorporation of all spells of economic activity (including employment and periods of sickness/disability) into the LS. This would extend and enhance the JUVOS link. It would provide information about job switching and job mobility. The DWPLS could also be used to incorporate information on income, NI and pensions contributions. This would provide information on the ‘quality’ of employment outcomes in terms of remuneration. Linkage between sources is possible using existing data matching facilities at NHSCR.

(ii) Department for Education and Skills (DfES) database

A link to the DfES database would allow individual education records (in school and possibly beyond) to be incorporated into the LS. Information on schooling, educational progression and achievement, *etc.* could then be studied in conjunction with information

[90] At the time of writing the data linking had been established but was undergoing beta testing to provide robustness checks on the data linkage. Initial estimates suggest that the matching of claimants to LS members was successful in over 97 per cent of cases.

[91] This list is not exhaustive. One can also easily envisage links to criminal records, department for transport/DVLA records, migration information and so on.

relating to socio-economic characteristics of the household. Linking LS and DfES data would similarly present practical problems in terms of finding a common identifier. Linkage between sources is possible using existing data matching facilities at NHSCR.

(iii) Hospital Episode Statistics (HES)

As has been shown in Scotland, basic information on hospital admissions and discharges could be linked onto the LS. This would supplement existing census data on health (self reported health and long standing illness) as well as cancer registration data. This would enhance the LS as a resource for the study of health phenomena and health inequality in relation to socio-economic factors. The linking LS and HES can be facilitated without extensive practical problems via an NHSCR identifier.

It is noted that approval has also been granted (in July 2006) to link and make available for substantive research Health Authority postings data from 1991, drawn from the NHSCR Central Health Registry Inquiry Service (CHRIS). These new data will provide information on internal migration for LS members. Health Authority de-registrations, also to be linked, may serve as a proxy for unobserved embarkation from England and Wales, thus filling a known shortfall in international migration data. Testing of these data is currently underway prior to their incorporation in the LS database, planned for 2007.

(iv) Inter-Departmental Business Register (IDBR)

A link to the IDBR would open up the potential to create a database which relates individual characteristics to workplace characteristics. This opens upon potential avenues of research in terms of analysing employment patters and structures within firms. It is noted, however, that such a link would be an arduous labour-intensive task and would probably need to compare 'work address' on the census form with the firms' addresses in the IDBR.

(v) Survey data

As well as linking the LS to mainstream administrative data, it is possible in principle to link LS members' information in sample surveys back to the LS. Examples of relevant surveys in this respect might include: the New Earnings Survey (NES), British Household Panel Survey (BHPS) and Labour Force Survey (LFS). It is noted, however, that even for large surveys the number of people linkable (based on one of the four birth dates) might be quite small. One would also need to be convinced of the value of he various links.

6.5 Summary

The ONS Longitudinal Study (LS) in England and Wales and their embryonic counterpart studies in Scotland and Northern Ireland provide a unique dataset in the UK for monitoring new births and deaths, combined with an unrivalled database for studying social and economic change in its own right or mortality, health, household change, migration, education and the labour market in relation to underlying socio-economic factors. The sample size is large, covering more than a half a million people at one time, and dominates that of other longitudinal studies such as the birth cohort studies or British Household Panel Survey (BHPS) in terms of width of coverage (although it is noted that these studies have advantages in terms of frequency of observation and range of questioning).

An additional potential strength of these longitudinal studies is that they provide potential as a platform and vehicle for the matching on of other administrative data, as explored in the previous sub-section of the report. The latent potential for linking on other potentially very large datasets offers potentially unlimited scope for innovation longitudinal based research into work, education and health. However, practical and legal considerations have stalled most potential linking exercises for the time being. The general issues of data sharing and matching are considered in the next section of the report.

7 Issues in data access and sharing

7.1 Introduction

During the past few years there has been a growing interest in the potential for administrative data sources to be better utilized and exploited by interested researchers. Administrative data sources, such as those listed in previous chapters, contain a wealth of information relevant to the policy research interests of central and local governments and to the wider research interests of the scientific community. The Cabinet Office's Performance and Innovation Unit (2002) made recommendations about how more widespread and coordinated use of personal data could be promoted by government. This has been followed up recently by the cross government 'MISC 31' committee on data sharing[92]. Similarly, the potential for administrative data sources was flagged up in the Allsopp Review (2003) which suggested:

> "Administrative data appear to offer opportunities to increase the quality and analytical power of key national statistics, as well as reducing costs. More generally, within the important constraints of adequate protection for sensitive information and limiting use to solely statistical purposes we believe there is considerable scope for the government to make better use of the information it holds. The ONS and the government should explore the extent to which tax and other administrative sources could replace business survey data. They should propose the necessary action to overcome legal and other barriers where information is held within government that is of sufficient quality to improve statistical provision, or where quality can be increased to meet statistical needs, while maintaining adequate safeguards of confidentiality." Allsopp (2003)

The vision for the future is to make use of administrative data for two purposes. The first is to increase access to administrative data so that data collected for administrative purposes are better utilised, both for research and to inform service delivery. This is particularly relevant given technical developments over recent years which now make handling of complex datasets much more viable than say even five years ago. Steps have been made recently towards better provision and access to administrative data, in particular education and business data provide two examples of good practice. This chapter discusses some general issues regarding access to personal datasets and the use of such data for research purposes.

The second possibility regards the scope for linking large scale administrative datasets to address key policy issues[93]. Examples include the potential for linking of administrative datasets held separately and under independent stewardship by different government departments or bodies, along the lines of that which has already been achieved by integrating DWP and HMRC data, as outlined in chapter 3. Beyond this there is the possibility of linking existing administrative datasets onto the longitudinal census data, described in chapter 6 or, more radically, using administrative data as a means of removing the need for a census. Issues surrounding the feasibility of such an undertaking and potential barriers including issues of legality are also discussed in this chapter.

[92] For the recently released 'Data Sharing Vision statement' see DCA (2006).

[93] An example of a recent policy initiative which champions the need for data sharing is the Social Exclusion Action Plan. See http://www.strategy.gov.uk/work_areas/social_exclusion/index.asp

7.2 Benefits of data sharing[94]

Data sharing is a generic term which refers to the situation whereby government departments (or similar bodies) allow access to microdata under their stewardship. The benefits from encouraging improved data sharing in the UK are broadly two-fold. The first relates to the potential that administrative data offers in terms of providing a more comprehensive evidence base to assist in policy making. The second aspect relates to the potential for using administrative data as a means of achieving cost savings, in terms of utilizing data which would otherwise remain dormant as well as avoiding duplication of data collection through surveys. These two aspects are discussed in more detail below.

7.2.1 Benefits of administrative data

This section lists some of the perceived benefits of administrative data above and beyond survey (or census) data, which are collected with considerable effort and often with large cost[95]. These include:

(i) **100 percent coverage of target population** – unlike survey data which are by their nature based on a small sample (usually for reasons relating to practical administration and costs) administrative systems will often contain data for all (or close to all[96]) of the target population. Examples of wide coverage include education data (chapter 2), benefits data (chapter 3) and hospital episode data (chapter 4).

(ii) **Large sample sizes** – on a related issue, given 100 percent coverage of the target population (as opposed to a sample survey) the amount of information which can be collected is potentially vast compared to survey data, with resulting datasets often being very large indeed. As an illustration, the DWPLS will contain approximately 100 times more observations of economic activity/employment/earnings each quarter than the corresponding UK Labour Force Survey (LFS) which is based (approximately) on a one percent sample.

(iii) **Attrition is minimized** – the classic problem with longitudinal survey data is one of attrition. As individuals are tracked through time many of the original cohort members are lost in follow-up surveys. Moreover, the loss of individuals tends to be systematic and depend on personal characteristics[97]. The nature of administrative data collection along with census coverage ensures that attrition will be much less of a problem.

(iv) **Accuracy** – due to the nature of the data collection administrative data are likely to be more accurate than survey data, and in particular less subject to recall error or mis-reporting on the part of the individual. An example of this can be found in the

[94] For a general discussion on data sharing see: GRO (2005).

[95] This section lists the benefits of administrative data compared to survey data. It is recognized however, that converse arguments exist which are not listed here. Not least of all survey data can achieve much more detailed and finely tuned data collection that currently possible using administrative data.

[96] It is noted that 100 percent coverage may be not always be achieved even if this is the underlying aim of data collection. Administrative data may be subject to missing observations due to problems of tracking individuals, changing personal details, non response, etc. These will depend on the nature of data collected and means of collection.

[97] The Youth Cohort Study (YCS) provides an excellent example of large scale systematic attrition in follow up surveys.

ASHE earnings data collected through the PAYE system (see chapter 5) which is collected from tax records. In contrast, earnings data in the Labour Force Survey (LFS) is based on individual recall.

(v) **Timely data** – administrative data sources are regularly (and sometimes continuously; for example DWPLS) updated through automated data collection. This means that repeated surveys need not be initiated. It also means that time lapses are minimized and data can relate to very recent time periods.

(vi) **Non intrusive** – because administrative data are collected on a regular basis using established means of data collection (for example via tax, national insurance or school records), the collection procedure is not seen as intrusive or burdensome by the target population. The opposite may be true for survey data.

(vii) **Historical data** – may well be available from administrative records so that policy relevant analysis can be undertaken using existing data covering previous time periods. This may remove the need for either data collection over future periods or unnecessary time lapses in collecting and analyzing data.

(viii) **Data are linkable** – much administrative data are potentially linkable on to other data sources. For example it may be possible, as discussed, to link administrative and census data or link independent administrative datasets (for example on employment and health) using a common identifier. This kind of linkage potentially allows social and health researchers to answer many interesting and complex questions. Issues regarding data linking are addressed later in this section.

7.2.2 Cost savings

In addition to the considerations listed above, there is also the additional consideration of potential cost savings arising from the better utilization and integration of administrative data into national data collection and provision. Survey data often replicates (to some extent) what has already been collected from administrative sources[98]. A pertinent point regarding administrative data is that they are already being collected. Therefore utilising administrative data to replace survey data or avoiding duplication with the system could result in potentially large cost savings, with little additional cost beyond the extraction, cleaning and standardisation of data.

Two possible scenarios for future policy with respect to the development of administrative data sources are as follows. The less ambitious vision would involve a plan for looking at the degree of overlap between administrative and survey data resources. In addition a plan could be formulated for better utilizing and integrating public-sector databases which exist separately and under the stewardship of different departments. Integrating and possibly linking government databases in a more systematic way would help to rationalize data collection and hence deliver cost-savings. It would also help to provide a common platform for more effective collection and management of microdata relating to individuals. Beyond this a more ambitious agenda exists in relation to how administrative datasets may supplement or even replace the collection of census data. These issues are the subject to internal debate within ONS that is ongoing at the time of writing.

[98] This could be said of survey data relating to labour market, health or education.

It is clear that at this point plans to utilize administrative data more systematically are still evolving and have been the subject of recent consultancy documents (see: ONS 2003a). However, the precise degree of overlap between administrative and survey data, and therefore the extent of duplicated resources and potential cost savings, has not been well quantified at this point. In addition, cost saving aspects of using administrative data need to be assessed in a broader cost benefit analysis exercise which considers interests of departments, stakeholders and the wider public.

7.3 Legal aspects of data sharing[99]

7.3.1 Accessing departmental data

Data sharing by government departments (or similar bodies) is restricted by their legal responsibility towards the data under their stewardship. However, at this point in time it is well recognized that the legal framework around allowing access to microdata (for non-contracted researchers[100]) is not well defined[101]. The current interpretation of the law relies heavily on common law obligations and interpretations of a number of statutes which in some cases vary from department to department. It is fair to say that the many areas of legislation are imprecise, in particular with respect to rules on disclosure of microdata, and the legislation has not been tested in court.

The result of the complex legislative situation is that government departments are prone to interpret the legislation differently, depending upon the nature of the information they hold - with some departments taking a more permissive view on data sharing (such as DfES, with respect to their release on NPD data) whereas some departments take a more conservative stance (such as DWP with respect to their longitudinal study). Overall, however, the uncertainly currently prevailing can give rise to conservative culture of research governance whereby departments might be prone to minimize risk of potential of legal action rather than embrace openness for what might be perceived as limited gain[102].

The current legal framework is governed, by and large, by three main legislative threads. These are:

[99] For an extend discussion of the legal aspects of data sharing from the perspective of various departments and bodies see: ONS (2005), AMS (2006), DOH (2005) and Scottish Executive (2004) .

