UK Commission for Employment and Skills

LMI for All

Developing a Careers LMI Database:

Phase 2A Report

Career Database Project Team

Warwick Institute for Employment Research

Jenny Bimrose, Rob Wilson, Sally-Anne Barnes, David Owen, Anne Green, Yuxin Li, Peter Millar, Luke Bosworth, Andy Holden

Pontydysgu

Graham Attwell, Philipp Rustemeier

Raycom

Raymond Elferink

Rewired State

Adam McGreggor, Debbie Wicks, Angi Long

External Consultant

Andy Dickerson

9 July 2013

Contents

Executive Summary 1

1. Introduction 6

1.1. Understanding the context 6
1.2. Project overview.. 8
1.3. Project aims and objectives.. 10

2. Phase 2A data developments 11

2.1. Approach to providing data 11
2.2. Description of data in the LMI for All database 12
2.3. Enhancing the database: additional data sources 16
2.4. Summary of data sources and indicators 23
2.5. Potential future enhancements... 24

3. Accessibility and open data: Phase 2A technical developments 26

3.1. Activity in Phase 2A 26
3.2. Platform and database 27
3.3. Extract, Transform and Loads (ETLs)... 29
3.4. Data security and data disclosure 30
3.5. Wiki for tracking project development.. 30
3.6. Improved API reliability... 31
3.7. Improved API documentation and Client Code Generation 32
3.8. LMI for All web portal 33
3.9. Improved search 34
3.10. Terms and Conditions of Service 34
3.11. FAQ 34
3.12. Implications for Phase 2B 35

4. Stakeholder engagement and communication... 40

4.1. Testing the database API.. 40
4.2. The developers 40
4.3. The careers stakeholders.. 41
4.4. The hack day 41
4.5. The modding day 44
4.6. Next steps for stakeholder engagement and communication 48

5. Summary of Phase 2A progress and recommendations 49

5.1. Issues and resolutions 49
5.2. Implications for costing and timescales 49
5.3. Recommendations 50

Annex A: Main data sources included in Phase 2A 53

Annex B: O*NET 94

Annex C: Review of potential of other data to be considered for Phase 2B .. 103

Annex D: Use cases for hack day developers.. 130

Annex E: Summary of careers stakeholder pre-event questionnaire............. 132

References 135

List of tables and figures

Figure 1 Representation of LMI for All database, web portal and API 9

Figure 2 Overview of data and variables in the LMI for All database 12

Figure 3 Current LMI for All platform and database 27

Figure 4 LMI for All database dashboard 31

Figure 5 Representation of the LMI for All platform and database recommended for Phase 2B 38

Table 1 Descriptions of apps and interfaces, plus stakeholder feedback......... 42

Figure 6 JobBungee/On Demand screenshots 45

Figure 7 JobQuest screenshot 46

Figure 8 LMI everywhere widget screenshot... 46

Table A.1 Typical Earning Function Results 57

Table A.2 Broad Sectors (SIC2007) 88

Table A.3 Industry Groups (SIC2007).. 89

Table A.4: SOC2010 Major Groups and Sub-major Groups 90

Table B.1 Mapping from SOC 4-digit categories directly to O*NET 96

Table B.2 Data layout of ‘Skills.txt’ and ‘Abilities.txt’ 97

Table B.3 Abilties.txt 98

Table B.4 Skills.txt... 99

Table B.5 Alternative steps to improving the matching 100

Table B.6 The O*NET content model.. 102

Figure C.1 Labour market questions in 2011 Census of Population 118

Figure C.2 Journey-to-work questions in 2011 Census of Population 119

Table C.1 Mapping from ISCO08 to SOC2010 123

Table C.2 Map from ISCO 88 to SOC2010 at 2-digit level 124

Glossary

API

API, an abbreviation of application program interface, is a set of routines, protocols, and tools for building software applications. A good API makes it easier to develop a program by providing all the building blocks. A programmer then puts the blocks together.

App

An App or application is a computer software application that is coded in a browser-supported programming language (such as JavaScript, combined with a browser-rendered mark-up language like HTML) and reliant on a common web browser to render the application executable. Apps are accessed by users over a network.

ASHE

The Annual Survey of Hours and Earnings, from the Office for National Statistics, provides information about the levels, distribution and make-up of earnings and hours worked for employees in all industries and occupations.

BRES

Business Register and Employment Survey collects data to update local unit information and business structures on the Inter-Departmental Business Register (IDBR) and produce annual employment statistics, which are published via the NOMIS website. It replaces the Business Register Survey and the Annual Business Inquiry.

CEN

Chancellor of the Exchequer's Notice is required to access potentially disclosive data.

Data cube

A data cube is commonly used to describe a time series of image data representing data along some measure of interest. It can be 2-dimensional, 3-dimensional or higher-dimensional. Each dimension represents some attribute in the database and the cells in the data cube represent the measure of interest. Queries are performed on the cube to retrieve decision support information.

DLHE

Destinations of Leavers from Higher Education is a survey of qualifiers from higher education (HE) institutions, which is conducted in two parts. The first stage asks what leavers were doing six months after they qualified from their HE course. The second stage or longitudinal survey is a follow-up survey that looks at the destinations of leavers three and a half years after they qualified. Managed by the Higher Education Statistics Agency (HESA).

ESS

The Employer Skills Survey conducted by UKCES provides information on business management, recruitment, skills gaps and vacancies. The surveys are designed to be representative of the employer population across geography and sector.

ETLs
Extract, Transform and Load processes are for database usage, including: extracting data from external sources; transforming it to fit operational needs, which can include quality levels; plus loading it into the end database.

Hack day
Hack days (also known as Hackathons or Appathons) bring together experts and developers to collaborate or work alone rapidly prototyping software or hardware, building mobile and web apps or quick models for new ideas and features.

ILO
The International Labour Organization is devoted to promoting social justice and internationally recognised human and labour rights. It helps advance the creation of decent work and the economic and working conditions that give working people and business people a stake in lasting peace, prosperity and progress. Its main aims are to promote rights at work, encourage decent employment opportunities, enhance social protection and strengthen dialogue on work-related issues.

JACS
JACS (Joint Academic Coding of Subjects) is the subject classification system used to describe the subject content of courses at UK Higher Education institutions. JACS3 is used from 2012/13. This was developed jointly by HESA (Higher Education Statistics Agency) and UCAS.

JCP
Jobcentre Plus, part of the Department for Work and Pensions (DWP). It provides services that support people of working age from welfare into work, and helps employers to fill their vacancies. Main supplier of vacancy data.

JSON
JavaScript Object Notation is a lightweight data-interchange format. It is a text format that is language independent using familiar conventions that can be found in the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python and others.

LFS
The Labour Force Survey, conducted by ONS, is a quarterly sample survey of households living at private addresses in the UK. Its purpose is to provide information on the UK labour market.

LMI
Labour market information is data, graphs and statistics that describe the condition of the past and current labour market,

as well as make future projections.

Modding day
The modding day follows a hack day. Its aim is to take forward the developments of the hack day and to produce a more useable and defined product.

MySQL
MySQL is a type of database management system that enables data to be added, accessed and processed in a database. It is open source. MySQL is supported by Microsoft and Oracle.

NOMIS
Web-based database of labour market statistics from ONS, includes statistical information on the UK labour market (i.e. Employment, Unemployment, Earnings, Labour Force Survey and Jobcentre Plus vacancies).

NUTS1
Nomenclature of Units for Territorial Statistics. This is a geocode standard for referencing the subdivisions of countries for statistical purposes. The standard is developed and regulated by the European Union. There are three levels of NUTS defined. In the UK, NUTS1 represents the regions of England, plus Wales, Scotland and Northern Ireland.

O*NET
The Occupational Information Network is a US program providing a primary source of occupational information. Central to the project is the O*NET database, containing information on standardised and occupation-specific descriptors. Information from this database forms the heart of O*NET OnLine http://www.onetonline.org/, an interactive application for exploring and searching occupations.

ONS
The Office for National Statistics is an Executive Office of the UK Statistics Authority. It is responsible for the collection, compilation, analysis and dissemination of a range of economic, social and demographic statistics relating to the UK.

RAS
RAS is an iterative procedure where the rows and columns of preliminary estimates of a two dimensional array are iteratively changed using proportions that are based on ‘target’ row and column totals (see Section A.8).

Relational database
A relational database is the predominant choice in storing data that conforms to relational model theory.

Scala and Scalatra
Scalatra (using Scala) is a web micro-framework that helps the developer quickly build high-performance websites and APIs.

SDS

The Secure Data Service provides safe and secure remote access by researchers to data previously deemed too sensitive, detailed, confidential or potentially disclosive to be made available under standard licensing and dissemination arrangements.

SIC

The Standard Industrial Classification is used to classify business establishments and other statistical units by the type of economic activity in which they are engaged. The latest version in SIC2007.

SOC

The Standard Occupational Classification is a common classification of occupational information for the UK. Jobs are classified in terms of their skill level and skill content. The latest version is SOC2010. SOC 4 digit provides a list of occupations at a more detailed level.

SPARQL

A recursive acronym for SPARQL Protocol and RDF Query Language. This is an RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format. SPARQL is a format favoured by linked data proponents as it allows advanced queries and the ability to query between different datasets.

SQL server

This is a relational database server, developed by Microsoft. It is a software product designed to store and retrieve data as requested by other software applications.

Standard server, web container of servlet container

This is the component of a web server that interacts, is responsible for managing servlets, mapping a URL to a particular servlet and ensuring that the URL requester has the correct access rights.

TTWA or Travel-To-Work-Area

TTWA indicates an area where the population would commute to another area for the purposes of employment.