[100] The legal framework governing contracted researchers (i.e. those undertaking work on behalf of the department) is somewhat different from that described in this section. Data contracts between government departments and researchers specify the level of access may be granted and the use to which data can be put.

[101] Under the Freedom of Information Act (FIA) 2000, any individual or organisation now has the right to make a request for any information held by a public authority. In the main course of events data, data not relating to the individual making the request is released at a suitably aggregated level. For more information on FIA see: http://www.opsi.gov.uk/acts/acts2000/20000036.htm

[102] Walley (2006) and Warlow C (2005) suggest that the interpretation of UK confidentiality laws may be seen currently as overzealous from the point of view of medical researchers.

(i) Data Protection Act (DPA) 1998

Section 33 of the DPA states that administrative data may be used, where appropriate, for "statistical or historical research". However, the terms and conditions of this use are not well defined. More generally, any handling of data relating to individuals is subject to the auspices of the Data Protection Act 1998. To this end sharing of administrative micro data must, in the first instance, protect individual confidentiality[103].

(ii) Human Rights Act (HRA) 1998

The HRA 1998 similarly protects individuals' right of confidentiality with respect to release of data. Departments must in particular be aware of the individual's "right to respect for his private and family life, his home and correspondence," under Article 8 of the act. The state may only override these considerations in the following exceptional circumstances: national security; public safety; protection of the economy; prevention of crime and disorder; protection of health or morals; or the protection of the rights and freedoms of others.

(iii) Common law of confidentiality

In addition to these two pieces of legislation departments are also bound under confidentiality by common law. Again, the nature of common law as it applies to data disclosure is not well defined since relevant test cases have not yet set strong legal precedence. However, it is noted that a breach of confidentiality may provide grounds for a civil action for damages.

In addition to these general legal guidelines, departments may also be restricted by specific acts which govern their department's actions and responsibilities. Where relevant these have been outlined in the relevant chapters, but include: the Census Act 1920, Education Act 1996, Health and Social Care Act 2001, the Statistics of Trade Act 1947 and the Statistics Act, 1938.

The application of the legal framework described above means that departments must be careful to ensure that data sharing is lawful and that confidentiality is maintained. In practical terms this means anonymising (of pseudonymising) data before release. Anonymisation requires the removal of name, address, full postcode and any other detail or combination of details that might support identification[104]. In addition the department must also show due regard to disclosure risk whereby individuals (and their confidential data) may be recognized by inference even after anonymisation. Issues of disclosure risk are discussed in the next section.

7.3.2 Cross departmental sharing

The situation with respect to sharing data between government departments is more difficult. At the time of writing there is no legal gateway of any sort which facilitates the sharing of data between government departments or with external organizations. As a rule it is therefore currently the case that government departments do not share data.

[103] It is noted that if any person discloses individual estimates or returns he/she will be liable to imprisonment and/or a fine

[104] Pseudonymised differs from anonymised data in that the original provider of the information may retain a means of identifying individuals.

Moreover, the lack of a legislative framework has brought to a halt previous initiatives to link administrative data[105]. In instances where data sharing has taken place this has been done under the auspices of specific acts of parliament. The best example of this in the context of this report is the sharing of data between DWP and HMRC which was facilitated by Section 122 of the Social Security Fraud Act 2000 and for the specific purpose of tracking fraudulent benefits claims. However, in this case permission was specifically sought by DWP by the addition of clauses to the bill in its transition through parliament, and under the act data sharing is one directional (such as is fit for purpose): from HMRC to DWP. In a similar fashion, the DfES has recently (January 2004) tabled a Statutory Instrument (SI) requesting departmental access to the Child Benefit data.

Clearly such *ad hoc* fit-for-task means of initiating data sharing arrangements are in general unsatisfactory. What is needed in order to create a legal gateway for data sharing is an explicit piece of legislation or general statistics act giving explicit permission for bilateral sharing.

7.3.3 Barriers to data sharing

The ill defined legal framework and the issues surrounding the perceived risk of data disclosure create effective barriers to data sharing. In particular the ONS publication 'Data Sharing for Statistical Purposes' (see ONS, 2005) lists the following barriers to data sharing:

(i) Administrative powers – it is acknowledged that there is a marked reluctance in departments to use implied administrative powers to enable data sharing due to the nature of the legal environment described above. One way forward in this respect, according to the report, is for ministers "[to] instruct their departments to audit their powers to share data with ONS for statistical purposes, and make the outcome of such audits available across the GSS."

(ii) Duty of consent – under common law the department holding data must direct or indirectly gain consent from individuals before sharing it. In practical terms this means that departments must have a statutory gateway that implies consent or, according to the report, "rely upon them concluding that the data access is in the substantial public interest" (such as issues of crime and fraud).

(iii) Data Protection Act – with respect to data sharing the DPA is ill defined and in need of clarification. To this end the ONS "would welcome any amendments to the Act that assists the sharing of data for statistics."

(iv) Cultural and technological – it is fair to say that over recent years the technology to link and match data has evolved much more rapidly than a supporting legal structure. Whilst integrating data systems is now readily achievable, whereas it was not ten years ago, powers that enable this are not sufficiently in place. The report suggests that "new primary legislation is required if the potential of these technological advancements are to be realised full", but also recognised that cultural practices and systems need to evolve within departments to embrace the possibility of change[106].

[105] Recent initiatives to link DWP and ONS data have failed in this respect.

[106] For a fuller picture of how technology enables data sharing and its implications see: Cabinet Office (2005a) and Cabinet Office (2005b), and CST (2005).

7.3.4 The Statistics and Registration Service Bill

The Statistics and Registration Service Bill, was introduced to parliament on 21 November 2006. The bill is intended primarily to create a new independent 'Statistics Board' (outside of Ministerial control) to replace the Office for National Statistics[107]. However, it contains clauses which, if enacted, may help clarify some of the legal issues regarding access to administrative data.

The bill contains clauses that should facilitate more sharing of data between government departments, and potentially provide an opportunity for extending access to administrative data for statistical purposes. Specifically, Clause 21 states that "the Board may promote and assist statistical research, in particular by providing access where it may lawfully do so to data held by it." The bill suggests practical arrangements for providing access to offical data for "approved researchers'. In addition, Clause 44 enables the Treasury to make regulations which authorise a public body to disclose information to the Statistics Board (when prevented from doing so by existing legislation) in order that the board can conduct its functions.

7.4 Disclosure risk and access to data

A primary consideration in data sharing is the control of disclosure risk. Disclosure is the releasing of confidential information relating to individuals (or businesses) which breaches the legal obligations of confidentiality described above. It is noted that managing risk of disclosure is not as simple as anonymising datasets before release, although this is an obvious first requirement. In general, anonymisation is not in itself an adequate strategy for protecting against disclosure. A more general risk exists whereby individuals may be identified by inference from a dataset by specifying a filter on the data, across several dimensions or variables, so that the number of individuals in the resulting cell is equal to one.

The aim of disclosure control is to ensure that no unauthorised individual, technically competent with public data and private information could, according to ONS: either (i) identify any information not already public knowledge and supplied in confidence to ONS (such as survey returns) with a reasonable degree of confidence, or 2) associate that information with the supplier of the information.

Consider an example of this in the context of a labour market dataset. Let us suppose that an anonymised dataset has variables for gender, age, local geography and occupation. Imagine restricting the dataset, via a filter, to the following:

"gender = Male; age= 47; geography = Shetland Isles; occupation = plumber".

The resulting number of individuals of that description may be very small indeed and perhaps only one. In this case the anonymised dataset may be reduced to a single individual who may be readily identified. Moreover, confidential information on earnings and so on may be readily obtained. Similar instances of disclosure may also apply, for example, to education, demographic or health data where issues of confidentiality may be very important indeed.

[107] For more information on the bill see: http://www.hm-treasury.gov.uk/consultations_and_legislation/statistics_bill/statistics_bill_index.cfm

Disclosure risk through reducibility of the dataset of this sort needs to be managed and organisations such as the ONS have paid a great deal of attention to data matching and disclosure risk (see ONS, 2003d) and the issue has generated some academic debate. What is clear, however, is that the task of quantifying risk precisely is difficult (in terms of deriving a workable metric[108]) and risk depends to a large extent on the nature of the dataset, the variables contained, the number of identifiable categories per variable and the relative sparseness of observations on selected dimensions of the data. In some case these issues require discussion and judgment.

The ONS policy on disclosure risk is as follows:

- To identify possible situations where disclosure risk could occur;
- To identify key variables which may be used to indirectly identify individuals;
- To implement probabilistic methods for estimating the disclosure risk measures using DIS/SUDA methods[109];
- To set threshold rules on level of aggregation in output tables derived from microdata.

With respect to the final point on rules for reporting based on microdata, the ONS applies the following rules:

(i) Threshold Rule – no cells based upon less than 10 observations;
(ii) Dominance Rule – the sum of all but the largest two units exceeds at least 10% of the value of the largest unit;
(iii) Secondary Disclosure – ensures that summary data cannot be made disclosive by means of differencing within or across tables to reveal new information.

7.5 Issues in data linking

As has been highlighted throughout this report, there are notable benefits to be gained from linking administrative datasets. The main obvious benefit in linking administrative data lies in being able to bring together existing resources to create new datasets which in turn would in turn allow researchers to address complex research questions relating to health, socio-economic status, education and so on. In addition to this, however, linking of administrative data into other survey datasets or into the UK census could be potentially of use in order to fill gaps in data collected through surveys which suffer from problems of non-response.

Data linking involves taking information about the same individual (or business unit, depending on the data source) from two or more datasets and merging these datasets to produce a new dataset containing matched individual records. As already documented, little work has been done at this point in linking large administrative datasets relating to individuals, other than the example of linking census data (chapter) and linking DWP and HMRC datasets (chapter 3). However, considerably more work has been done on matching business datasets where the legal restrictions are less prohibitive.

[108] The ONS have recently launched a research project to examine metrics on disclosure risk.

[109] For details of these methods with particular reference to how they were implement by ONS for 2001 Census SAR release, see www.ccsr.ac.uk/sars/events/2004-09-30/gross2.pdf

Matching data from two independent sources based on common identifier(s) is not an exact science. Matching relies on the merging of datasets using one or more shared fields such. For individual-based data, relevant fields are likely to be: Name, Sex, Address/ Postcode, Date of Birth, and perhaps a common administrative identifier such as National Insurance Number (NINO). For firm based datasets relevant fields are likely to be: Name and Address of firm, and a common administrative identifier based on the IDBR reference number. The problem is that there are likely to be issues surrounding the precision with which the data can be matched and the likelihood of errors and mismatches arising in the process of data linking. Some of the main considerations when matching data are outline below.

(i) Method of matching[110] – the method by which records are matched will determined the likely degree of error and/or attrition in the resulting matched dataset. Data may matched using either (a) *Deterministic* (or exact) matching – this method relies on all-or-none matching whereby if the all fields do not match exactly in the two datasets then the match is rejected; or (b) *Probabilistic* matching whereby the rules (or tolerance levels) are applied to maximize the probability that two records belong to the same individual - rather than matching by chance or due to coding error. In simple terms the rules determine whether or not an almost-exact match, where the number of variables that agree are used to decide whether or not two records should be linked. Examples of methods of matching based on examples cited in this report include: linked census data which uses method (a), whereas the Generalised Matching Service (GMS) used by the DWP in its construction of the DWPLS which uses method (b).

(ii) Attrition in matching – Records which do not match (whether in full for deterministic matching or under probabilistic rules) will be rejected. This creates (potentially) attrition in linked data and large amounts of data may be lost. Estimates based on linking of business datasets using IDBR common identifiers suggests that only approximately 70 to 80 percent of records match exactly based on simple rules of shared common identifiers (see Harris, 2001 and Davies, 2006).

(iii) Error in matching – this applies to probabilistic matching method where records are assigned to each other on the basis of likelihood rather than exact matching. In this case, errors in matching may be of two types: (a) "false negatives" i.e. the procedure does not match together records that do belong to the same person or (b) "false positives" *i.e.* the procedure matches records together erroneously due to the poor quality or commonality of the identifying fields. Using a probabilistic matching method, the likelihood of type (a) and type (b) errors can be controlled by adjusting rules and tolerance levels on how matches are accepted or rejected.

(iv) Quality of administrative data – the quality of the match will be dependent on the quality of the fields chosen as common identifiers. The problem, potentially, with administrative data is that they may be collected to varying standards and that coding standards and/or methods of transcribing data may vary across departments. In practical terms common identifiers such as name, address, and date of birth may be entered under a different ordering or formatting. Moreover, data entries maybe subject to spelling mistakes, anomalous ordering or mis-recording. More generally administrative data might be of variable quality across departments and subject to missing observation, discontinuities, changes in systems of coding, and so on.

[110] For a discussion of regarding methods for data linkage see Gill (2001).

These considerations highlight the point that linking data from independent administrative source is far from simple. If the datasets contain a unique, high quality identifier such as the National Insurance Numbers (NINOs) then the task is likely to be less onerous, in other cases when matching relies on identification through other (additional) fields then the result is likely to be one of high rates of error and/or data attrition. In addition to this, much additional effort will be needed to clean datasets before merging in addition to work in forming common standards, geographies and time periods. This is likely to be both time consuming and resource intensive.

7.6 An Integrated Population Statistics (IPS) system

7.6.1 The concept of an IPS[111]

The most ambitious vision in terms of data linking is in creating an Integrated Population Statistics (IPS) which would draw together a number of administrative and other liked census and survey data sources in order to produce a population statistics database that links together and contains individual-level information on individuals. The IPS, once developed, could then form the basis of UK social statistics and population research, resulting in improved, standardized and consistent statistical coverage of the population which is updated regularly and automatically as new data is collected. The IPS could form the basis of the Neighbourhood Statistics system and underpin the mid-year population estimates series.

The IPS would result in cost saving along the lines described earlier, as duplication of data collection is avoided. Moreover, it could be held and controlled from one single source (*e.g.* under the stewardship of the ONS).

Under ONS proposals (see: ONS, 2003a) the key elements of the IPS, with rep would be a:

(i) **Population register** – census based (or similar) register of all individuals. At present this could be based either on the National Health Service Central Register (NHSCR) or, for those of working age, the National Insurance Number (NINO). Alternatively this could be facilitated by the establishment of a National Identity Register.

(ii) **Address register** – covering all addressed and linked to the population register by integrating information on household composition[112]. At present no comprehensive database of this kind exists for the UK.

In addition to a system for tracking individual, the system could also be enhanced by cross referencing to:

(iii) Business register. In this instance the IDBR is a readily available (although it is not exhaustive in that it excludes small businesses). Linkages could be made based on the employment details by firm recorded using NINO.

[111] For a more detailed discussion of the Integrated Population System see: ONS (2003a).

[112] One major practical problem here is that household composition is very fluid and individuals moving households and with relatively high frequency. Notification of change of circumstances within the system might be difficult to manage.

Once established, with the appropriate registers in place, the IPS could also provide a vehicle for integrating more extensively other administrative datasets such as those described in this report, relating to education, health, the labour market, and so. For example, the wider use of the NINO would enable existing resources such as those held by DWP and HMRC to be integrated into the system. In essence, the existence of an IPS would provide the framework and spine for linkage of data from these different sources[113]. Moreover, without this kind of resource data linkage is likely to be difficult and more imprecise.

Such a system would need explicit and supporting legislation. It is noted that without radical legislative reform governing the sharing of data across government an IPS will not be possible. In addition, there are a number of significant risks associated with the implementation of the integrated population statistics system. Strong safeguards would have to be put in place ensure confidentiality of individual records and to ensure that rules governing access to the data is well defined. Above and beyond this, an IPS system would have to gain public acceptability. This in itself might be a sticking point given the worries about encroaching government.

7.6.2 Integrated systems in other countries

Other countries outside the UK have successfully implemented an IPS type model. In particular the Scandinavian countries have established national data registers over the past three decades or so. The so called Scandinavian model (for more details see: Smith *et al.* 2004) effectively replaces the census system and minimizes the need for population surveys by introducing an IPS type model based on a standard and unique ID number for each person[114].

Denmark has had a nation-wide Central Population Register (CPR) established since 1968, and has since been supported by the later integration of a Register of Buildings and Dwellings, as well as various other administrative datasets[115]. Using the unique identifier information collected via administrative data is gathered and assigned to individuals in a standardized and consistent fashion. In particular this means that data can be effectively merged without undue concerns regarding accuracy, error or double counting, as is still the case in the UK. Under this system registration of birth and death registrations is compulsory, as is registration of change of address. Therefore the census of population is replaced by a purely register-based system. Finland and Sweden have similar IPS systems with data collection integrated into a single system.

The Netherlands operates a similar model where the Social Statistics Database has been established since 1996, and is supported by new legislation which allows individual records to be linked for statistical purposes. The database uses the population register as a spine. It contains an array of information from administrative data sources, such as benefits data and employee insurance schemes, plus various surveys such as the labour force survey and health interview survey.

[113] Equally household surveys could be added in the same way.

[114] For a discussion of alternatives to the census see ONS (2003b). For a discussion of using administrative records to replicate the US census see Judson *et al.* (2002a).

[115] Denmark's Statistics Act obliges public bodies to make available any information required by Statistics Denmark for statistical purposes.

7.7 Summary

This chapter of the report has demonstrated the potential of administrative data for enhancing data collection and provision in the UK. The attraction of administrative data is that they are already collected and their potential for use may lie dormant. In particular, better use and linking of large scale administrative data sources could offer the potential for substantial cost savings for government. It can also open up new and exciting avenues for statisticians and social researchers. In many respects the UK lags behind many of its continental counterparts, most particularly the Scandinavian countries, which have already established integrated population systems that encompass and integrate administrative data sources.

The major barrier to moving this agenda forward in the UK revolves around the issue of legality. As has been pointed out the legal framework for data sharing is not well defined in the UK and legal developments now lag behind technological developments. This leaves an uncomfortable situation whereby government departments are effectively left to make unilateral decisions on data sharing based on their established protocols and interpretations of the legal constraints placed upon them. Certainly there is no statistical policy as guidance and a UK general statistics act would be very helpful in this regard. It is hoped that the Statistics and Registration Service Bill, currently making its way through parliament, may be useful in this regard.

Other significant barriers to data sharing at this time include resource constraints, given that cleaning, anonymising, linking and/or facilitating access to departmental data all require a significant investment of time. A further and final consideration is the public perception and acceptability of extensive data sharing along the line of the Scandinavian model. Given the increasing concerns about encroaching government, the UK public would need to be convinced that their confidentiality was not being compromised and that data sharing was being used only for legitimate statistical purposes.

8 Conclusions

It should be apparent to even the most casual reader of this report that administrative data have huge potential as research resources, not just in their own right but also through their ability to inform, extend and complement other data resources designed for research purposes. Some administrative data are being developed as important research resources in their own right. Most have potential to extend and add value to existing studies, to validate survey sources and to reduce the interview burden on census and survey respondents. The use that is already being made of such data, recognised through the wide selection of research reports and studies referenced in this report, clearly indicates their value to inform research.

In summarising the findings of this report a number of key observations stand out:

- the scale of administrative data resources available within government departments is extensive. While this selective audit was focussed within a few departments, a large number of different resources has been revealed. A considerable amount of research has already been conducted using these resources, yet they remain relatively undiscovered within the wider social science community.
- By their very nature, existing administrative data resources are growing continually and their complexity is increasing via linkage between them. Furthermore, new developments are underway (*e.g.* 'Connecting for Health') which will have a major impact upon the extent of future resources.
- Access to these resources for research purposes varies from department to department. Some 'gatekeepers' have given careful consideration to access conditions and regulate these conditions tightly. Some have a more relaxed approach to access for *bona fide* research. In some instances, the ESRC has established good links (*e.g.* Celsius), or has funded preliminary work to explore these resources further (*e.g.* PLASC). In other areas, no such links or preliminary work have been established.
- At the time of writing, there remains some uncertainty about whether or not an independent ONS will have the powers to centralise access to all such data resources. While legislation may give the proposed Statistics Board the authority to link data across department resources, this may only happen on a case by case basis and through an Order in parliament.
- Knowledge about the suitability of different administrative data resources for various research purposes is limited and fragmented. There is a need to harness and coordinate this knowledge in a systematic manner.
- The most obvious barriers to the use of these resources for research purpose are:
  - lack of knowledge across the research community about their existence, scope and potential;
  - the difficulty of gaining access for research purposes (knowledge of procedures, gatekeepers);
  - lack of technical skills to handle large scale datasets and/or to establish appropriate linkage mechanisms between difference sources;

  - the general lack of emphasis on 'quality control' procedures, particularly the question of whether or not a specific source of administrative data will provide the additional information required to address a specific research purpose.

It is proposed, therefore, that the ESRC, in collaboration with a number of key government departments (ONS, DWP, DfES, NHS, Home Office) should seek to address these problems directly. One approach would be to establish an **Administrative Data Resources Service**. The role of the service would include:

- agreeing procedures for research access where such procedures do not currently exist;
- harmonising these procedures where possible;
- studying the feasibility of data linkage between different data sources and testing possible linkage procedures;
- promoting the use of administrative resources for research purposes via conference, workshops, websites etc;
- assisting researchers who wishes to make use administrative data resources for research purposes;
- undertaking quality reviews of administration data resources to establish their strengths and weaknesses as research resources;
- assisting in the preservation of linked data sets via archiving and data curation, subject to conditions laid down by data owners;
- liaising with the ONS Virtual Microdata Laboratory and with departments, to study the feasibility of holding administrative datasets within the VML, whilst remaining under the control of the originating department;
- liaising with the Economic and Social Data Service to determine whether or not a wider range of historical and anonymised administrative datasets can be lodged therein.

The proposed **Administrative Data Resources Service** would not seek to become a repository for administrative data, which would remain within the control and protection of the agency responsible for their safe-keeping. It would operate in a fashion similar to CELSIUS, but would work across a variety of government departments and agencies to promote access to and appropriate use of administrative data sources. It would not seek to replicate work being undertaken elsewhere (*e.g.* the SUS), working instead to assist the research community to make better and more use of these rich resources.

References

Aghion Philippe, Rachel Griffith & Peter Howitt (2006). 'The U-shaped relationship between vertical integration and competition: theory and evidence'. IFS Working Papers W06/12, Institute for Fiscal Studies.

Allsopp, C. (2004). '*Review of Statistics for Economic Policymaking*'. Treasury, London.

AMS (2006) '*Personal data for public good: using health information in medical research*'. London: Academy of Medical Sciences.

Axhausen, K. W. (1995). 'Car availability change in England and Wales 1971-1981 - first results from the OPCS Longitudinal Study'. *Transportation* 22 (2): 151-164.

Babb, P. and A. Bethune (1995). 'Trends in births outside marriage'. *Population Trends*; (81): 17-22.

Bartley, M. and I. Plewis (2002). 'Accumulated labour market disadvantage and limiting long term illness: data from the 1971-1991 ONS Longitudinal Study'. *International Journal of Epidemiology*; 31: 336-341.

Ball, J. and M. Marland (1996). 'Male Earnings Mobility in the Lifetime Labour Market Database', Department of Social Security, Analytical Services Division Working Paper No 1.

Barnes, M., Jonathan Haskel and Andrew Ross (2001). 'Understanding productivity: new insights from the ONS business data bank'. Paper presented to the Royal Economic Society Conference, April.

Barnes, Matthew and Ralf Martin (2002). 'Business Data Linking: An introduction', *Economic Trends* No. 581 April.

Barnes Matthew and Jonathan Haskel, (2002). 'Job Creation, Job Destruction and the Contribution of Small Businesses: Evidence for UK Manufacturing'. Working Papers 461, Queen Mary, University of London, Department of Economics.

Barrow, P., P. Waller and L. Wise (2006). 'Comparison of Hospital Episodes with 'Drug-Induced' Disorders and Spontaneously Reported Adverse Drug Reactions'. *British Journal of Clinical Pharmacology*, 61 (2): 233-237 Feb.

Bertoud, R. (2004). 'The profile of exits from incapacity related benefits over time'. DWP Working Paper, Number 17.

Bird, Derek (2004). 'Methodology for the 2004 Annual Survey of Hours and Earnings'. *Labour Market Trends*, Vol 112, no 11, pp 457-464.

Blackwell, L. (2000). 'Gender and ethnicity at work: occupational segregation and disadvantage in the 1991 British Census'. *Sociology*.

Blackwell, L. (2001). 'Occupational sex segregation and part-time work in modern Britain'. Gender, Work and Organisation; 8 (2): 146-163.

Blackwell, L., K. Lynch, S. Jones (2001) 'Science teaching: the demographic squeeze', *Labour Market Trends*, October, 485-493.