Ubuntu Linux LTS

This is a popular open source operating system for servers and cloud computing.

UKDA

The UK Data Archive is curator of the largest collection of digital data in the social sciences and humanities in the UK.

Universal Jobmatch service

Universal Jobmatch is the Department for Work and Pensions (DWP) online service, which is open to all jobseekers, regardless of whether or not they are claiming a benefit. It works by matching jobseekers to jobs based on their skills and CV.

Visual Basic (VB) Visual Basic is a third-generation programming language from Microsoft. It enables rapid application development of graphical user interface applications and access to databases.

Working Futures Detailed historical and projected employment estimates produced on behalf of UKCES (for details see: http://www.ukces.org.uk/ourwork/working-futures)

XCRI XCRI stands for eXchanging Course Related Information. It is the UK standard for describing course information.

Executive Summary

1. LMI for All is a web portal, which provides access to a comprehensive and rich set of labour market information (LMI) that IT developers and others can exploit to produce a range of applications to help inform career choices and decisions. It is being funded and managed by the UK Commission for Employment and Skills (UKCES) and supports the programme of reforms that have been designed by Government to build a stronger and more balanced economy by improving the quality and range of labour market information available.

2. The prime objective of LMI for All is to create a comprehensive repository of LMI to support careers decision-making. It builds on a successful feasibility study carried out early in 2012, which tested the feasibility of using existing sources of data to develop a prototype LMI database for careers (also managed by UKCES). Although the initial focus is on careers guidance and advice, it has the longer-term potential to inform a much wider audience concerned with labour market behaviour and economic performance. By linking and opening up careers focussed LMI to support people make better decisions about learning and work, it provides a resource to improve the effectiveness and efficiency of careers support offered to users.

3. Working with a range of partners led by the Warwick Institute for Employment Research, the second phase of the project, Phase 2A, (October 2012 – May 2013) has culminated in the launch of the first release of LMI for All on the 31 May 2013. The **main site** can be found at http://www.lmiforall.org.uk/. This contains information about LMI for All and how it can be used. The **API web explorer for developers** can be accessed at http://api.lmiforall.org.uk/. The API is currently open for testing and the development of applications. **Technical details about the data tool** can be found at http://collab.lmiforall.org.uk/. Here, information will be found about the data included in the current release and how this can be used. There is also a Frequently Asked Questions section.

4. The LMI for All web portal supports the Open Data policy agenda, by opening up existing data sources. However, it is not simply about 'open data'. Rather, it is about making useful LMI available at a much more detailed level than currently exists, taking account of concerns about disclosure, privacy and confidentiality, and statistical reliability.

5. The current version of the LMI for All database contains key data from the following data sets, which for the first time, are available from a single access point:
    - Employment and Replacement Demands from *Working Futures;*
    - Pay based on the Annual Survey of Hours and Earnings and the Labour Force Survey*;*
    - Total hours based on the Annual Survey of Hours and Earnings*;*
    - Unemployment rates based on the Labour Force Survey*;*
    - Skills shortage vacancies based on the Employer Skills Survey; and
    - Current job vacancies based on the Universal Jobmatch database, using a schematic to illustrate the data that are available through the portal.

Data overview – LMI for All Phase 2A

	Earnings	**Employment (Historical)**	**Employment (Projected)**	**Employment (Replacement Demand)**	**Hours**	**Vacancies**	**Unemployment Rates**
Sources	Annual Survey of Hours and Earnings (ASHE)/Labour Force Survey (LFS)	*Working Futures* - Business Register and Employment Survey (BRES)/Labour Force Survey (LFS)	*Working Futures* - Business Register and Employment Survey (BRES)/Labour Force Survey (LFS)	*Working Futures* - Business Register and Employment Survey (BRES)/Labour Force Survey (LFS)	Annual Survey of Hours and Earnings (ASHE)	Employer Skills Survey (ESS)	Labour Force Survey (LFS)
Indicator	Average full-time earnings	Number of jobs (employee self-employed)	Number of jobs (employee self-employed)	Number of job openings between selected years (employee self-employed)	Average weekly hours	Number of Vacancies, Hard-to-fill vacancies, Skill shortage vacancies, Occupation	ILO Unemployment rate
Dimensions*	Occupation, Industry, Qualification, Geography, Gender, Status	Occupation, Industry, Qualification, Geography, Gender, Status	Occupation, Industry, Qualification, Geography, Gender, Status	Occupation, Industry, Qualification, Geography, Gender, Status	Occupation, Industry, Geography, Gender, Status	Occupation, Industry, Geography	Occupation, Industry, Qualification, Geography, Gender, Status
Period	2012	2000-2010	2010-2020	2010-2020	2012	2011	2011/2012

Notes: * Occupation (SOC2010 4-digit), Industry (SIC2007, 75 industries), Qualification (NQF 0-8), Geography (UK countries and English regions), Gender, Status (full-time or part-time employee and self-employed).

6. The LMI currently available includes detailed information about the labour market, which covers the following dimensions:
    - Detailed 4-digit occupational categories;
    - Industry;
    - Countries and English regions within the UK;
    - Level of highest qualification held;
    - Gender;
    - Employment status (full-time and part-time employees, self-employment); and
    - Universal Jobmatch vacancy data is included in the API through a fuzzy searching process.
7. A key challenge for Phase 2A of LMI for All has been around the connected matters of disclosure, confidentiality and statistical reliability. A practical solution has been developed that satisfies the demand for detail and open access, while at the same time recognising the very real concerns about privacy, disclosure and the need for statistical robustness.
8. In addition to the focus on data, the use of technology for LMI for All has been an integral and crucial part of the success of the second phase. Key achievements have included the development of a secure and robust infrastructure for the data, which provides a secure environment that can be extended as new data becomes available and the number of end-users increases and adheres to open standards; modern and flexible software tools to allow the querying of the database by external users; and software tools and spaces for documenting the process and allowing public access.
9. The third strand, essential to the overall success of the project, involves stakeholder engagement and communication. The approach during this second phase has involved the organisation of a hack day, followed by a modding day. For the hack day, ten developers were recruited by a process of open competition. Their task was to design and develop prototype applications that used the database, testing the API. In addition, ten career stakeholders, from different sizes and types of career organisations, attended the event to judge the applications produced. Four applications were selected, which went forward for further development to the modding day. The cycle of design, development and judging by career stakeholders was repeated. The aim of the hack and modding days was to demonstrate the potential of the database through the production of prototype applications.

Recommendations

- ❖ This exercise has demonstrated the practical feasibility of developing a data portal to serve the needs of the careers guidance and advice community. LMI for All should be further developed to meet the LMI needs of these groups (as well as other potential users in the longer term).
- ❖ The main indicators in the LMI for All database in Phase 2A (October 2012 – May 2013) should continue to be used in the next phase of the project (Phase 2B, June 2013 – March 2015), including:
    - Employment and employment forecasts based on *Working Futures* (these include

information on qualifications and replacement demands);

  - Unemployment rates (using the International Labour Organization definition of unemployment[1]) based on the LFS;
  - Pay (estimates based on a combination of ASHE and LFS data);
  - Hours worked (ASHE);
  - Vacancy estimates (based on ESS and Universal Jobmatch);
  - Vacancies (based on a fuzzy search from Universal Jobmatch);
  - Occupational descriptions (ONS).

- ❖ Various refinements to the way these estimates are generated are proposed, some of which can be implemented in Phase 2B (e.g. focusing on medians/ deciles, rather than means). Others involve work outside the project (e.g. refining the projections of employment at the 4-digit occupational level, which will require an extension to the current *Working Futures* database).
- ❖ Further consideration of use of “raw” survey data as opposed to estimates/predictions.
- ❖ The full, revised O*Net dataset, including Skills and Abilities, as well as a number of other skill related indicators, should be implemented in Phase 2B of the project.
- ❖ Other possible indicators and enhancements to the LMI for All should be considered, including:
  - Further work to integrate UJM vacancy data into the database more fully, once mapping to occupational categories has been resolved.
  - Making greater use of data from higher education, such as HESA information on the destination of graduates (however, this will require detailed negotiation with data owners).
  - Course information - a great deal of information is available about courses of study and links to different career paths, but this is not well coordinated or consistent. More work needs to be done to bring this into the database.
  - The UK Census of Population, especially local labour market information (since there is limited other sub-regional information), including some commuting and workplace data);
  - NOMIS, using the API to include workforce jobs data at regional level, the unemployment claimant count and data from the APS;
  - Cedefop pan-European employment database – equivalent to UK *Working Futures*, (but only available at 2-digit occupational level), move to the revised ISCO08 data as soon as they are available (early 2014) and exploit more detailed information (if and when it is published);

[1] The ILO definition of unemployment covers people who are: out of work; want a job have actively sought work in the previous four weeks and are available to start work within the next fortnight; or out of work and have accepted a job that they are waiting to start in the next fortnight.

- The following should not be included in the database in Phase 2B: ONS Vacancy Survey (no occupational detail); Annual Population Survey (does not add much to LFS); Jobcentre Plus vacancies (historical data only – series discontinued); and EULFS (problems with availability and detail).

- Early discussions need to take place in Phase 2B regarding technical priorities and server capacity. The development and maintenance of a vibrant web portal with support services for users and developers will promote uptake. Consideration needs to be given to the amount of resources this will require, not only in technical terms, but in design, moderation and intervention to respond to and support developers and users. Such resources have to be balanced with priorities for further data and technical development.

- Continuous encouragement and support should be given to organisations with an interest in using the early release of the web portal and API, which is part of the approach to testing, evaluating and improving the pilot tool, as well as demonstrating the benefits to a wider audience. The nature and level of this support should be discussed.