Bottle, A. and P. Aylin (2006). 'Mortality Associated with Delay in Operation after Hip Fracture: Observational Study'. *British Medical Journal*, 332 (7547): 947-950 Apr 22.

Boyle Paul, Andrew Cullis, Robin Flowerdew and Vernon Gayle (2004). *UK Data Audit; Phase II*. A Report to the ESRC and the Research Resources Board. May.

Bradley, Steve and Jim Taylor (2004). 'Ethnicity, educational attainment and the transition from school,' Manchester School, Blackwell Publishing, Vol. 72(3), pages 317-346, 06.

Brassett-Grundy, A. (2003). *Researching Households and Families Using the ONS Longitudinal Study*. LS User Guide 20.

Brown, A., S. Harding, A. Bethune and M. Rosato (1998). 'Incidence of Health of the Nation cancers by social class'. *Population Trends*; 90 (Winter 1997): 40-47.

Brown, J., S. Harding, A. Bethune and M. Rosato (1998). 'Longitudinal study of socio-economic differences in the incidence of stomach, colorectal and pancreatic cancers'. *Population Trends*; (94): 35-41.

Burchardt, Tania and Abigail McKnight (2006). *Employment, Welfare and Exclusion*. Centre for Analysis of Social Exclusion (CASE) research programme. http://sticerd.lse.ac.uk/CASE/research/employment.asp.

Burgess S., D. Wilson and A. Briggs (2005). 'The Dynamics of School Attainment of England's Ethnic Minorities'. CMPO Working Paper 05/130.

Buxton, Julian; Lynda Clarke, Emily Grundy and C. E. Marshall (2005). 'The long shadow of childhood: associations between parental social class and own social class, educational attainment and timing of first birth'. *Population Trends* 121, Autumn. Office for National Statistics.

Cabinet Office (2005a). *Transformational Government Enabled by Technology*. Cabinet Office. London, November 2005.

Cabinet Office (2005b). *Making a Difference: Safe and Secure Data Sharing Between Health and Adult Social Care Staff*. Cabinet Office, London.

Cabinet Office (2002). *Privacy and data-sharing: The way forward for public services* Cabinet Office, London.

Cathcart, P., J. Van Der Meulen and J. Armitage et al. (2006). 'Incidence of Primary and Recurrent Acute Urinary Retention between 1998 and 2003 in England'. *Journal of Urology*, 176 (1): 200-204.

Chesher and S. Neisham (2006). 'Review of Literature on the Statistical Properties of Linked Datasets'. DTI Occasional Paper No. 3.

Criscuolo C., J. Haskel and R. Martin (2003), *Building the evidence base for productivity policy using business data linking*. Economic Trends, 600 pp39-51 (ONS, London)

Clarke, L. and H. Joshi (2003). 'Children's changing families and family resources'. In: Jonsen, A.-M., McKee, L., editors. *Children and the Changing Family: Between Transformation and Negotiation*. Falmer Press.

Covizzi I, Gutierrez-Romero R, Noble M. (2006) Evaluating England's 'New Deal for Communities' Programme using the difference in difference Method. *Journal of Economic Geography*. Forthcoming.

Criscuolo C., J. Haskel, R. Martin (2004). 'Import competition, productivity, and restructuring in UK manufacturing', *Oxford Review of Economic Policy* 20 (3): 393-408.

Connolly S., M. Gregory (2002). 'The National Minimum Wage and hours of work: Implications for low paid women', *Oxford Bulletin of Economics and Statistics*, 64: 607-631, Suppl. S.

CST (2005). *Better use of personal information: opportunities and risks*. Council for Science and Technology.

Dale, A., C. Holdsworth (1998). 'Why don't minority ethnic women in Britain work part-time?' In: O'Reilly, J, Fagan, C, editors. *Part-time Prospects: an International Comparison of Part-Time Work in Europe*, North America and the Pacific Rim. London: Routledge. 77-95.

Davies, Elizabeth, Paul Williamson and Clare Holdsworth (2006). 'The leaving of Liverpool: an examination into the migratory characteristics of Liverpool'. Mimeo. University of Liverpool.

Davies, Rhys (2006a), 'Business Structure Database User Guide Version 1'. Mimeo Social and Economic Micro Analysis Reporting Division, Office for National Statistics.

Davies, Rhys (2006b). 'Examining the Nature of Links made between Workplaces Included in the 2004 Workplace Employment Relations Survey and the ONS Annual Respondents Database'. Mimeo. Social and Economic Micro-Analysis Reporting Division, Office for National Statistics.

DCA (2006) *Information sharing vision statement September*. Department for Constitutional Affairs. London. http://www.dca.gov.uk/foi/sharing/information-sharing.pdf

DCA (2003). *A toolkit for Data Sharing*. Department for Constitutional Affairs. London. November 2003. http://www.dca.gov.uk/foi/sharing/toolkit/index.htm.

Dickens, R. (2006) 'Labour Market Mobility in the UK'. Works and Pensions Economics Group (WPEG) Conference.

Dickens R., P. Gregg, and J. Wadsworth (1999). 'New Labour and the labour market', *Oxford Review of Economic Policy*, Vol.16 No.1. 16: 95-113.

Dickens R., P. Gregg, and J. Wadsworth (1997). 'Male earnings mobility in the lifetime labour market database'. CEP Working Paper.

Disney R., J. Haskel and Y. Heden (2003). 'Entry, exit and establishment survival in UK manufacturing', *Journal of Industrial Economics* 51 (1): 91-112 Mar 2003.

Dixon, T., M. Shaw and P. Dieppe (2006). 'Analysis of Regional Variation in Hip and Knee Joint Replacement Rates in England using Hospital Episodes Statistics'. *Public Health*, 120 (1): 83-90.

DOH (2005). *Research governance framework for health and social care*. 2nd ed. London: Department of Health.

Donkin, A. and L. Hattersley (2001). 'Using the Longitudinal Study Cancer Data for Research'. *LS User Guide 19*. Office for National Statistics.

Ekert-Jaffe, O., H. Joshi, K. Lynch, R. Mougin and M. S. Rendall (2002). 'Fertility, timing of births and socio-economic status in France and Britain: social policies and occupational polarisation'. *Population-E*; 3: 475-508.

Emmerson C., Christine Frayne, Sandra McNally and Olmo Silva (2005). 'Evaluation of Aimhigher: Excellence Challenge. The Early Impact of Aimhigher: Excellence Challenge on Pre-16 Outcomes: An Economic Evaluation'. Department for Education and Skills (DfES) Research Report RR652.

Emmerson C., Christine Frayne, Sandra McNally and Panu Pelkonen (2003). 'Economic Evaluation of Excellence in Primary Schools'. DfES Research Report.

Endean R. (1999). 'Patterns of Work, Low Pay and Poverty - Evidence from the Lifetime Labour Market Database and the British Household Panel Study' in Hills J (ed) *Persistent Poverty and Lifetime Inequality: the evidence*, CASE report 5, Centre for the Analysis of Social Exclusion, London School of Economics.

Ewens, D. (2005a). *The National and London Pupil Datasets: An introductory briefing for researchers and research users*. Greater London Authority, Data Management and Analysis Group.

Ewens, D. (2005b). *Moving Home and Changing School: Widening the analysis of pupil mobility*. Greater London Authority, Data Management and Analysis Group.

Fielding, A. J. (1992c). 'Migration and social mobility - South East England as an escalator region'. *Regional Studies*; 26 (1): 1-15.

Fielding, A. J. (1998). 'Counterurbanisation and social class'. In: Boyle, P. J., Halfacre, K. H., editors. *Migration in Rural Areas: Theories and Issues*. Chichester: John Wiley.

Fielding, A. J. (1998). 'Immigrant workers and the class structure of England and Wales: a longitudinal analysis of the social mobility of Britain's black and Asian populations'. In: Sato, M., Fielding, A. J., editors. *The political economy of international labour migration in the post cold-war era*. Tokyo: Dubunkan.

Fieldhouse, E. and E. Hollywood (1999). 'Life after mining: hidden unemployment and changing patterns of economic activity amongst miners in England and Wales, 1981-1991'. *Work Employment and Society*; 13 (3): 483-502.

Galindo-Rueda, Fernando and Jonathan Haskel (2005). 'Skills, workforce characteristics and firm-level productivity in England'. Institute for the Study of Labor (IZA) Discussion Papers 1542.

Garout, M., P. Tekkis and A. Darzi, et al. (2005). 'Comparison of Hospital Episode Statistics with the Association of Coloproctology of Great Britain and Ireland Colorectal Cancer Database'. *British Journal of Surgery*, 92: 155-156 Suppl. 1 Apr 2005

Gill, L., (2001). 'Methods for Automatic Record Matching and Linking and their Use in National Statistics', National Statistics Methodology Series, No. 25, London: Office of National Statistics.

Girma S. and H. Görg (2005). 'Blessing or Curse? Domestic Plants' Survival and Employment Prospects after Foreign Acquisition', CEPR Discussion Paper Number 2994.

Goodson, N., J. Marks and M. Lunt et al. (2005). 'Cardiovascular Admissions and Mortality in an Inception Cohort of Patients with Rheumatoid Arthritis with Onset in the 1980s and 1990s'. *Annals of the Rheumatic Diseases*, 64 (11): 1595-1601.

Griffith R. (1999). 'Using the ARD establishment level data to look at foreign ownership and productivity in the United Kingdom'. *Economic Journal* 109 (456): F416-F442.

GRO (2005). *Citizen Information Project Better sharing of citizen data across the public sector.* General Register Office. London. http://www.gro.gov.uk/cip/.

Grundy, E. (2000). 'Co-residence of mid-life children with their elderly parents in England and Wales: changes between 1981 and 1991'. *Population Studies*; 54: 193-206.

Grundy, E., D. Mayer, H. Young and A. Sloggett (2004). 'Living arrangements and place of death of older people with cancer in England and Wales: a record linkage study'. *British Journal of Cancer*; 91: 907-912.

Gustavo Crespi, Chiara Criscuolo and Jonathan Haskel (2006). 'Productivity, Exporting and the Learning-by-Exporting Hypothesis: Direct Evidence from UK Firms'. CEP Discussion Paper No 726.

Hakim, C. (1994). 'A century of change in occupational segregation 1891-1991'. *Journal of Historical Sociology*; 7 (December): 435-454.

Hamnett, C. (1987). 'Labour and housing market change in London: a longitudinal analysis'. In: Harns, R, Pratt, G, editors. *Housing Tenure and Social Class. The National Swedish Institute for Building Research.*

Harding, S. (1998). 'The incidence of cancers among second generation Irish living in England and Wales'. *British Journal of Cancer*; 78 (7): 958-961.

Harding, S. and E. J. Allen (1996). 'Sources and uses of data on cancer among ethnic groups'. *British Journal of Cancer Supplement*; 29: S17-21.

Harding, S. and M. Rosato (1999). 'Cancer incidence among first generation Scottish, Irish, West Indian and South Asian migrants living in England and Wales'. *Ethnicity and Health*; 2 (1 / 2): 83-92.

Harley, M., M. Mohammed and S. Hussain et al. (2005). 'Was Rodney Ledward A Statistical Outlier? Retrospective Analysis Using Routine Hospital Data to Identify Gynaecologists' Performance'. *British Medical Journal*, 330 (7497): 929-932 Apr 23 2005.

Harris R. and M. Trainor (2005). 'Plant-level analysis using the ARD: Another look at Gibrat's law'. *Scottish Journal of Political Economy* 52 (3): 492-518 Jul 2005.

Harris, Richard; Qian Cher Li. and Catherine Robinson (2005). The productivity impact of skills in English manufacturing, 2001: evidence from plant-level matched data. National Institute of Social and Economic Research Conference Paper. September 2005.

Harris R. (2002). 'Foreign ownership and productivity in the United Kingdom - Some issues when using the ARD establishment level data' *Scottish Journal of Political Economy* 49 (3): 318-335.

Harris, R. I. D. and C. Robinson (2001). 'Research Project on DTI Industrial Support Policies. Contract A: An Analysis of Current DTI Industry Support Patterns'. Mimeo. Department for Trade and Industry. See: www.gla.ac.uk/economics/staff/pdfs/harris_DTI_A.pdf.

Hasluck, C. and J. Bimrose with S-A. Barnes, J. Brown, L. Marris, G. McGivern, M. Orton and R. White. (2006) Evaluation of Skills Coaching trials and Skills Passports DWP Research Report 391.

Hart R. A. (2006). 'Worker-job matches, job mobility and real wage cyclicality'. *Economica* 73 (290): 287-298.

Haskel, Jonathan (2006). 'Unions and Productivity Again: New Evidence from Matched WERS and Business Census Data'. ONS Analysis of Enterprise Microdata Conference Paper.

Hattersley L. and R. Creeser (1995*). Longitudinal Study 1971 - 1991: History, organisation and quality of data*. Series LS No.7 HMSO, London.

Hewson, P. (2005). 'Investigating Population Level Trends in Head Injuries amongst Child Cyclists in the UK'. *Accident Analysis and Prevention*, 37 (5): 807-815 Sep 2005.

Hijzen H., Richard Upward and Peter Wright (2005). 'The Earnings Cost of Business Closure in the UK'. Globalisation and Labour Markets Research Paper Series Research Paper 2005/31.

Holmans, A. E. (2000). 'Divorce, Remarriage and Housing: the Effects of Divorce, Remarriage, Separation and the Formation of the New Couple Households on the number of Separate Households and Housing Demand and Conditions'. London: Department of the Environment, Transport and Regions.

Jenkins A., Rosalind Levacic and Anna Vignoles (2006a). 'Estimating the Relationship between School Resources and Pupil Attainment at GCSE'. DfES Research Report RR727.

Jenkins A., Rosalind Levacic, Anna Vignoles, Fiona Steele and Rebecca Allen (2006b). 'Estimating the Relationship between School Resources and Pupil Attainment at Key Stage 3'. DfES Research Report RR679.

Johnston R, Deborah Wilson and Simon Burgess (2006) 'Ethnic segregation and educational performance at secondary school in Bradford and Leicester'. Centre for Market and Public Organisation Working Paper 06/142.

Jones, Gareth (2000). 'The Development of the Annual Business Inquiry'. *Economic Trends,* no 564.

Joyce, Lucy, Diana Kasparova and David Wilkinson (2006). Evaluation of Basic Skills Mandatory Training Pilots: Synthesis Report. DWP Research Report 385.

Judge, A., J. Chard and I. Learmonth et al. (2006). 'The Effects of Surgical Volumes and Training Centre Status on Outcomes Following Total Joint Replacement: Analysis of the Hospital Episode Statistics for England'. *Journal of Public Health*, 28 (2): 116-124 Jun 2006.

Judson, D. H., and Mark Bauder (2002a). 'Evaluating the Ability of Administrative Records Databases to Replicate Census 2000 Results at the Household Level'. Proceedings of the American Statistical Association, American Statistical Association.

Kalwij A. S. and M. Gregory (2005). 'A panel data analysis of the effects of wages, standard hours and unionization on paid overtime work in Britain'. *Journal Of The Royal Statistical Society*, Series A-Statistics In Society 168: 207-231 Part 1.

Kirby, Simon and Rebecca Riley (2003). 'The employment effects of full participation in ONE'. DWP Research Report 183.

Lam, K., Catrin Ormerod, Felix Ritchie and Prabhat Vaze (2006). 'Do company wage policies persist in the face of minimum wages? An analysis of earnings data for low-paid individuals linked with the characteristics of their employer'. *Labour Market Trends*, vol 114, no 03, pp 69 - 81.

Lindsay. G., S. Pather., and S. Strand (2006). 'Special educational needs and ethnicity: Issues of over- and under-representation. DfES Research Report 757'. Nottingham: DfES.

Liffen, Maslen and Price (1988). *HES Book*. Department of Health.

Lowrance W. (2002). *Learning from experience: privacy and the secondary use of data in health research.* The Nuffield Trust, London.

Machin, S., Shqiponje Telhaj and Joan Wilson. (2006). 'The Mobility of English School Children'. Centre for Economic Performance. Discussion Paper.

Machin, S., S. McNally and C. Meghir (2004). 'Improving Pupil Performance in English Secondary Schools: Excellence in Cities'. Journal of the European Economics Association. Vol. 2, No. 2-3, Pages 396-405.

Mackenbach, J. P., M Huisman and O. Andersen (2004). 'Inequalities in lung cancer mortality by the educational level in 10 European populations'. *European Journal of Cancer*; 40 (1): 126-135.

Maheswaran, R., D. P. Strachan, B. Dodgeon and N. G. Best (2002). 'A population-based case-control study for examining early life influences on geographical variation in adult mortality in England and Wales using stomach cancer and stroke as examples'. *International Journal of Epidemiology*; 31 (2): 375-382.

Martin, Ralf (2005). 'UK plant level energy data'. ONS Analysis of Enterprise Microdata Conference Paper 2005.

Martin Jean, John Bynner, Graham Kalton, Paul Boyle, Harvey Goldstein, Vernon Gayle, Samantha Parsons and Andrea Piesse (2006). *Strategic Review of Panel and Cohort Studies*. A Report to the ESRC and the Research Resources Board.

McCabe, J., A. Jibawi and P. Javle (2005). 'Defining the Minimum Hospital Case-Load to Achieve Optimum Outcomes in Radical Cystectomy'. *BJU International*, 96 (6): 806-810 Oct 2005.

Mitchell, R., S. Gleave, M. Bartley, R. Wiggins and H. Joshi (2000). 'Do attitude and area influence health?' *Health and Place*; 6 (2): 67-79.

Morris M., Simon Rutt and Tilaye Yeshanew (2005). 'Evaluation of Aimhigher: Excellence Challenge Pupil Outcomes One Year on'. Department for Education and Skills (DfES) Research Report RR649.

Morris, Marian and Simon Rutt (2005). 'An Analysis of Pupil Attendance Data in Excellence in Cities (EIC) Areas and Non-EIC EAZs'. Department for Education and Skills (DfES) Research Report RR657.

Moser, K. A. and P. O. Goldblatt (1991). 'Occupational mortality of women aged 15-59 years at death in England and Wales'. *J Epidemiol Community Health*; 45 (2): 117-24.

Nickell S. and G. Quintini (2003). 'Nominal wage rigidity and the rate of inflation'. *Economic Journal* 113 (490): 762-781.

Nicholls M., M. Marland and J. Ball (1997). 'The Department of Social Security's Lifetime Labour Market Database' in "*Jobs, Wages and Poverty: Patterns of Persistence and Mobility in the New Flexible Labour Market*", ed Paul Gregg. Centre for Economic Performance.

Nuttall, M., J. Van Der Meulen and M. Emberton (2006). 'Charlson Scores Based on Icd-10 Administrative Data were valid in Assessing Comorbidity in Patients Undergoing Urological Cancer Surgery'. *Journal Of Clinical Epidemiology*, 59 (3): 265-273 Mar 2006.

Nuttall, M., P. Cathcart and J. Van Der Meulen et al. (2005). 'A Description of Radical Nephrectomy Practice and outcomes in England: 1995-2002'. *BJU International*, 96 (1): 58-61 Jul 2005.

ONS (2005). *Data Sharing for Statistical Purposes. A Practitioners' Guide to the Legal Framework*. Office for National Statistics. London.

ONS (2003a). 'Proposals for an Integrated Population Statistics System', Office for National Statistics Discussion Paper.

ONS (2003b). *Alternatives to a Census: Review of international approaches. Census strategic development Review*. Office for National Statistics.

ONS (October 2003c). *Alternatives to a Census: Rolling Census. Census strategic development Review*. Office for National Statistics, London.

ONS, (2003d). 'Protocol on Data Matching' Draft for Consultation, National Statistics Code of Practice, Office for National Statistics.

ONS (1999). *Tracking People: A guide to longitudinal social sources*. Office for National Statistics. London.

ONS (1998). 'Review of the ONS Longitudinal Study'. ONS Occasional paper 1. Office for National Statistics London

Page, James, Eleanor Breen and Jayne Middlemas (2006) 'Gateway to Work New Deal 25 Plus pilots evaluation' Department for Work and Pensions Research Report No 366

Peach, C. and M. Byron (1993). 'Caribbean tenants in council housing: race, class and gender'. *New Community*; 19 (3): 407-423.

Platt, Lucinda, Ludi Simpson and Bola Akinwale (2005). 'Stability and change in ethnic groups in England and Wales'. *Population Trends* 121, Office for National Statistics.

Rendall, M. S., H. Joshi, J. Oh and G. Verropoulou (2001). 'Comparing the childrearing lifetimes of Britain's 'divorce-revolution' men and women'. *European Journal of Population / Revue Europeenne de Demographie*; 17 (4): 365-387.

Ridley, Kate and Lesley Kendall (2005). 'Evaluation of Excellence in Cities Primary Pilot 2001-2003'. Department for Education and Skills (DfES) Research Report RR675.

Rincon, Ana, Catherine Robinson and Michela Vecchi (2001). 'The Productivity impact of E-Commerce in the UK, 2001: Evidence from microdata'. The National Institute of Economic and Social Research NIESR Discussion Paper No. 257.

Ritchie, F. J. (2005). 'Statistical Disclosure Control in a Research Environment'. Mimeo, Office for National Statistics, London.

Robjohns, J. (2006). 'ARD2: the new Annual Respondents Database', *Economic Trends*, no 630, pp 43-51.

Rogers, Mark (2005). 'R&D and Productivity in the UK: evidence from firm-level data in the 1990s'. Global Conference on Business & Economics, Oxford.

Rudge, J. and R. Gilchrist (2005). 'Excess Winter Morbidity among Older People at Risk of Cold Homes: A Population-Based Study in a London Borough'. *Journal of Public Health*, 27 (4): 353-358.

Roland, M., M. Dusheiko and H. Gravelle et al. (2005). 'Follow Up of People Aged 65 and over with a History of Emergency Admissions: Analysis of Routine Admission Data'. *British Medical Journal*, 330 (7486): 289-292 Feb 5.

Sadun R. and J. Van Reenen (2005). 'Information Technology and Productivity: It ain't what you do it's the way that you do I.T.', EDS Innovation Research Programme Discussion Paper Series EDSDP0002.

Schagen, I. and S. Schagen (2005). 'Combining multilevel analysis with national value-added datasets: a case study to explore the effects of School diversity' *British educational Research Journal* 31(3), 309-328.

Scottish Executive (2004). *Data Sharing: Legal Guidance for the Scottish Public Sector*. Scottish Executive, Edinburgh 2004.

Sharp C., Ian Schagen and Emma Scott (2004). 'Playing for Success: the Longer Term Impact: A Multilevel Analysis'. Department for Education and Skills (DfES) Research Report RR593.

Sloggett, A. (2004). 'Socio-economic and socio-demographic inequalities in cancer incidence and survival in the older population of England and Wales'. Mimeo. Centre for Population Studies.

Smith George, Michael Noble, Chelsie Anttila, Leicester Gill, Asghar Zaidi, Gemma Wright, Chris Dibben and Helen Barnes (2004). *The Value of Linked Administrative Records for Longitudinal Analysis*. A Report to the ESRC and the Research Resources Board.

Stockdale, Brian (2003). 'UK Innovation Survey 2001: Preliminary results of the UK Innovation Survey 2001 which covers enterprises across the whole of the UK business sector (1998 to 2000)'. *Economic Trends*, no 580.

Taylor, J. and A. Nguyen (2006). 'An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value?' *Oxford Bulletin of Economics and Statistics*, 68:2, 203-224.

Walley, Tom (2006). 'Using personal health information in medical research. Overzealous interpretation of UK laws is stifling epidemiological research'. Editorial. *British Medical Journal*, Volume 332.

Warlow C. (2005). 'Overregulation of clinical research: a threat to public health'. *Clinical Medicine*; 5:33-8.

White, C., R. Wiggins, D. Blane, A. Whitworth and M. Glickman. (2005). 'Person, place or time? The effect of individual circumstances, area and changes over time on mortality in men, 1995-2001'. *Health Statistics Quarterly*; 28 (Winter): 18-26.

Williams, M. (2000). 'Migration and social change in Cornwall 1971-91'. In: Creeser, R, Gleave, S, editors. *Migration within England and Wales Using the Longitudinal Study*. ONS Series LS, No. 9. London: The Stationery Office; p. 30-39.

Young, H. (2002). 'Breast Cancer Survival in England and Wales: the Influence of Socio-Economic Status, Social Support and Parity'. Report submitted in partial fulfilment of the requirements of the Master of Science degree in Medical Demography, London School of Hygiene & Tropical Medicine, University of London.