- A more strategic use of social media and dissemination at key events should be adopted throughout Phase 2B, to ensure the web portal and API are promoted to create a market for the product and to maintain the momentum of interest. This will complement the planned dissemination events (three workshops and a conference), to be delivered throughout the Phase 2B.

- The successful format of the hack and modding days should be adopted in Phase 2B, since these were successful in not only proving the viability of the database, but also enabled career stakeholders to contribute to the development process.

- Active participation of key stakeholder representatives throughout the project should be carefully designed to ensure stakeholder engagement. This will be achieved through key stakeholder participation in future hack and modding days, alongside a conference and seminar events for the different stakeholders groups (namely careers representatives, policy makers and developers). Throughout Phase 2B, there will be an on-going dialogue with organisations that have expressed an interest in using the API and their feedback gathered in order to inform further refinements and amendments to the database and API.

- Communication of the web portal concept should go beyond traditional dissemination methods (e.g. newsletters, professional publications, presentations at various events, etc.). Visual representations of potential applications should be made available to various audiences, in response to advice on priority target groups and their career needs collected from key stakeholders (e.g. the National Careers Service; Careers Wales; the National Apprenticeship Service; TAEN, etc.).

1. Introduction

The genesis of LMI for All came from Sir John Holman, particularly through work for the Gatsby Charitable Foundation. In February 2010, the Gatsby Foundation funded a short review of STEM (Science, Technology, Engineering and Mathematics) careers, led by Sir John Holman, who was then National STEM Director (Holman & Finegold, 2010). This review complemented the publication of a report by the Science and Society Expert Group (Science for Careers Expert Group Report, 2010). Both the report and review shared a concern about the quality of support available for those wishing to develop a career in science. The focus was on secondary and tertiary education levels.

Following the STEM Review, a seminar was held by the National STEM Centre on 28 September 2011. The seminar highlighted the critical importance of robust and accessible LMI for young people, parents, careers professionals and teachers.

To explore this further UKCES commissioned a partnership, led by the Warwick Institute for Employment Research, to develop a prototype database drawing on existing data sources, with the specific purpose of testing the feasibility in practice of creating a comprehensive repository of careers labour market information (LMI) over the longer term.

A key objective from the research and development work was to advise on the feasibility of developing a database of robust LMI, which can be opened up for multiple interfaces for a range of users. The prototype was developed during the period January – March 2012 and the conclusion from this feasibility phase of the project was that further development of the prototype was viable, both from the technical and data perspective.

Subsequently, UKCES commissioned a second phase of work (October 2012 – May 2013), with the same partnership, to create a comprehensive repository of LMI, to support career decisions. Although the initial focus was to be on careers guidance and advice, it would have the longer-term potential to inform a much wider audience concerned with labour market behaviour and economic performance. This second phase of the project was to establish LMI for All as a web portal, to provide access to a rich set of labour market information (LMI) that IT developers and others can exploit to produce a range of applications to help inform career choices and decisions. This phase was completed successfully, with the launch as scheduled, on 31 May 2013. The main site can be found at http://www.lmiforall.org.uk/. This contains information about LMI for All and how it can be used. The API web explorer for developers can be accessed at http://api.lmiforall.org.uk/. The API is currently open for testing and the development of applications. Technical details about the data tool can be found at http://collab.lmiforall.org.uk/. Information can be found here about the data included in the current release and how this can be used.

This report relates to the second phase of the development and research project.

1.1. Understanding the context

As part of the Government's Plan for Growth, a commitment has been made to 'create an improved careers information portal as part of the National Careers Service' (HM Treasury and DBIS, 2012, p. 7). This is seen by Government both as part of a wide-ranging programme of economic reforms and investment in infrastructure to help build a stronger and more balanced economy in the medium term (HM Treasury and DBIS, 2012, p. 1). Opening up careers focussed labour market information (LMI) to optimise access to, and use of, core national data sources that can then be used to support people make better decisions about learning and work not only contributes to the creation of an improved careers information portal, but also supports the Open Data policy agenda. The recent White Paper on Open Data states the clear intention to use 'the data we hold more effectively, and by pushing that data into the public domain' (HM Government, 2012, p.12). Determination to shift the culture of the public sector to improve data sharing is also indicated – where this is in the public interest and within legislative boundaries – by using the latest technology (HM Government, 2012, p. 6).

A key policy objective for providing high quality LMI is to improve the efficiency of labour markets by facilitating adjustments and helping match supply with demand. LMI can play a key role in reallocating resources from alleviating unemployment in declining industrial sectors of the economy and addressing labour and skill shortages in other sectors or regions. By facilitating the reallocation of labour from declining to expanding areas, LMI can also foster local development. It can also improve the integration of new entrants to the labour market, including immigrants and young people making school-to-work transitions. LMI additionally plays a key role in helping people make better-informed decisions about investments in human capital. Not only is it a crucial component of a skills or human capital development strategy that is increasingly regarded as necessary to foster productivity and competitiveness, it can also inform decisions with respect to life-long learning, including apprenticeship decisions and subsequent adult training.

Millions of people across the UK make important decisions about their participation in the labour market every year. This extends from pupils in schools, to students in Further and Higher education institutions and individuals at every stage and phase of their career and learning journeys. Whether these individuals are in transition from education and/or training, in employment and wishing to up-skill, re-skill or change their career direction, or whether they are outside the labour market wishing to re-enter currently or at some future date, high quality and impartial LMI is crucial to effective and efficient career decision-making. UKCES has already made a significant, and recent, investment in exploring the characteristics of effective career support, the role of new technologies in enhancing this support and the part that Government and national agencies might play in securing greater quality and impact from the public investment in career services (UKCES, 2011). In particular, commissioned research has examined the inter-relationships amongst LMI, information communications and technologies (ICT) and information, advice and guidance (IAG) (UKCES, 2010a) and considered what drives the provision of LMI in an online career context (UKCES, 2010b).

It would, however, be difficult to argue that there is insufficient LMI – a myriad of information currently exists and developments in ICT have ensured that much is easily accessible. However, the existence and dissemination of LMI is a necessary, but not sufficient, condition to meet the objective of enhancing individuals' decision-making processes. LMI, as any other type of information, must be meaningful and resonate with individuals' particular situations to be transformed into knowledge. Personal values, beliefs, perceptions and emotions play a critical role in the ways people use and process LMI. Indeed, much existing robust LMI can be regarded as 'coded knowledge', with these data needing to be interpreted to become useful information that will then be transformed into relevant knowledge. Here, the role of mediation as part of the career guidance process in converting information into relevant knowledge is critical.

1.2. Project overview

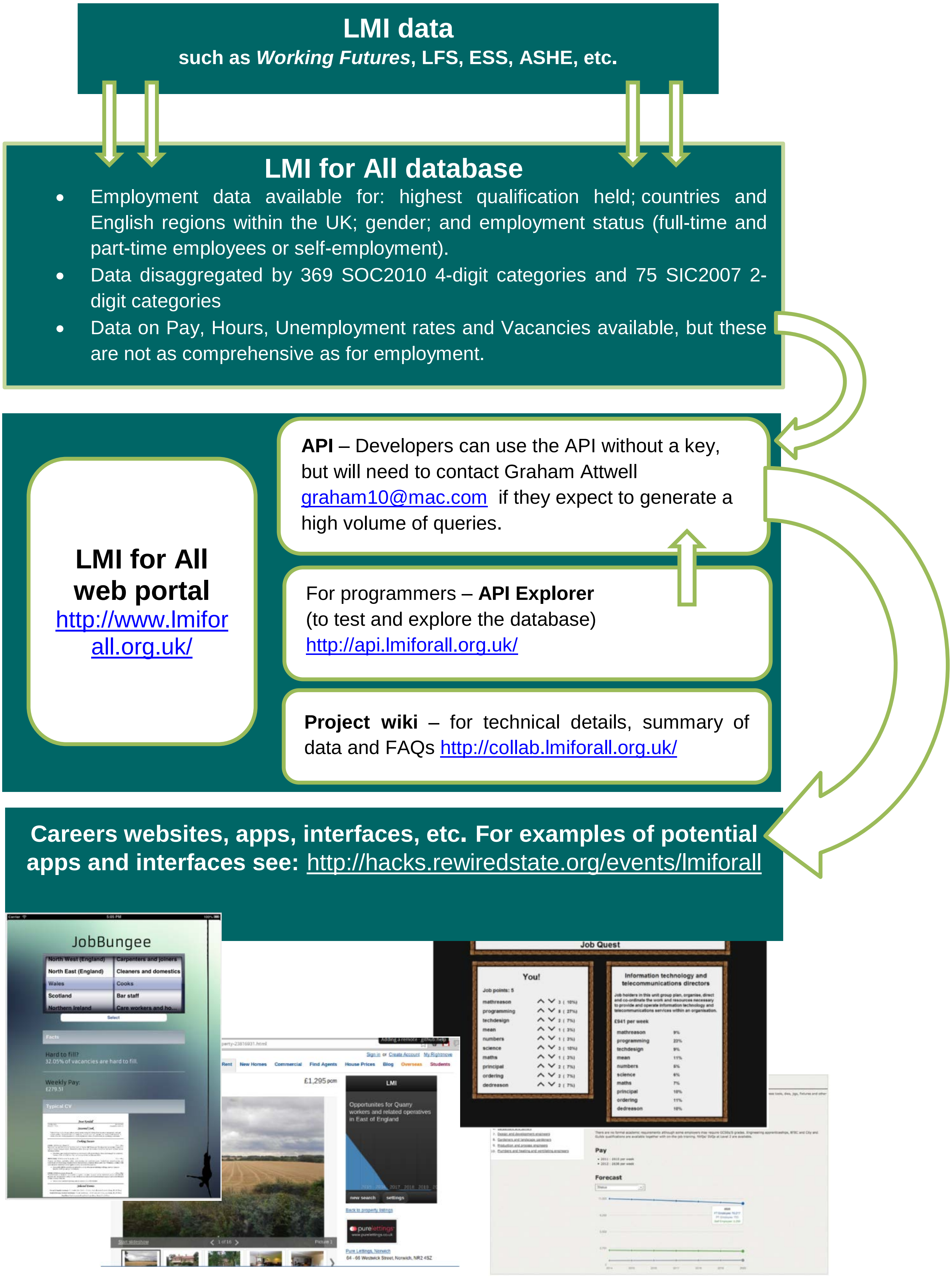
The key purposes of Phase 2A of this project have been twofold: first, to identify and investigate which robust sources of LMI can be used to inform the decisions people make about their learning and work; second, to bring these sources together in an automated, single, accessible location so that they can be used by developers to create websites and applications for career guidance purposes. Integral to the success of the project was testing the concept of creating an accessible point to access career LMI, as well as evaluating the appropriateness of the data tool created directly with key stakeholders. It builds on findings from the prototype project (Phase 1), carried out early in 2012 by the UKCES (Bimrose et al., 2012). For this, the UKCES commissioned the Warwick Institute for Employment Research (IER), working in partnership with Pontydysgu, Raycom and Rewired State, to pilot the feasibility of developing a careers LMI database that would be accessible through an API (Application Programming Interface)[2]. The API succeeded in bringing existing selected data sources together, such as from some of our key national surveys, to open up careers focused LMI. Because the pilot demonstrated the feasibility of the original concept of an LMI database for careers, UKCES commissioned the second phase of this research and development project (Phase 2A).

LMI required for the database in the second phase of the project (Phase 2A) includes data on employment level, employment forecasts, vacancies, earnings and qualification levels from diverse data sources, all focussed on occupations. Data sources include: the Labour Force Survey; *Working Futures* employment database; Annual Survey of Hours and Earnings; UKCES Employer Skills Survey; and the O*NET skills database (limited in Phase 2A). Also included in the database are the ONS occupational descriptions. The project is being developed in-line with open data principles, with outputs from the project allowing community, voluntary, public and private sector organisations to access these data free of charge, so that they can use the data to develop tools and interfaces that meet the needs of different customer groups in ways that will augment existing careers guidance and information being used to support the career decision process of individuals in transition. Figure 1, below, provides a representation of how the LMI for All database is accessed.

Relevant labour market data have been organised by occupational category using the 2010 Standard Occupational Classification (SOC) at unit group (4-digit) level as a framework. An index of c.28,000 job titles mapped to SOC provides the basis for the end-user to search, and gain access to, data of interest and relevance in an intuitive fashion.

[2] API, an abbreviation of application program interface, is a set of routines, protocols, and tools for building software applications. A good API makes it easier to develop a program by providing all the building blocks. A programmer then puts the blocks together.

Figure 1 Representation of LMI for All database, web portal and API



LMI data

such as *Working Futures*, LFS, ESS, ASHE, etc.

LMI for All database

- Employment data available for: highest qualification held; countries and English regions within the UK; gender; and employment status (full-time and part-time employees or self-employment).
- Data disaggregated by 369 SOC2010 4-digit categories and 75 SIC2007 2-digit categories
- Data on Pay, Hours, Unemployment rates and Vacancies available, but these are not as comprehensive as for employment.

LMI for All web portal http://www.lmiforall.org.uk/

API – Developers can use the API without a key, but will need to contact Graham Attwell graham10@mac.com if they expect to generate a high volume of queries.

For programmers – **API Explorer** (to test and explore the database) http://api.lmiforall.org.uk/

Project wiki – for technical details, summary of data and FAQs http://collab.lmiforall.org.uk/

Careers websites, apps, interfaces, etc. For examples of potential apps and interfaces see: http://hacks.rewiredstate.org/events/lmiforall

1.3. Project aims and objectives

1.3.1. Aims

The overall aims of this project were twofold:

- To identify and investigate which robust sources of LMI can be used to inform the decisions people make about learning and work; and
- To bring these sources together in an automated, single, accessible location (referred to as the LMI for All database), so that they can be used by developers to create websites and applications for career guidance purposes.

1.3.2. Objectives

These are represented in three separate, but inter-related work strands, as follows:

Data development:

- To identify the key information that is used in making decisions about learning and work.
- To explore the feasibility of including UK wide data where this is available.
- To prepare the data and bring these together with other data sources as part of a single access point.

Accessibility and open data:

- To produce an initial version of the data tool (this refers to the LMI for All database, platform, web portal and API), based on lessons learned from the feasibility project (conducted January – April 2012).
- To develop subsequent iterations of the data tool, in-line with stakeholder feedback, to be gathered as part of the project process.

Stakeholders and communication:

- To test the data tool, through two separate iterations (for the first and second phases of the project) of hack and modding days.
- To consult with stakeholders in the broad community of career guidance practice, through three workshops.
- To disseminate findings to a wider audience, through a conference.

The three work strands are divided across two phases, Phase 2A (November 2012 to end May 2013) and Phase 2B (June 2013 to March 2015). This report focuses on the progress of Phase 2A activities, highlighting issues and solutions, together with recommendations for next steps.

2. Phase 2A data developments

This section outlines data developments in Phase 2A, reviews additional data sources for Phase 2B and identifies implications for the future of the project.

2.1. Approach to providing data

The initial approach to developing the LMI for All database focussed on using the APIs from official sources in order to facilitate quick and automatic updates. However, it soon became apparent that there were a number of problems and pitfalls with this approach. The main difficulties arise because many of the official data sources that it was intended to tap into were not designed with the purpose of providing very detailed labour market information for careers guidance purposes in mind. A key issue is around the connected matters of:

- Disclosure;
- Confidentiality; and
- Statistical reliability.

Many of the official statistics are collected under the terms of strict legal instruments, which ensure confidentiality for those providing the data and guarantee that these data will not be published in such a manner as to disclose commercially sensitive or other confidential information about the companies or individuals concerned. The Office for National Statistics (ONS), which is responsible for collecting and publishing the information, has strict rules in place to ensure that this is the case. This poses quite severe limits on the level of detail that can be placed into the public domain. It should also be noted that key data owners (such as ONS) do not currently have APIs in place that allow easy access to data on indicators such as employment and pay.

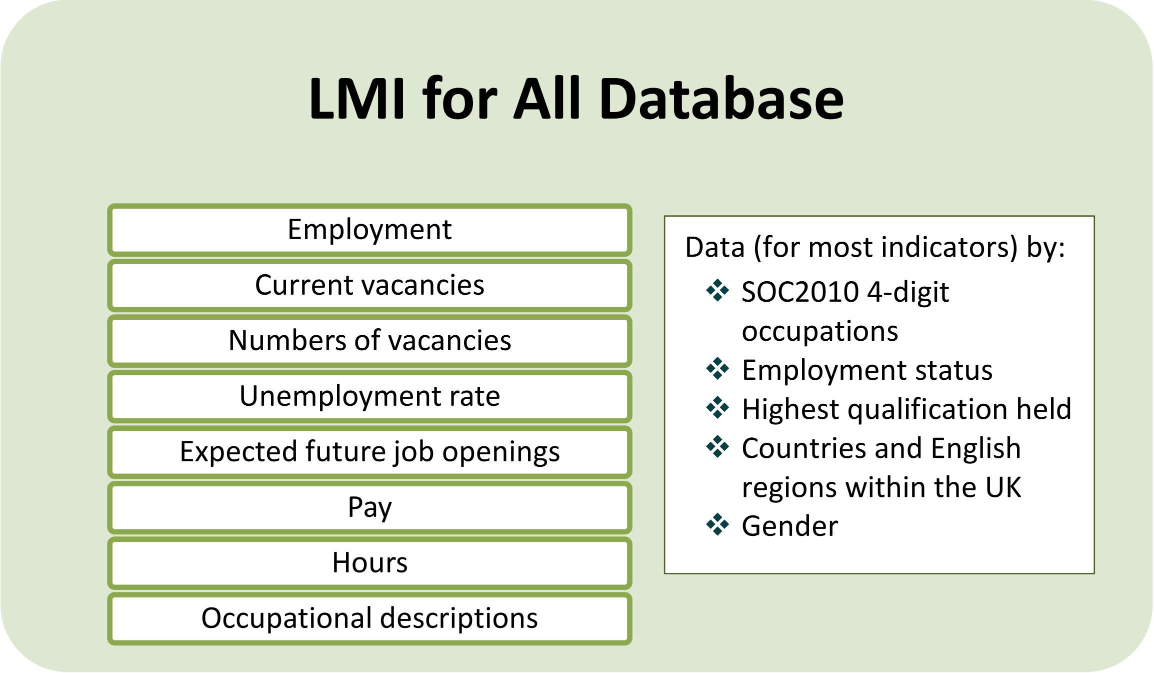
The other important consideration is statistical reliability. This is essentially a matter of the sample size on which the statistics are based. Many of the official sources are based on samples, which while large in statistical terms, are not large enough to provide robust information at a very detailed level. This applies to both the Business Register and Employment Survey (BRES), which is the main source of information on employment by industry, and the Labour Force Survey (LFS), which is the main source of information on the structure of employment by occupation, qualification and employment status. Reliance on the raw survey data would, therefore, severely limit the level of detail that could be provided.