Glossary

ABI	Annual Business Inquiry
ACL	Adult and community learning
AFDI	Annual Inquiry into Foreign Direct Investment
ARD	Annual Respondents Database
ASHE(PD)	Annual Survey of Hours and Earnings (Panel dataset)
BDL	Business Data Linking
BERD	Business Enterprise Research and Development
BHPS	British Household Panel survey
BSCI	Business Spending on Capital Items
BSD	Business Structure Database
CELCIUS	Centre for Longitudinal Study Information and User Support
CIS	Community Innovation Survey
CMPO	Centre For Market And Public Organisation
DfES	Department for Education and Skills
DoH	Department of Health
DPA	Data Protection Act
DWP	Department for Work and Pensions
EiC	Excellence in Cities
EMA	Education Maintenance Allowance
ESF	European social funding
ESRC	Economic and Social Research Council
ESS	Employers Skills Survey
ETP	NHS Electronic Transmission of Prescriptions
FE	Further Education
FSM	Free School Meals
GCSE	General Certificate of Secondary Education
GMS	Generalised Matching Service
HE	Further Education
HES	Hospital Episode Statistics
HESA	Higher Education Statistics Agency
HMRC	Her Majesty's Customs and Excise
HRA	(Human Rights Act
IC	NHS Information Centre for Health and Social Care
ICT	
IDBR	Inter-Departmental Business Register
ILR	Individual Learner Record
IOP	Index of Production
IPS	Integrated Population Statistics
ISR	Individualised Student Record
JUVOS	Joint Unemployment and Vacancies Operating System
KS	(National Curriculum) Key Stage
LEA	Local Education Authority
LFS	Labour Force Survey
LLMD	Lifetime Labour Market Database
LS	ONS Longitudinal Study
LSC	Learning and Skills Council
LSRB	LS Research Board

MPI	Monthly Production Inquiry
NCIC	National Cancer Intelligence Centre
ND	New Deal programme
NES	New Earnings Survey
NFER	National Foundation for Educational Research
NHS	National Health Service
NHSCR	National Health Service Central Register
NHSCRS	NHS Care Records Service
NI	National Insurance
NINO	National Insurance Number
NIRS	National Insurance Recording System
NPD	National Pupil Database
NPfIT	National Programme for IT
NWCS	NHS-Wide Clearing Service
OLS	Ordinary Least Squares
ONS	Office for National Statistics
ORLS	Oxford Medical Records Linkage Study
PAS	Patient Administration System
PAYE	Pay As You Earn
PENSIM	DWP Pensions Simulation Model
PLASC	Pupil Level Annual School Census
QCES	Quarterly Capital Expenditure Survey
QFI	Quarterly Fuels Inquiry
ROCR	NHS Review of Central Returns
SEN	Special Education Needs
SFR	Statistical First Release
SIC	Standard Industrial Classification
SLS	Scottish Longitudinal Study
SOC	Standard Occupational Classification
SUS	Secondary Uses Service
TLRP	ESRC Teaching and Learning Research Programme
UCAS	Universities and Colleges Admissions Service
UPN	Unique Pupil Number
VAT	Value Added Tax
VML	Virtual Microdata Laboratory
WASD	DWP Working Age Statistical Databases,
WBL	Work based learning
WERS	Workplace Employment Relations Survey
WPLS	DWP Works and Pensions Longitudinal Study

Annex 1 Education Datasets

Annex 1.1 Local Education Authority (LEA) Reference Table Data

Variable
LEA Number
LEA Name
Post Code of LEA
Address and Contact details of LEA
Government Office Region Indicator

Annex 1.2: School Reference Table Data

Variable
School Unique Reference Number
Local Education Authority (LEA)
DfES Establishment Number
Further Education/Higher Education Identification Number
School Name
School Locality
Post Code of School
Open/Closed Indicator
School Opening Date
School Closure Date
Head teacher
Type of Establishment Indicator
Further Education Type Indicator
Phase of Education Indicator
Statutory Lowest Age of Pupils
Statutory Highest Age of Pupils
Gender Indicator
Total Number of Girls
Total Number of Boys
Religious Character Indicator
Approved Number of Day Pupils at Special School
Admissions Policy Indicator
Special Classes Indicator
Boarders Indicator
Nursery Provider Indicator
Total Number of Government Funded Children at EY Setting
School Capacity
Diocese Indicator
Urban / Rural Indicator
Government Office Region Indicator
Parliamentary Constituency Indicator
Easting Grid Reference
Northing Grid Reference
Ward Indicator
District Indicator
Police Area Indicator
Travel to Work Area Indicator
Learning Skills Council / Connexions Area Indicator
Specialist School Indicator
Special Measures Indicator
Education Action Zone Indicator
Beacon Status Indicator

Variable
Excellence in Cities Indicator
Excellence In Cities Group Indicator
Learning Skills Unit Indicator
EiC City Learning Centre Indicator
EiC Action Zone Indicator
Fresh Start Indicator
Training School Indicator
Investors in People Indicator
Early Excellence Centre Indicator
Private Finance Initiative Indicator
Sixth Form Indicator
Early Year Setting Indicator
PRU Teenage Mothers Indicator
PRU Childcare Indicator
PRU SEN Indicator
PRU EBD Indicator
PRU Number of places
PRU FT Provision
PRU Pupil Educated by Others
Lowest Age of Pupils on ASC
Highest Age of Pupils on ASC
Approved Number of Boarders at Special School
Combined Specialism Indicator
Leading Edge Partnership Indicator
Leadership Incentive Grant Indicator

Annex 1.3: Foundation Stage Profile Data

Variable
Academic Year Unique Pupil Number (UPN) Unique Establishment Number LEA Name Date of Birth Gender Personal, Social and Emotional Development (PSE) Scores: • Communication, Language and Literacy (CLL) • Mathematical Development (MAT) • Knowledge and Understanding of the World • Physical Development • Creative Development

Annex 1.4: Pupil Level Annual School Census (PLASC) Data

Variable
Academic Year Unique Pupil Number Unique Establishment Number LEA Name Date of Birth Gender Age at start of academic year Ethnicity Pupil's Post Code Special status indicators: • Part Time • Boarder National Curriculum year group First Language Free School Meal Eligibility Special Educational Needs (SEN) Status and SEN type Permanent Exclusion Indicator In Care indicator and Care Authority (if applicable) Key Stage 4 and 5 Qualifications • A levels • GCSEs • GNVQ Marker • NVQ Marker • Other Qualification Marker

Annex 1.5: Key Stage Attainment Data

Variable
General Academic Year Unique Pupil Number (UPN) Unique Establishment Number LEA Name Date of Birth Gender School Type Source Country and language (England or Wales)
Key Stage 1 Overall teacher assessment. Passed relevant NC level (yes/no). Total test point score. English (or Welsh): • Speaking and listening target • Reading attainment target • Writing attainment target Maths: • Using and applying maths target • Number and algebra attainment target • Shapes, space and measures attainment target Science: • Experimental and investigative science attainment target • Life processes and living things target • Materials and their properties target • Physical processes target Key Stage 1 average point score
Key Stage 2 Overall teacher assessment. Passed relevant NC level (yes/no). Total test point score. English (or Welsh): • Reading test • Writing test • Handwriting test • Spelling test Maths: • Maths test A • Maths test B • Maths extension test Science: • Science test A • Science test B • Science extension test Key Stage 1 average point score Key Stage 2 average point score
Year 7 Progress Test English test level and mark English reading test level and mark English writing test level and mark English handwriting test mark

Variable
English spelling test mark
Maths test A mark
Maths test B mark
Mental arithmetic test mark
Key Stage 3
Overall teacher assessment.
Passed relevant NC level (yes/no).
Total test point score.
English (or Welsh) reading test
English (or Welsh) writing test
English extension test
Maths test
Paper 1
Paper 2
Arithmetic paper
Science test
Paper A
Paper B
Key Stage 1 average point score
Key Stage 2 average point score
Key Stage 3 average point score
Key Stage 4 (GCSE)
Level 1 threshold indicators at GCSE/GNVQ
Level 2 threshold indicators at GCSE/GNVQ
1 or More A* to G GCSE or equivalent Indicator
5 or More A* to G GCSE or equivalent Indicator
Number of Full GCSE Entries
Number of GCSE Entries in vocational subjects.
Number of Full Intermediate GNVQ Entries.
Number of Full Foundation GNVQ Entries.
Number of Part 1 Intermediate GNVQ Entries.
Number of Part 1 Foundation GNVQ Entries.
Number of Intermediate Language Unit GNVQ Entries.
Number of Foundation Language Unit GNVQ Entries.
Number of Entry Level Qualification Entries.
Number of Key Skills Level 1 Entries.
Number of Key Skills Level 2 Entries.
Number of GCSE passes by grade
Number of GCSE vocational course passes by grade
Number of Intermediate GNVQs passes by grade
Number of Intermediate GNVQs passes by grade
Number of Foundation GNVQs passes by grade
The Highest Grade English level achieved
The Highest Grade Maths level achieved
The Highest Grade Science level achieved
Total A or A* GCSE passes (range from 0 to 18)
GCSE grade by subject
5 or more A*-G GCSE/GNVQ passes by subject
Key Stage 1 average point score
Key Stage 2 average point score
Key Stage 3 average point score
GCSE median points score
By subject:
Candidate Serial Number
Awarding body exam subject code

Variable
Awarding Body Code Qualification and Assessment Code (3 digit code) Exam Serial Number Exam Grade
Key Stage 5 (GCSE) Number of GCE A level entries by grade Number of GCE AS level entries by grade Number of VCE A level entries by grade Number of passes at GCE A level Number of passes at VCE A level Number of passes at GCE AS level Grade achieved at GCE A level by subject Grade achieved at VCE A by subject Grade achieved at GCE AS by subject By subject: Candidate Serial Number Awarding body exam subject code Awarding Body Code Qualification and Assessment Code (3 digit code) Exam Serial Number Exam Grade

Annex 1.6: HESA datasets list of major variables

STUDENT DATASET
A/AS-level/Highers points score and tariff points score
Age
Classification obtained by first degree qualifiers
Cost centre
Disability
Domicile
Ethnicity
Expected length of study programme
First year indicator
FTE - Student full-time equivalence
Gender
Highest qualification on entry
Level of study/Qualification aim
Location of institution
Major source of tuition fees
Mode of study
Subject/Subject area of study
FIRST DESTINATION DATASET
Activity
Qualification required for job
Location of employment
Employer size
Institution of further study
Type of qualification of further study
Mode of further study
Standard Industrial Classification (SIC)
Standard Occupational Classification (SOC)

Annex 1.7: Widening Participation in Higher Education: a quantitative analysis: part of the economic and Social Research Council (ESRC) Teaching and Learning Research Programme (TLRP)

Anna Vignoles, University of London, Institute of Education; Alissa Goodman, Institute of Fiscal Studies; Stephen Machin, London School of Economics; Sandra McNally, London School of Economics

Project Summary

The drive to widen participation in higher education in the UK is longstanding yet there is still remarkably little in the way of large-scale quantitative analysis of many aspects of this issue. This research aims to develop a theoretically based quantitative empirical analysis of the higher education experience of different students, particularly disadvantaged students, ethnic minorities, women, those entering HE without A levels and mature students. The analysis will be multi-disciplinary, building on economic, education and geographical theories of educational attainment, and will provide a life course perspective to the issue of widening access to higher education. The research will use a variety of data sets, largely longitudinal, and will provide an international perspective to some issues. In addition, we propose to combine two key UK administrative data sets to provide us with a uniquely comprehensive account of the educational attainment of an entire cohort of young people in the UK . Our objective is to determine at what age interventions aimed at widening access to higher education need to be focused? In particular, when do socio-economic, ethnic and gender gaps in educational attainment emerge? Do students from different family backgrounds and with differing prior attainment have different perceptions of their ability to benefit from higher education? For those who do enter HE, do they experience a very different form of higher education from their more advantaged peers, particularly in terms of subject studied and quality of institution? How does gender affect degree choice and how does this compare to other countries? Can we better understand these choices using economic models of educational investment? For students who do enter higher education, how do progression rates vary across different types of student and across different subjects and quality of institution?

Annex 2 Labour Market Datasets

Annex 2.1 Main Variables in WPLS

INDIVIDUAL CHARACTERISTICS
Age
Gender
Ethnicity
Place of residence (post code and house number)
BENEFITS DATA
Child Benefit
Disability Living Allowance
Attendance Allowance, Carers Allowance
Industrial Injuries Benefits
Income Support
Job Seekers Allowance
Overseas Invalidity Benefit
Pensions
Incapacity Benefit
Pension Credit
Winter Fuel Payments
Maternity Allowance
Bereavement Benefit
Housing Benefit*
Council Tax Benefit*
Home Office Prisoner Data (England and Wales) *
London Electricity and Veterans Agency*
EMPLOYMENT SPELLS
Employment records (employer based and self assessment)*
New Deal and other work-related activity
Earnings*
SAVINGS AND PENSIONS PROVISION (INDIVIDUALS AGE 60 OR OVER ONLY)[116]
Individual Savings Accounts (ISAs) *
Personal Equity Plans (PEP) *
Tax Exempt Special Savings Account (TESSA) *
Private pensions*
Savings accounts information*

Notes: (1) * Provided by HMRC; (2) The database also holds a 1 percent sample of all NI contributions from 1975 onwards.