This issue has been addressed previously in the context of developing the *Working Futures* (WF) employment database (See Wilson and Homenidou, 2012a, 2012b). The solution adopted there has been to combine the various official sources and to create estimates of employment at a more detailed level than it is possible to obtain from the official surveys alone. This has been combined with putting in place checks to ensure that the data generated are robust (in a general statistical sense) and that they do not breach confidentiality nor are disclosive. ONS previously required all those with access to the data to sign up to a Chancellor of the Exchequer's Notice (CEN). Detailed discussions with ONS were focused on two main issues. First, that the aggregation of information on employment by industry to some 75 industries could avoid problems of disclosure without the necessity for a CEN;[3] and second that as long as sources such as the LFS and the Annual Survey of Hours and Earnings (ASHE) were used to produce estimates for general groups rather than revealing information on individual cases, then this should not breach confidentiality.

Further details of how the official sources have been used to generate detailed estimates of Employment, Pay and Hours are set out in Annex A. In addition, for pay, supplementary information is provided showing variation by age, based on a parametric approach.

2.2. Description of data in the LMI for All database

Figure 2 Overview of data and variables in the LMI for All database



LMI for All Database

- Employment
- Current vacancies
- Numbers of vacancies
- Unemployment rate
- Expected future job openings
- Pay
- Hours
- Occupational descriptions

Data (for most indicators) by:

- SOC2010 4-digit occupations
- Employment status
- Highest qualification held
- Countries and English regions within the UK
- Gender

2.2.1. Indicators

Employment (WF historical estimates based on LFS, BRES, etc.)

The LMI for All database requires detailed data if it is to be useful for careers guidance. Individuals and their advisers have an interest in knowing which jobs are available, distinguishing sector, occupation and typical qualifications required, (as well the typical pay and hours) associated with those jobs. Ideally, the full set of detail required is as follows:

- Occupation (up to the 4-digit level of SOC2010, 369 Categories);[4]
- Sector (up to the 2-digit level of SIC2007, about 80 categories);
- Geographical area (12 English regions and constituent countries of the UK);[5]

[3] For details, see Annex A.

[4] Some have argued for an even more detailed breakdown to the 5-digit level of SOC, but this is not feasible given data currently available.

- Gender and employment status (full-time, part-time employees and self-employed).

For reasons discussed above, the use of the raw data from BRES and the LFS does not provide a suitable source of the kind of detailed data needed to populate the database.

It is important to emphasise that individual observations from these official surveys on Employment (or Pay or Hours worked) are not required. What is needed is general information on ‘typical’ pay or general employment opportunities in particular areas for people with selected characteristics. The official data are a means to this end rather than being required for their own sake.

The level of detail required in the LMI for All database can be obtained by replacing the official ‘raw’ data by *estimates* or *predictions*. For *employment*, the *Working Futures* employment database has been used. The *Working Futures* database includes historical information on employment by both Occupations and Qualifications. The latter shows the numbers employed by highest level of qualification held using the National Qualification Framework (NQF) system of classifying levels of qualification. The measure of employment used is workforce jobs rather than a head count of people in employment.

The standard *Working Futures* employment database only provides information up to the 2-digit level of the Standard Occupational Classification (SOC2010). This has been extended for the LMI for All database to the 4-digit level by combining the database with additional information on the patterns of employment at this more detailed level using LFS data. These estimates are constrained to match the main *Working Futures* database using an extended version of the algorithm developed to produce the main *Working Futures* database (For details see Wilson and Homenidou, 2012b).

Although estimates can be generated for the full level of detail shown at the start of this section, not all of these are reliable and robust. In order to rule out such information, the API censors results that fall below a certain threshold and flags up cases where the estimates may be less reliable. These criteria are based on rules of thumb developed for the main *Working Futures* database. The rules of thumb used are:

1. If the numbers employed in a particular category/cell (defined by the countries/ regions, gender, status, occupation, qualification and industry) are below 1,000, then a query returns ‘no reliable data available’ and offers to go up a level of aggregation across one or more of the main dimensions (e.g. UK rather than region, aggregation of industries rather than the most detailed level, or SOC 2-digit rather than 4-digit).
2. If the numbers employed in a particular category/cell (defined as in (1)) are between 1,000 and 10,000 then a query returns the number but with a flag to say that this estimate is based on a relatively small sample size and if the user requires more robust estimates they should go up a level of aggregation across one or more of the main dimensions (as in 1).

This also applies to estimates of replacement demands as well as employment levels. Full

[5] Plus for some purposes additional information on: age; gender; status; and qualification (highest held).

details are given Annex A.

Pay (estimates based on a combination of ASHE and LFS)

In the feasibility study (Bimrose et al., 2012) information on pay was extracted from the LFS. UKCES were keen to make use of data on pay from ASHE as this is thought to be more reliable (because information is provided by employers, rather than being the subject of individuals' recall) and based on a larger sample. However, despite this, it is still not able to deliver robust information at a very detailed level (i.e. for individuals classified by a combination of detailed industry, occupation and region). This is partly because of concerns about disclosure, but also because the limited sample size means that estimates have a high degree of uncertainty. This issue is exacerbated if information on variations in pay by age is also required. A further problem is that ASHE does not have any information on pay by qualification.

In order to get around these problems, the LMI for All database is based on a set of estimates/predictions of pay rather than the raw survey estimates. Analysis of pay using earnings equations is a well-established way of understanding the key factors that influence pay. This is done using a combination of both ASHE and LFS data. In order to ensure that the predicted pay figures match up with the published official data, an algorithm to constrain the data to match agreed 'targets' has been developed. This is analogous to the procedure used to generate the detailed *Working Futures* employment data, described in the previous section.

Queries to the LMI for All database about Employment and Pay (and Hours) also check the implied sample sizes to see if the estimates are likely to be unreliable. In the case of Pay (and Hours) the API interrogates the part of the LMI for All database holding the employment numbers to do the checks, as in (1) and (2) above, but then reports the corresponding Pay or Hours values as appropriate.

Again, full details are given in Annex A.

Hours worked (ASHE)

As in the case of Pay, relevant information is available from the LFS or ASHE, but in both cases very detailed data cannot be extracted because of concerns about disclosure, confidentiality or statistical reliability. The ASHE data are regarded as the more reliable (for the same reasons as Pay) and are therefore used here.

This problem has been addressed in a similar way to Pay, by producing predictions for Hours in place of the raw survey data. In principle, a regression equation could be used to produce these estimates although there is no direct equivalent to the well-established 'earnings equation'.

In practice, because the ASHE data reclassified using SOC2010 are not yet available via the Secure Data Service (SDS), a non-parametric method has been used based on the published data.

As for Pay, the API checks for reliability and where necessary, suppresses unreliable data. Again full details are given in Annex A.

Unemployment (LFS)

The unemployment rate is an extremely important indicator from a careers guidance perspective, representing the probability of a worker of a given type or living in a particular location being unemployed. The unemployment rate can be calculated by age, gender and occupation for statistical regions from the LFS. The ONS publishes time series of unemployment rates.

While the LFS microdata can be used to calculate unemployment rates for SOC 4-digit occupations, the sample sizes involved can be very small (resulting in problems of breaching confidentiality and statistical reliability of estimates). Estimates of the unemployment rate have therefore been generated, using the End User Licence version of the LFS microdata. In principle, these allow detail up to the same level as shown for employment at the start of this section, but in practice, there are many gaps in the data and the results for many categories are based on sample sizes too small for the results to be reliable. The same rules of thumb are used to suppress unreliable estimates as for Employment and Pay. The Census of Population provides an alternative source for the unemployment rate which has much greater geographical detail, but this is only available for March 2011 (the Census date).

Vacancies (UKCES ESS)

The number of vacancies is another key indicator from a careers guidance perspective, as a measure of the number of jobs potentially available to job-seekers. Historically, DWP have generated a set of information on vacancies notified to Jobcentre Plus by occupation that would ideally form part of the database (this source is discussed in the next section below). This series has recently been discontinued and is being replaced by a new series generated by DWP/Monster (this new series is also discussed below), but no occupational data coded to the SOC is currently available from this source.

Thus, at present there is only one source that can be used in the LMI for All database. This is the Employer Skills Survey (ESS), carried out once every two of years since 2001, and now managed by UKCES. The normal published results only provide detail to a very broad occupational level (1 or 2-digit). The survey company IFF have provided a more detailed version of the dataset with information at a 4-digit level, which has been used for this project.

Because it is based on a sample of around 1 in 20 employers, data from the ESS is subject to statistical uncertainty, which increases as the number of observations on which an estimate of vacancy numbers is based. The API therefore only returns vacancy estimates based on 50 or more observations. This means that data is not available for many smaller occupations (the effect of which is greatest for 4 digit occupations).

Another limitation of this source from a careers guidance and advice perspective is that it does not provide a picture of jobs currently available – but a measure of the number of vacancies employers had when the survey was conducted. The latest data relate to 2011. Data from the 2013 survey will become available towards the end of 2013. Nor is it comprehensive, focusing on up to six occupations in the sampled firms. However, until the new series produced by DWP/Monster can be linked in to the database it provides the best indication of job availability. The ESS data complements the official count of vacancies by providing an indication of the matching of supply and demand in particular occupations (showing occupations in which vacancies are hard to fill and subject to skill shortages).

Occupational descriptions (ONS)

ONS have collated information on detailed job descriptions for SOC2010 4-digit categories. This is very useful from a careers guidance perspective, because the description details methods of entry into an occupation including the qualifications required and a list of task involved in the job. It is, therefore, included in the LMI for All database. Detailed information is provided for each SOC2010 4-digit category.

2.3. Enhancing the database: additional data sources

There are many other data sources that could be exploited to enhance and extend the LMI for All database. These are considered in this section. The discussion is deliberately succinct, with more detailed information provided in Annex C.

As with a number of the sources discussed in the previous section there are many technical problems linked to the fact that these sources were not designed with the particular purpose of providing data suitable for careers guidance and advice.

2.3.1. The UK Census of Population

The decennial Census of Population provides a very rich source of labour market information. This is collected with various uses in mind, including general social science research. It is of considerable interest to labour market analysts. Annex C provides a comprehensive description of the various data available from the Census, including the timetable for delivery of results announced by ONS.

Many of these data are probably of more value to general labour market analysts than those concerned specifically with careers guidance and advice. Annex C sets out a long list of potentially interesting indicators including:

- Labour market and employment data (employment, unemployment, economic activity);
- Commuting and workplace data (distance travelled and mode of transport).

The key advantage of the Census is the provision of data for small geographical areas and the information it provides on the distance workers have to travel to different types of job.

Its main disadvantage from a careers guidance and advice perspective is that it is not very timely (most results being published more than two years after the Census is taken) and it refers to just a single point of time (March 2011). For further Details see Annex C.

2.3.2. ONS Vacancy Survey

The ONS Vacancy Survey provides a count of the total number of vacancies in the UK economy. It provides information by sector but not by occupation. It could be used to provide some indication of the general state of the job market. However, given that the main focus of the LMI for All database is on careers guidance and advice it is recommended NOT to include this source as a priority in Phase 2B. However, it should be noted that the marginal cost of adding it would probably be trivial. For further details see Annex C.

2.3.3. Annual Population Survey

The Annual Population Survey (APS) is a boosted version of the LFS, providing sufficient sample numbers in each local authority district for statistically reliable labour market measures to be derived. The APS dataset provides access to the same range of variables available from the LFS, but at a more detailed geographical scale, and thus (in principle) is a better choice for the LMI for All database than the LFS.

The cross-tabulation of variables from APS microdata can yield a large amount of information on the employment characteristics (either of residents or workplaces) of a sub-regional geographical area, but there are restrictions placed on its use by ONS because of concerns about confidentiality. Restrictions on access become greater as the level of detail increases and limit the ability of analysts to distribute data from the APS to third parties. APS data could only be incorporated within the LMI for All database via a route not subject to such restrictions (e.g. the generation of an extract by government statisticians or using the APS data from NOMIS which is already in the public domain).

The current use of the LFS in the LMI for All database has been narrowed down to providing unemployment rates (see Section 2 of the main report). The extra value of the APS in this regard is that it can provide this data for smaller geographical areas and a more detailed occupational breakdown.

The implications of ONS licence restrictions upon the data which can be generated from the APS needs to be further investigated in order to decide whether additional APS data can be included and distributed publicly by including it in the database. This will require further negotiations with the UK Data Archive (UKDA) and ONS. The marginal benefits are therefore probably modest and the marginal costs quite high.

For further details see Annex C.

2.3.4. NOMIS

The National Online Information System (NOMIS) is a repository for a range of data sources. It holds the full range of labour market related ONS and DWP statistical outputs available at the sub-regional scale. It provides extremely easy access to a time series of data going back to 1982 for the majority of datasets and to 1971 for a few others (e.g. employment and June unemployment[6]). Most NOMIS datasets cover either Great Britain or the whole of the UK. They include data from BRES, APS, Census of Population and the LFS.

The NOMIS datasets are accessible via a ‘Restful’ API interface[7].

Employment data (based on the Business Register and Employment Survey (BRES)),

[6] NOMIS holds unemployment data monthly from July 1978 onwards, but extended this series back for each year from 1971 to 1978 for June only in order to provide a time series which links to the employment time series (also referring to June each year at that time), from which an annual unemployment rate (U/(U+E))*100 can be calculated. June was chosen because this is the month in which seasonal effects are least.

[7] This is a web API implemented using HTTP and REST principles, which is often JSON. The API is hypertext driven.

include:

- Location of jobs by industry;
- Industrial profile of employment in an area;
- Location of part-time jobs.

However, access to the BRES based employment data is problematical, because the data are collected under the Statistics of Trade Act 1947 which promises to maintain the confidentiality of data provided by survey respondents. Hence all users have to apply for a CEN from the Department for Business Innovation and Skills in order to use the data.

Data from the Annual Population Survey including:

- Occupational profile of employment;
- Qualification profile of employment;
- Labour market participation by age group, gender, ethnicity and nationality.

The advantages of using APS data from NOMIS is that because NOMIS uses a pre-specified set of standard cross-tabulations, there are no problems of access, it is available for different levels of aggregation and there are statistical flags attached which identify whether data is reliable.

Unemployment data are based on the claimant count and are available for a long time series and small geographical areas. Although these data are coded by occupation, they use the SOC2000 classification and are therefore of limited value for the LMI for All database.

Jobcentre Plus (JCP) data on notified and unfilled vacancies and the duration of vacancies are also available, classified to occupations using the SOC2000 classification. This is now a historical series, because data collection ended in October 2012 when Monster.co.uk took over from Jobcentre Plus.

Census of Population: NOMIS also provides very easy access to data from the 2011 Census of Population via a simple query system and bulk downloads. Census data is valuable mainly for providing contextual information about local labour markets, the characteristics of jobs located in an area and information on the geographical matching of labour supply and labour demand, through information on commuting patterns.

NOMIS provides access to a rich variety of data on employment and the labour market and is a source, which is invaluable for any general labour market analysis application. However, the LMI for All database has a narrower focus on the availability of opportunities for current job seekers and most of the official statistics it encompasses do not directly address this need.

2.3.5. *Working Futures* projections at 4-digit level

The current published *Working Futures* database provides projections by occupation at a 2-digit level of SOC2010 (although 3-digit level results have also been produced historically).

In principle, more detailed projections are technically feasible but this is limited by the quality of the available data upon which the analysis is based (primarily the LFS).

In the pilot database, the possibility of using common growth factors applied to all 4-digit unit groups within a 2-digit category (i.e. assuming fixed shares) was explored but not fully implemented. This has been taken to an operational level in Phase 2A.

The LFS enables reasonably robust estimates of the shares of employment in SOC 2-digit categories that are employed in the 4-digit unit groups they contain at the all industry level. With a small amount of additional work, trends within the 4-digit occupation could be considered and used to develop more realistic projections. This could also make use of data from the Census of Population (which was not available when the *Working Futures* projections were undertaken). There is also some scope for considering variations by industry (but sample sizes in the LFS preclude doing this at much more detailed a level than the six broad sectors used in *Working Futures*).

As long as these results are clearly presented as projections based on simple assumptions rather than precise predictions, then it is feasible to generate such numbers. This is the spirit in which even more detailed occupational projections are made in the US by the Bureau of Labor Statistics (See Wilson (2010) for more detailed discussion).

2.3.6. Cedefop database

For the past 5 years, IER, in collaboration with others, have developed an historical employment database and projections at a pan European level on behalf of Cedefop. This replicates many of the same features of the *Working Futures* employment database.

In principle, the data can be used to generate employment information, including replacement demands, for each of the 27 EU Member States plus a few additional countries such as Norway and Switzerland.

In practice, there are a few issues:

- The data are currently classified using ISCO 88 which is not directly comparable with SOC2010 – however, a broad brush mapping can be derived (see below).
- The new data to be published in 2013/2014 will use ISCO 08. This is broadly compatible with SOC2010. IER and ONS have been working on developing mappings.
- The current Cedefop projections are primarily focused on the 2-digit level. Development of information at a more detailed level is being explored, but data limitations are problematic. Information at a 4-digit level is unlikely to be available in the foreseeable future.

On balance, it would be useful to add such information to the database in order to provide a broad perspective on job opportunities across Europe but it would not be a top priority from a careers guidance perspective, given the lack of occupational detail and the difficulties in making a simple mapping of occupational categories.

2.3.7. Other European sources

A range of other European sources have also been considered, including the European LFS as well as other regular European surveys (such as the Eurobarometer surveys, the European Values Survey, European Social Survey and the European Working Conditions Survey). These can also provide useful contextual information on issues such as attitudes towards labour migrants in different countries, working conditions, etc. These are briefly summarised and discussed in Annex C.

In practice, although they all contain some interesting and useful data they are generally not suitable for inclusion in the LMI for All database because the sample sizes are inadequate to provide reliable data at a detailed and consistent level by occupation. The information they provide is also generally not particularly relevant for careers guidance and advice. They would have more value if the database were to be extended to cover the needs of other users such as more general labour market analysts.

2.3.8. Skills and abilities (O*NET Skills data)

The feasibility study (Bimrose et al., 2012) suggested that the US O*NET database could be exploited in the UK to provide useful information about the skills involved in carrying out different jobs. The US database has been developed over many years and contains a very rich set of information classified using the US equivalent to SOC2010. The feasibility study used some mappings developed in an earlier study to link SOC2010 occupational categories to the US ones. It showed that this could then be used to exploit information on STEM skills developed in the US based around two particular areas entitled ‘Abilities’ and ‘Basic Skills’ in the O*NET database.

The present project has reassessed the mappings and also explored the other areas covered by the O*NET system. This includes a much richer set of skills and related attributes. These would add considerable value from a careers guidance perspective and should therefore be included at an early stage in Phase 2B.

2.3.9. Vacancies, Universal Jobmatch (DWP/Monster)

The data collected by Monster on behalf of DWP replace the former series of vacancy by occupational information, which was based on vacancies notified to Jobcentre Plus (a subset of unknown size of all vacancies in the economy). In practice, the data currently available via the DWP/Monster website use a system of classification based on job titles that does not match any UK occupational standard. Unless some mapping can be made between the categories used by DWP/Monster and SOC2010 4-digit categories used in the LMI for All database, this information is of limited value. Using the current DWP/Monster categories it can be included in LMI for All but not fully integrated, which requires reclassification using the standard SOC2010 categories.