Annex 2.2 100% Benefit Databases (Working Age Statistical Databases, WASD)

These databases were created to evaluate the impact the introduction of Jobcentre Plus offices had on the Labour Market. The database holds information about client's claims and spells on the main DWP benefits from June 1999 and is updated every six weeks. The data contained comes from the different benefits systems and covers the following benefits - JSA, IS, IB, SDA, ICA, WB/BB, DLA, PC, RP and AA.

- The 'National Database' holds claim level information on what benefits a person has claimed (GB only)

- The 'Spells Database' holds a record for each spell someone has had on benefit (i.e. one or more consecutive benefit claims)

[116] The provision of this information is facilitated for pension planning purposes under the Pensions Act, 2004.

- The 'Client Group History Table' contains a complete history of which client groups a person has belonged to during their time of benefit. It also records what their previous and next client group was.

- The 'JCP Database' is a subset of the 'National Database' and contains claim level information for client's of the 17 JCP pathfinder areas.

- 100% Benefit Database – This database is used for the evaluation of benefit claimants.

Annex 2.3 DWP Programme and Other Related Datasets

1. Access to Work

This dataset is used in the evaluation of Access to Work. The Access to Work programme provides advice and practical help for disabled people with the additional employment costs which result from disability, for example, in travelling to work, adapting the workplace, obtaining special equipment and support workers.

2. Basic Skills Programme

The dataset includes individuals who engage (via jobcentre intervention) in any part of the basic skills process. The dataset is used in the evaluation of client literacy and numeracy levels and the numbers of those screened for basic skills need, attending independent assessment or starting a basic skills course. From April 2001, Jobcentre Plus has been tasked with carrying out additional activity aimed at those with Basic Skills needs. Clients who need help with their basic skills can access suitable training which can be job focused to ensure they can develop the skills they need to find and retain work.

3. Childcare Barriers to work

The Childcare Barriers to work dataset will be used to monitor and evaluate the use of the Childcare Partnership marker. With the introduction of the Childcare Partnership Managers, it became apparent that no measure of the barriers effecting lone parents was being recorded. To combat this Jobcentre Plus introduced a new marker, to be used as a management tool, to determine areas where childcare barriers are present for Jobcentre Plus customers. From April 2004 it is mandatory for an advisor to record details of the barriers the customer has in finding work and the type of childcare required.

4. Employment Retention & Advancement Scheme (ERA)

The Employment Retention & Advancement Scheme Dataset data includes Jobcentre Plus clients eligible for New Deal 25 Plus and those volunteering for New Deal for Lone Parents. Also included are Lone Parents on Working Tax Credit (WTC) working part time at low wage jobs. The ERA scheme tests a new strategy for improving job retention and advancement for low wage workers in Britain. Using a random assignment model, ERA is directed at individuals in three low income groups known to have difficulty retaining jobs or advancing to better positions.

5. Employment Zones

The Employment Zones dataset includes individuals eligible for the employment zones programme Employment Zones were introduced in April 2000 to 15 areas with consistently high levels of long term unemployment. They pool funds for training, Jobcentre Plus support and the equivalent of benefit to maximise flexibility and give individuals more say in the choices which affect them. They are designed to help long term unemployed people to find sustainable employment.

6. Ethnic Minority Outreach (EMO)

The Ethnic Minority Outreach provider dataset contains details on all customers who have engaged with an EMO provider and their progress. Formerly known as Minority Ethnic Outreach, EMO started in April 2002. It is a programme design to engage and assist ethnic minorities into the labour market and to promote the benefits of a diverse work force to businesses. The programme is delivered by external contractors.

7. The European Social Fund (ESF)

The European Social Fund Dataset is used to supply information to the nine Jobcentre Plus regions which enables them to produce the Interim Claims and Project Closure Reports necessary to enable Jobcentre Plus to receive the appropriate ESF funding.

The European Social Fund (ESF) provides money to organisations (known as Co-financing Organisations) which have successfully bid for funding to be used to enhance and improve employment opportunities for several groups including; young people, people unemployed six months or longer, and women. Each of the nine English Jobcentre Plus Regions are operating as Co-financing Organisations whereby ESF money and domestic funding is brought together and administered by them to provide appropriate training and support for eligible customers.

8. Extended Schools Childcare

The Extended Schools Childcare dataset is used for the evaluation of three elements of the pilot, childcare chats, childcare tasters and extended schools childcare. This pilot is aimed at predominantly lone parents, but also any workless parents and partners. These elements encourage parents to consider returning to work, by addressing their childcare issues and provide enough formal childcare places to satisfy the needs of the parents. The childcare chat is available to all parents that are unemployed or working less than 16 hours living in the pilot areas. The Childcare Taster allows the parent to experience five days worth of childcare. This is available for customers participating in New Deal for Lone Parents (NDLP)/New Deal for Partners (NDP) and lone parents within Employment Zones (EZ). Extended Schools Childcare is the expansion of formal childcare places. Schools are encouraged to develop new and existing provision, to be made available outside the normal school day. Jobcentre Plus advisors can refer lone parents to these formal places and is available to all parents that are unemployed or working less than 16 hours living in the pilot areas.

9. Incapacity Benefit Reforms

The Invalidity Benefit Reform dataset contains all those customers for whom additional advisory support may be beneficial. Incapacity Benefit Reform activity will engage individuals and provide effective support. IB Personal Advisers (IB PA) equipped with a much broader set of skills and using a 'Screening Tool' identify those customers for whom additional advisory support is not appropriate because they are only likely to be on benefit for a short period. IB Reforms started on 27th October 2003 in three districts. A further four districts started on 5th April 2004.

10. Job Retention and Rehabilitation Pilots (JRRP)

The Job Retention and Rehabilitation Pilots dataset will be used by Social research for evaluation of the JRRP pilot. Job Retention and Rehabilitation Pilots will run from early 2003 to 2005. The pilots will test different strategies to help people to return to work. These include a healthcare intervention, a workplace intervention and a combined approach. Eligible people (those who have been off work for between 6 and 26 weeks and are deemed to be at risk of losing their jobs) will be randomly assigned to one of three intervention groups or a control group.

11. Jobcentre Plus pathfinder

This database is used to evaluate Jobcentre Plus Pathfinders. In April 2002 the Jobcentre Plus scheme was introduced to achieve the government's aim of bringing together the Employment Service and the Benefit Agency, building on the ONE scheme which started before this integration. The role of the JobCentre Plus scheme was to deliver a labour market focused service through Work Focused Interviews and encourage and help people back into work.17 pathfinder pilot sites were set up and this database records the activity within these pilot sites (i.e. meetings booked, attended, job submissions, numbers entering and exiting JCP). The database has recently been extended to include data on all rolled out offices.

12. Jobcentre Plus vacancies and Jobcentre Plus employer

This dataset is used to support analyses of Jobcentre Plus performance in terms of vacancy taking, vacancy handling, and vacancy filling. Details of all vacancies notified to Jobcentre Plus (since Jan 2000) are kept, including information on whether the vacancy is permanent or temporary, full-time or part-time, and by occupation. Details of all employers who have notified vacancies to Jobcentre Plus (since Jan 2000) are recorded, including information on industrial sector.

13. Joint Claims

The Joint Claims Eligible Population dataset contains all couples identified from benefit systems that are eligible for the Joint Claims programme. From 19th March 2001, couples without dependent children and where both partners are over the age of 18, and at least one was born on or after 19th March 1976 have been required to make a Joint Claim for JSA. Where the couple makes a Joint Claim, both partners are required to be available for work and to seek work actively. From 28th October 2002, this requirement was extended to couples where at least one partner was born after 28th October 1957.

14. New Deal 50 plus (ND50+)

The New Deal 50 plus dataset provides information to assist in the evaluation of the New Deal 50+ program. It is currently being redesigned to better assist analysts. New Deal 50 plus (ND50+) is a voluntary programme where customers aged 50 and over can receive specialist advice in assisting them back into work.

15. New Deal for Disabled People (NDDP)

This dataset includes all clients who volunteer for NDDP. The data is used in the evaluation of NDDP and in the production of National Statistics. New Deal for Disabled People (NDDP) offers support to help people with health conditions and disabilities to find and keep jobs. NDDP is voluntary and offers eligible people access to a network of Job Brokers from private, voluntary and public sector organisations. A number of datasets are derived from the NDDP data.

16. New Deal for Lone Parents (NDLP)

New Deal for Lone Parents (NDLP) aims to help lone parents to overcome barriers into work and improve their job readiness. NDLP was introduced as a voluntary programme nationally in October 1998 for those lone parents claiming Income Support. It was extended in November 2001 for all lone parents who are not working or who are working less than 16 hours per week.

17. New Deal for Long Term Unemployed (NDLTU)

New Deal for Long Term Unemployed Dataset is used for the evaluation of the New Deal programme and for the publication of National Statistics. New Deal for Long Term Unemployed (NDLTU), was introduced nationally on 29 June 1998 for those who had been claiming Jobseekers Allowance (JSA) for at least 2 years. From April 2001, NDLTU was extended and enhanced to provide clients with access to a greater and more tailored range of support and provision. Eligibility has been extended to include those who had been claiming Jobseekers Allowance for 18 months or more.

18. New Deal for Partners (NDP)

The New Deal for Partners Participants database contains all partners that have gone through the New Deal for Partners programme. New Deal for Partners (NDP) was re-launched in April 2004. The program offers non-working partners of working age persons claiming Jobseekers Allowance (excluding Joint Claims couples), Income Support, Incapacity Benefit, Severe Disablement Allowance or Carer's Allowance a similar level of help and support currently available to Lone Partners through the New Deal for Lone Parents programme.

19. New Deal for Young People (NDYP)

New Deal for Young People Dataset is used for the evaluation of the New Deal programme and for the publication of National Statistics. New Deal for Young People (NDYP) was introduced nationally on 6 April 1998 and is aimed at those aged 18-24 who have been claiming Jobseekers Allowance (JSA) for at least 6 months. Eligible clients will progress through a Gateway period, lasting up to 4 months, where advisers will work to improve employability and to find unsubsidised jobs for as many as possible. Those who do not find a job will move into one of four options: Subsidised employment, Full-time education/training, Environmental Task Force or Voluntary Sector jobs. Clients' reaching the end of their option without obtaining a job, will enter a follow-through period, where they receive intensive help to find a job.

20. Progress to Work

The Progress to Work Dataset includes eligible clients on working age benefits. Also persons out of work and not claiming benefits, who are eligible for progress2work help. The Progress2work LinkUP Dataset includes eligible clients on working age benefits. Also persons out of work and not claiming benefits, who are eligible for progress2work LinkUP help. Non Jobcentre Job Dataset is used to measure partner/provider job entries and by Jobcentre offices to monitor Jobcentre Plus Targets. This extract was created in June 2004 and contains all individuals who have gained a job through a partner/provider rather than through the jobcentre. It contains the dates of all job entries and referrals, including the status of a client and their reason for ending their claim. It also contains retention data by looking at whether someone has remained off JSA once they have entered a job.

Progress2work (p2w) provides additional support to help unemployed ex-drug misusers into work. It includes specialist support to help people through mainstream programmes p2w-Link-UP has been developed to provide further support for those facing the greatest disadvantages in the labour market. This includes recovering alcohol misusers, ex-offenders and homeless people. During Autumn 2002, p2w Link UP pilots were established in 9 locations including four of the designated street crime areas (Avon & Somerset, Merseyside, Metropolitan, and South Yorkshire). During 2003 pilots were established in a further 14 districts ensuring full coverage of all designated street crime areas. Building on the progress2work model, the initiative supports people by providing specialist advice and offering access to appropriate specialists.

21. Social Fund

Social Fund Dataset holds a 100% record of budgeting loan applications since 1999, giving the amount applied for, the factors governing the amount offered and the amount accepted by the client. It is used for regular monitoring of the social fund loans budget and for answering ad hoc queries such as parliamentary questions and requests for policy advice. Since 1999, the maximum amount of budgeting loan a client can receive has been governed by the client's family composition, benefit history and existing debt to the fund.