The authors understand that Monster are contractually obliged to provide data using standard classifications in the future, but until such data are available some kind of “fuzzy matching” based on reported job titles could be explored in Phase 2B to provide a feed of vacancy information from the DWP/Monster website. This would include details of actual vacancies rather than an attempt to quantify them.

2.3.10. Course information

Data on courses and training available across the UK are not held in any one central database, so discussions have been progressing throughout Phase 2A with various government departments and other relevant organisations to negotiate access to the information repositories which are accessed through the various search tools. It is evident that compiling a comprehensive list of further and higher education training and courses will be complex mainly due to the number and range available, as well as the variable quality of the data available. Accessing these data will be complex due to the way it is recorded and coded, with different coding systems that have been developed and evolved over time (i.e. JACS[8], XCRI[9]). In order to include course data in the LMI for All database, there would need to be comprehensive mapping of courses to occupational codes.

Although a central database of course data is not available, various stakeholders compile and use information from providers for various purposes. For university courses, these can be found on the UCAS (University and Colleges Admissions Service) website. This covers the whole of the UK. College-based provision is found on careers websites. Each of the four constituent countries of the UK has a careers website and these sites have been investigated for high quality course data.

In England, the Skills Funding Agency (SFA) maintains a Course Directory Provider Portal, which comprises learning and course provision data. The provider portal enables learning providers to view and update their course directory information. For learners, the Course Directory can be accessed on the National Careers Service website at https://nationalcareersservice.direct.gov.uk/advice/courses/Pages/default.aspx. The SFA disclosed that there have been problems with the quality of course data collected in the past, but this has greatly improved. An API is available, but this needs redeveloping. Discussions have also progressed with the Student Information Services Limited, a charity that runs the 'best course for me' website (http://www.bestcourse4me.com). This website provides information on university courses and possible career paths. Mapping of course codes to SOC have been undertaken and a range of APIs are available. During the discussions, the complex nature of coding and mapping was highlighted.

Skills Development Scotland obtain course data from a specialist service, called Gateway Shared Services (http://www.ceg.org.uk). This organisation collects and collates information about learning opportunities and careers throughout Scotland to produce a range of online services (such as MappIT, MerIT, PlanIT Plus, WorkIT) and reference books. It covers both further and higher education data, which are updated on annual basis.

In Wales, Careers Wales collects and updates course information and vacancy data for Wales. Access to these data has been negotiated and will be provided through an API in Phase 2B as data are maintained.

In Northern Ireland, there is no central database of course information. The Northern Ireland Course Directory (also known as NI Learning Opportunities Database) was developed and maintained by DCA Data Solutions. This was available on the Careers Service website and

[8] JACS (Joint Academic Coding of Subjects) is the subject classification system used to describe the subject content of courses at UK Higher Education institutions. JACS3 is used from 2012/13.

[9] eXchanging Course Related Information, or XCRI, is the UK standard for describing course information developed for further education.

fed in to the National Learning Directory managed by Learndirect. Access to the directory was removed from NI Careers website in early 2011. Currently, there are no plans to update and maintain a NI Course Directory. NI Careers recently confirmed that their advisers and clients currently access information about learning provision through http://www.indirect.gov.uk/careers, which links to further and higher education course providers. There are plans to procure software that would include access to a course directory with information from UCAS.

Further discussions are planned for Phase 2B with other organisations that collect course data (e.g. HESA). However, discussions so far have confirmed that compiling a comprehensive list of course information and data will be complex, time-consuming and likely to be resource intensive. There will need to be careful mapping and users will need to understand that such data will not be automatically updated on an annual basis. Manual input will be required.

2.3.11. Other information from Higher Education Agencies

There is also a rich source of relevant information on the passage of individuals through Higher Education (HE), including the first destinations of graduates. For example the data HESA collect in their graduate destination survey contains SOC classification and potentially allows mapping from courses studied to job destination in a straightforward way. Currently much of this kind of information is only made available subject to a charge. Such issues should be explored further in Phase 2B. Of course this is will require detailed consultation and negotiation with the data owners concerned.

2.4. Summary of data sources and indicators

Data source	Indicators	Variables	Available in the LMI for All database	Next steps
Working Futures (combination of LFS and BRES)	Total number of jobs by detailed type (historical estimates)	Where possible all data available at SOC2010 4-digit occupations. Also covers: Industry; region; gender; employment status; and highest qualification held.	Yes	
Working Futures (combination of LFS and BRES)	Expected future jobs and replacement needs (total job openings over a period)		Yes	Improve projections at 4-digit SOC level
UKCES ESS	Current vacancies			Currently not covered very comprehensively by the ESS data.
DWP Monster – Universal jobmatch	Types of vacancies		Yes, but currently available through search (using fuzzy matching); 2010 based data not available	Continue to review
ASHE/LFS	Typical pay (mean weekly pay)		Yes, but currently measured by the mean weekly earnings for full time employees only; LFS data used where ASHE not available.	Revision using ASHE once SOC2010 data available; Extend to include medians, deciles, part-time pay etc.
ASHE	Typical hours (mean weekly hours)		See note above	Explore use of regression analysis once SOC2010 data available.
LFS	Unemployment rates		Yes	
ONS Standard Occupational Classifications 2010 – Structure and descriptions	Occupational descriptions		Yes	ONS description of what is involved in undertaking a job in a 4 digit occupation
O*NET	Skills and abilities		No	Matching to UK SOC2010
Variety of sources	Course data		No	To be explored further; Problems with mapping to SOC, coverage and quality of data
HESA	Destination data		No	To be reviewed after discussions with data owners.
APS	Unemployment		No	To be reviewed, not considered high priority; Limited added value
NOMIS			No	Continue to review
Cedefop			No	Keep under review, but probably postpone inclusion until ISCO08 data are available (2014)

2.5. Potential future enhancements

- Phase 2A of this project has demonstrated that adequate data are available to populate the LMI for All database. However, this will require regular processing to keep the database up to date. Steps will need to be taken to maintain this process. This will involve developing a smooth workflow around processing the various key datasets (making the sources and procedures as efficient and transparent as possible so that updating the database is automated as much as it can be).

- Many sources considered are based on samples too small to provide useful information at the level of detail desired. Increases in sample sizes could help to make the data more useful. However, this will imply very significant costs and such developments are unlikely to happen quickly. In the meantime it is important to make the most of what is currently available.

- In the longer-term, it would be better if the predicted estimates used for the three key indicators in the database, employment, pay and hours, could be replaced by “raw” or “real” survey data, which could be updated automatically as they are published. This raises two questions:

    - If and when it will ever be possible to replace at least some of the predicted / estimated values used for some indicators by “real” survey values; and

    - Checks on the reliability robustness of some of the more detailed predictions/estimates.

- In principle, it is possible to use “real’ survey values where these are statistically robust and non-disclosive and to only use predicted values to fill in the many gaps. In practice, this would pose some problems of consistency. This is something that should be explored in more detail in Phase 2B. This will require further detailed consultation with ONS and the development of an agreed methodology for merging “real” and predicted values in a seamless fashion.

- Some data are not classified in a manner suitable for inclusion in the database, (the use of SOC2010 for classifying occupations is especially important). Steps need to be taken to ensure better harmonisation. This has two aspects:

    - It is partly about lobbying data providers to move to a common standard as soon as it is practicable (recognising that this has cost implications and may take time);

    - But it is also about carrying out more work in Phase 2B to further harmonise the data, and in particular to develop new mappings to non-standard classifications where possible.

- Work with data owners to encourage them to improve access to their data via APIs, with the ultimate aim of increasing automation and providing a more dynamic resource for data users, increasing commitment to open data principles, while recognising the practical barriers.

- Finally a number of other sources might add information that could be of value to a broader audience than those concerned with careers guidance and advice. Once the database is established thought should be given as to how it might be developed and enhanced to meet the needs of groups such as those concerned with local economic development and other users.

3. Accessibility and open data: Phase 2A technical developments

This section outlines the technical developments that have been undertaken during Phase 2A focussing on accessibility and open data issues. It highlights technical issues encountered and solutions found. It also discusses implications for Phase 2B of the project.

3.1. Activity in Phase 2A

The purposes of the technical developments in Phase 2A of the project were threefold:

- First, the development of a secure and robust infrastructure for the data providing a secure environment that could be extended in the future as new data becomes available and as the number of end-users increases.
- Second, the provision of modern and flexible software tools to allow the querying of the database by external users.
- Third, the provision of software tools and spaces for documenting the process and allowing public access to the LMI for All database.

Main outcome

The main outcome of the technical development under Phase 2A of the project is the LMI for All database (accessed through a web portal and API) that has been released on the open market. It builds on the data tool that was developed during the feasibility study in early 2012. The LMI for All database:

- Provides access to key data that individuals need to make decisions about learning, skills and careers;
- Provides data that have been quality-assured in terms of robustness, statistical quality and confidentiality;
- Includes data that have been cleaned and standardised;
- Has been tested with developers and further iterations made as required.