22. Step up

The Step Up dataset is used to create a sample for comparison of those who have entered Step up and also to monitor how many people would be eligible to enter it if it were no longer a pilot programme. Step Up Participants Data contains all individuals who have been identified as eligible for Step UP and who are now either in a stage of the Step UP programme or who are waiting to be placed in a job. The Step Up datasets contains data post April 2000. Step Up provides transitional jobs to act as a stepping stone for long term unemployed people moving from benefits into work. There are a total of Six Pilot areas.

23. Sustainability extract

The Sustainability Dataset is used to measure job entries and by Jobcentre offices to monitor Jobcentre Plus Targets. The sustainability extract was introduced in 1998 and contains every individual with an LMS record and job entry. It contains the dates of all job entries and referrals, including the status of a client and their reason for ending their claim. It also contains retention data by looking at whether someone has remained off JSA once they have entered a job.

24. Work Based Learning for Adults (WBLA)

Work Based Learning for Adults Dataset - This dataset is used for the publication of National Statistics and evaluation/analyses of Work Based Learning for Adults programme. Work Based Learning for Adults (WBLA) is a voluntary full-time training programme mainly aimed at people aged 25 and over who have been unemployed for six months or longer and are claiming Jobseeker's Allowance or another qualifying benefit. Jobcentre Plus became responsible for the delivery of WBLA from April 2001. As well as the transfer of responsibility from TECs, the programme was changed so that help was focused on those who most needed it, increasing in intensity with length of time unemployed and narrowing access to Basic Employability Training to people with basic skills levels below Entry Level.

25. Work Focused Interviews for Partners (WFIP)

The Work Focused Interviews for Partners Eligible Population dataset contains all spells where a claimant has included a partner in their benefit claim, back to June 1999. Work Focused Interviews for Partners Participants – This dataset contains all partners that have been identified for a WFIP from the Labour Market System and have gone through the WFIP process. Mandatory Work Focused Interviews for Partners (WFIP) were introduced in April 2004, in Jobcentre Plus areas only. WFIP is for partners of working age benefit customers claiming Income-based Jobseekers Allowance, Income Support, Incapacity Benefit or Severe Disablement Allowance. The claimant must have been continuously in receipt of benefit for a minimum of 6 months and receiving an Adult Dependency Increase for the partner at that time.

26. Workstep Programme

The Workstep dataset includes all persons engaged on the Workstep programme and is used for evaluation purposes. Workstep, introduced in April 2001, provides job support to disabled people who face more complex barriers to finding and keeping work. Workstep clients are able to make a valuable contribution in their job and, where appropriate, develop and progress to unsupported employment.

Annex 3 Health Related Data

Annex 3.1 HES Data Dictionary

PATIENT

Administrative and legal status of patient
Administrative category
Age at end of episode
Age at start of episode
Baby's age in days
Date of birth - patient
Date of birth check flag -
Ethnic origin
Finished consultant episode flag
Legal category of patient
Local patient identifier
NHS Number
Patient ID - HES generated (original)
Patient identifier - HES generated
Postcode district of patient's residence
Postcode of patient
Sex of patient

PERIOD OF CARE

Admission date check flag
Bed days within a year
Beginning of spell
Date episode ended
Date episode started
Date of admission
Date of decision to admit
Date of decision to admit check flag
Date of discharge
Destination on discharge
Discharge date check flag
Duration of spell
End of spell
Episode duration
Episode order
Episode start date check flag
Episode status
Episode type
First regular day or night admission
Hospital provider spell number
Intended management
Main specialty
Method of admission
Method of discharge
Patient classification
Source of admission
Treatment specialty
Waiting time
Ward type at start of episode

ADMISSIONS

Admission date check flag
Date of admission
Date of decision to admit
Date of decision to admit check flag
First regular day or night admission
Method of admission
Source of admission
Waiting time

DISCHARGES

Date of discharge
Destination on discharge
Discharge date check flag
Method of discharge
EPISODES AND SPELLS
Bed days within a year
Beginning of spell
Date episode ended
Date episode started
Duration of spell
End of spell
Episode duration
Episode end date check flag
Episode order
Episode start date check flag
Episode status
Episode type
Hospital provider spell number
Ward type at start of episode
HEALTHCARE RESOURCE
Dominant procedure
Healthcare resource group (HES generated)
NHS-generated HRG code
NHS-generated HRG code version number
CLINICAL
Date of operation
Diagnosis
External cause code - four characters
External cause code - three characters
External cause of injury or poisoning
Intended management
Main operation - three characters
Main specialty
Operation codes
Operation status code
Patient classification
Post-operative duration
Pre-operative duration
Primary diagnosis - four characters
Primary diagnosis - three characters
Treatment specialty
GEOGRAPHICAL
Census output area, 2001
Census output area, 2001 (6 character)
County of residence
Current electoral ward
Electoral ward in 1981
Electoral ward in 1991
Electoral ward in 1998
Government office region of residence
Government office region of treatment
Health authority of residence
Health authority of treatment
Local authority district
Local authority district in 1998
Lower super output area
Middle super output area
Ordnance survey grid reference
Patient's health authority of residence, provided by NHS
Patient's primary care trust of residence
Patient's strategic health authority of residence

Primary care trust of treatment Region of treatment Regional office of residence Rural / urban indicator Site code of treatment Strategic health authority area of treatment Super output
MATERNITY (IF APPLICABLE) Anaesthetic given during labour or delivery Anaesthetic given post-labour or delivery Antenatal days of stay Baby sequence number Baby's age in days Birth date (baby) Birth order Birth weight Change of delivery place Delivery method Delivery place (actual) Delivery place (intended) First antenatal assessment date Gestation period in weeks at first antenatal assessment Length of gestation Live or still birth Method to induce labour Mother's age at delivery Mother's date of birth Mother's date of birth check flag Neonatal level of care Number of babies Number of baby tails Number of previous pregnancies Postnatal stay Resuscitation method Sex of baby Status of person conducting delivery
CRITICAL CARE (IF APPLICABLE) ACP sequence number Augmented care location Augmented care period disposal Augmented care period end date Augmented care period indicator Augmented care period local ID Augmented care period number Augmented care period outcome indicator Augmented care period source Augmented care period speciality function code Augmented care period start date High-dependency care level Intensive care level days Number of augmented care periods within episode Number of organ systems supported
PSYCHIATRIC (IF APPLICABLE) Age at psychiatric census Carer support indicator Date detention commenced Date detention commenced check flag Detention category Duration of care to psychiatric census Duration of detention Legal group of patient

Legal group of patient (psychiatric) Legal status classification Marital status (psychiatric) Mental category Psychiatric patient status Status of patient included in psychiatric census V code indicator Ward type at psychiatric census date
ORGANISATION Commissioner code Commissioner code status Commissioner's regional office Commissioner's strategic health authority Commissioning serial number Health authority area where patient's GP was registered Primary care group Primary care trust area where patient's GP practice was registered Primary care trust of responsibility Provider type Regional office area where patient's GP practice was registered Strategic health authority area where patient's GP practice was registered
PRACTITIONER Code of GP practice Code of patient's registered or referring general medical practitioner Consultant code Person referring patient Referring organisation code
SYSTEM Combined grossing factor Coverage grossing factor Date data received by NHS-Wide Clearing Service Ethnic category (audit version) Ethnic character (audit version) NHS Number indicator Origin of primary care group Origin of primary care trust of responsibility Patient identifier (HES generated) - basis of match Record identifier

Annex 3.2 National Cancer Registrations Database

The following data fields are contained in the Cancer Registration Minimum Data Set (CRMDS) supplied from the regional registries to the ONS.

Registration details (identity number)
Patient's name
Patient's previous surname
Patient's address
Post code
Sex
NHS number
Marital status
Incidence date
Site of primary growth
Type of growth
Behaviour of growth
Multiple tumour indicator
Date of birth
Date of death (if dead)
Basis of diagnosis*
Death certificate only indicator*
Side (laterality)*
Treatment(s) (indicators)*
Stage*
Grade*
Optional Data :
Ethnic origin*
Country of birth
Patient's occupation
Patient's employment status
Patient's industry
Head of household's occupation
Head of household's employment status
Head of household's industry
Registration from screening*

* Available from 1993 onwards

Annex 3.3 Other Cancer Related Datasets

Dataset	Owner / Sponsor
Urology Dataset	BAUS (The British Association of Urological Surgeons)
Bolger Dataset (Gynae)	Brendan Bolger (2000)
Upper GI Datasets	AUGIS (Association of Upper Gastrointestinal Surgeons of Great Britain & Ireland)
Head & Neck Cancer Audit	BAHNO (British Association of Head & Neck Oncologists)
Breast Cancer Dataset	BASO (British Association of Surgical Oncologists)
Lung Cancer Dataset	RCP (Royal Colleges of Physicians)
Colorectal Cancer Dataset	ACP (Association of Clinical Pathologists)
NCHSPC Dataset	NCHSPC (National Council for Hospice & Specialist Palliative Care Services)
National Cancer Registry Dataset	National Cancer Registry Co-ordinator
Radiotherapy Dataset	National Cancer Services Analysis Team

Annex 4 Business Related Microdata

Annex 4.1 WERS 2004 Broad Topics

Management Survey	Union Survey	Employees Survey
Workforce composition	Structure of representation	Working hours
Personnel management / employment relations	Time spent on representative duties	Job influence
Recruitment and training	Means of communication with employees	Job satisfaction
Workplace flexibility and the organisation of work	Negotiation/ consultation	Working arrangements
Consultation and information	Redundancies, discipline and grievance matters	Training and skills
Employee representation	Collective disputes and industrial action	Information and consultation
Payment systems and pay determination	Relations with managers Union recruitment	Employee representation
Grievance, disciplinary and dispute procedures		Pay
Equal opportunities, work-life balance		
Workplace performance		

Annex 5 Demographic Datasets

Annex 5.1 Variables contained in the ONS LS Study[117]

CENSUS INFORMATION

- Age
- Sex
- Marital status
- Residence indicators (various)
- Country of birth
- Parent's country of birth (1971 only)
- Ethnicity (1991 and 2001 only)
- Religion (2001 only)
- Marriage and fertility history (1971)
- Family, household establishment type
- Housing and tenure, rooms and amenities
- Educational qualifications
- Economic activity
- Occupation
- Industry of work
- Social class
- Travel to work and migration
- Self-rated health
- Long-standing illness (1991 and 2001 only)
- Care-giving (2001 only)

LIFE EVENTS INFORMATION

- Fertility: live and still to women
- Infant deaths (deaths of infants born to sample mothers)
- Cancer registrations
- Mortality (deaths of sample members), including cause of death
- Widowhoods (deaths of sample members' spouses)

[117] It is noted that this is a shortened summary of main variables in the LS. The database structure is very complex. There are currently 4,500 variables held in the LS. Most of these related to or derived from the variables listed, with variable names, etc varying according to time period of census.

Annex 6 Project Participants

The following people helped to guide the project, whether directly by means of taking part in project related meetings, or by supplying relevant information and advice.

Name	Organisation
Susan Alpay	NHS Information Centre
Anna Barker	Department for Education and Skills
Iain Bell	Department of Works and Pensions
Louisa Blackwell	Office for National Statistics
Paul Boyle	Scottish Longitudinal Study
Michael Bright	Economic and Social Research Council
Simon Burgess	University of Bristol
Jon Burton	Cabinet Office
Rhys Davies	Office for National Statistics
John Elliot	Department for Education and Skills
Mark Foster	Department of Works and Pensions
Peter Goldblatt	Office for National Statistics
Miles Gould	NHS Information Centre
Lin Hattersley	Scottish Longitudinal Study
Simon Heathfield	NHS Connecting for Health
Andrew Jenkins	Institute of Education
Peter Knight	NHS Information Centre
Denise Lievesley	NHS Information Centre
Neil McIvor	Department of Works and Pensions
Jonathan Mosedale	Department for Transport
Jeremy Neathey	Economic and Social Research Council
Minda Phillips	Office for National Statistics
Felix Ritchie	Office for National Statistics
Andrew Roberts	Department for Education and Skills
Tim Shiles	Department for Education and Skills
Linda Shurlock	NHS Information Centre
George Smith	University of Oxford
Allan Sudlow	Medical Research Council
Ricky Taylor	HM Treasury
Andy Teague	Office for National Statistics
Anna Vignoles	Institute of Education