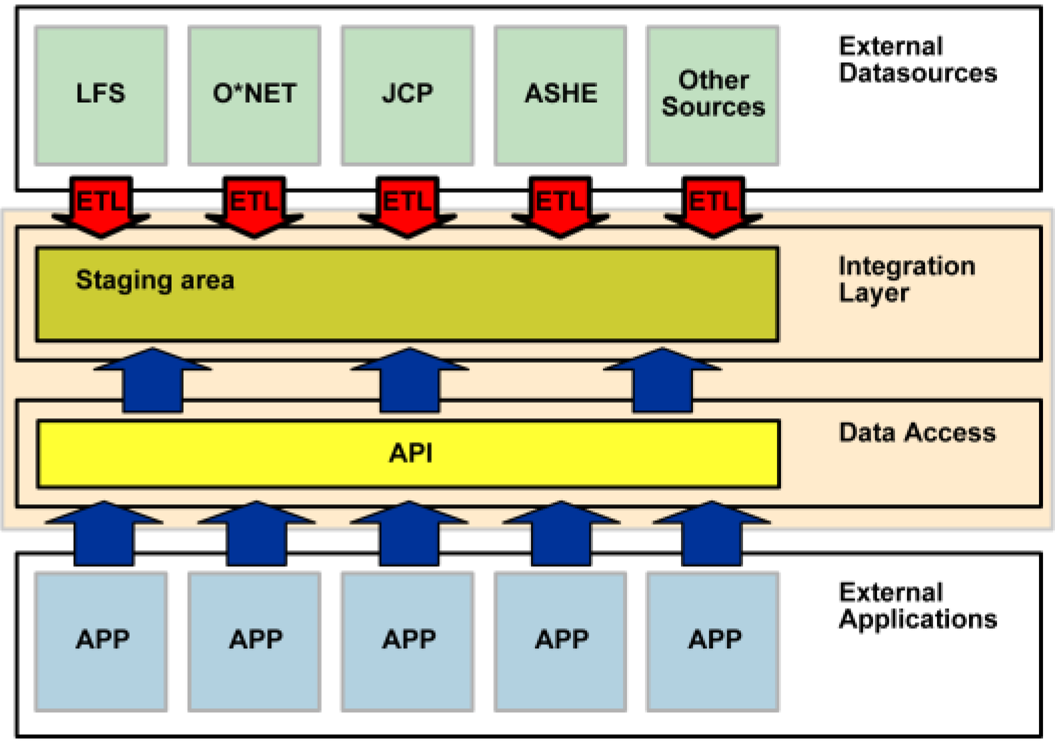
Access to the LMI for All web portal is at http://collab.lmiforall.org.uk/doku.php?id=start.

Although the project has faced a number of difficulties in making official data available to the wider public it has developed solutions and work arounds that are in the spirit of open access principles, use of open standards, etc, while at the same time recognising the very real concerns about issues such as disclosure, confidentiality, privacy and statistical robustness.

3.2. Platform and database

The original feasibility project was developed on a MySQL database[10]. The current LMI for All platform and database is represented in figure 3, below. Although a powerful environment, it was anticipated that a stronger industrial database would be required for implementing Phases 2A and 2B of the project; especially if development of more automated processes for the updating of data were to be developed. The main choices were between Microsoft SQL server and the Oracle SQL database. The decision to install a Microsoft SQL server was taken on the grounds of flexibility and price.

Figure 3 Current LMI for All platform and database



To host the database environment, a dedicated Windows 2012 Server with MS SQL Server 2012 Standard Edition was installed. Consideration was given to where and how to host the server. It was decided to host it at Dediserve Ltd. This is a trusted cloud-based server that can be extended when needed, so also offering flexibility. Cloud computing is the use of computing resources (hardware and software) that are available in a remote location and accessible over a network. Cloud computing allows organisations to get their applications up and running faster, with improved manageability and less maintenance. It enables IT to adjust resources more rapidly to meet fluctuating and unpredictable business demands. This is particularly important, as it is difficult to predict the physical size of the database, given that different data sources are a major topic being researched throughout the project. The environment allows the technical team dynamically to add resources and create new, or

[10] MySQL is a type of database management system that enables data to be added, accessed and processed in a database. It is open source. MySQL is supported by Microsoft and Oracle.

expand existing, servers. Additionally, it is very hard to estimate future demands on the server from external users, so again a Cloud solution offers more flexibility as it is relatively easy to add capacity to the server in a short time period. Selecting a provider who offered a secure solution with managed services for server maintenance and security was also a concern. Dediserve Ltd. met all of the requirements.

Currently, three separate servers are being run:

1. data.lmiforall.org.uk – This server holds the database environment. It runs Windows 2012 Server Edition and MS SQL Server 2012 Standard Edition.
2. api.lmiforall.org.uk – This server holds the API layer. It runs Ubuntu Linux LTS[11].
3. collab.lmiforall.org.uk – This server holds the collaboration environment and the public website. It runs Ubuntu Linux LTS.

The separation of the different servers allows dedicated environments and systems to be run for different platforms and usages. Importantly, it also provides greater security, as the database server can only be accessed through the API.

The LMI for All database has been designed so that no confidential/disclosive data are held anywhere in the LMI for All environment. No data on individuals or individual organisations are included.

3.2.1. Platform and database issues and solutions

Three main issues emerged during the development of the platform and LMI for All database:

- Managing the size of data;
- Ensuring the speed of querying the database; and
- Developing methods in which to query the database.

First, the size of the data in the original specification was underestimated. This was overcome through adding extra storage capacity to the server, but resulted in higher than estimated server costs. Second, there was the issue of speed in querying the database with some initial queries being slow. This was a consequence of the amount of data and the number of different sources/indicators included. This was overcome by improving the automatic indexing of data to achieve significantly faster searching. However, this also required more storage capacity. The final issue was around the different methods in which the database would need to be queried. This was resolved through discussions with the project data development team. It resulted in trade-offs between two different ways of developing and querying the database for the wage data. Wage rates are based on estimates to preserve anonymity but involve an iterative process. This cannot be undertaken dynamically or 'on the fly'.

The first solution included preparing dedicated datasets for the particular aspects of querying

[11] This is a popular open source operating system for servers and cloud computing.

that would be facilitated through the web portal and Application Programming Interface (API). This method would provide a simpler database structure, but would require significantly larger datasets. The second solution was to run the queries algorithmically and 'on the fly', i.e. in real time and dynamically. This, through performing live filtering and aggregation of data, would reduce the database size and potentially result in quicker returns. Although the second option was attractive, there were some concerns that it could introduce considerable issues in the stability of the API, especially with a high rate of simultaneous hits. However, it proved possible to reduce the number of files and thus the size and overhead on the server and to calculate a number of variables 'on the fly' in the API. These include estimates of pay by age and estimates of replacement demand.

3.3. Extract, Transform and Loads (ETLs)

Despite the Government's commitment to open data, there remain major challenges in developing applications based on these data. Although openly published, the formats of data are often not suited to being automatically read by machines. Furthermore, these data can require significant cleaning (such as removing invalid data or errors) or manipulation (such as resorting or mapping data to be easily queried) prior to being integrated into a database. Also, with data that is frequently updated, for instance LMI, the formats of the data frequently change between different releases (such as different classifications, new or deleted variables). This can require significant work in preparing data and also result in delays in the frequency of updates. The technical solution to this is to develop Extract, Transform and Load (ETL) processes for database usage, including: extracting data from external sources; transforming it to fit operational needs, which can include quality levels; plus loading it into the end database.

MS SQL Server provides a professional database environment that allows for the implementation of formal automated ETL processes. ETL processes for all current data sources have been implemented in order to minimise manual tasks in the import and integration processes.

The use of namespaces is important for the technical environment and have been introduced into the LMI for All database. Namespaces are separate sections within the database. Two separate namespaces in the MS SQL database environment are being used: 'raw'; and 'production'. The 'raw' namespace holds tables that store the raw imported data from the various data sources. The 'production' namespace holds the integrated production tables.

There are scripts in place that automate ETL processes for each data source that:

- Import the raw data from the import stream to the 'raw' namespace;
- Transform and clean the data;
- Integrate the data in the 'production' namespace.

Each table in 'production' has been properly indexed to improve performance and supply the API layer with quick results.

3.3.1. ETLs issues and solutions

Although the developed ETL processes are successful, these only minimise manual tasks in the import and integration process. Whilst data continues to be issued in different formats and due to the requirements over non-disclosure, there remains some considerable human overhead in the preparation of data. However, the present ETLs have greatly improved the manual processes used in the pilot. Additionally, there remains scope to further automate data management processes.

3.4. Data security and data disclosure

The database server cannot be directly accessed through the internet. Data can only be queried through the API layer, which itself sets strict rules as to what data can be accessed. This alleviates concerns about disclosure. The API layer is only able to access the production space and has no access to the raw namespace. None of the data in the database are disclosive or confidential and information that might be statistically unreliable is either supressed or flagged up as such.

3.4.1. Data security and data disclosure issues and solutions

In the early stages of the project, there were extensive discussions between the data development and technical teams to find robust solutions regarding data security. The issue was overcome through the adoption of two namespaces and strict API querying restrictions, as detailed above.

3.5. Wiki for tracking project development

A wiki has been developed which has two purposes. First, a private space is used for project communication and the management of the data between the data development and technical teams. Different datasets are made available at different points throughout the project. These datasets may be new or may be updates of previous releases. They will have various dimensions, classifications, values and coding. The datasets themselves are often too large for manual inspection. Thus, it is critical that each dataset is comprehensively described and documented for quality control purposes by the data development team. The wiki facilitates a system of process control, with documentation being signed off by the technical team, before data are downloaded and installed in the database.

Second, there is a public facing area in the wiki. This provides technical documentation and detailed documentation of the different data sets for LMI managers and application developers and was used in the hack and modding days.

Overall, the wiki provides a flexible and collaborative working environment, with a managed permission environment (i.e. access to particular pages or to the wiki as a whole can be restricted), which can be rapidly updated.

3.5.1. Project development issues and solutions

Prior to the development of the wiki, based on the open source docuwiki platform, a number of different collaborative tools and environments were trialled. It was important to find a tool that allowed sufficient functionality for collaborative working, but was also easy to use for non-technical partners from the project team. This was achieved with the wiki.

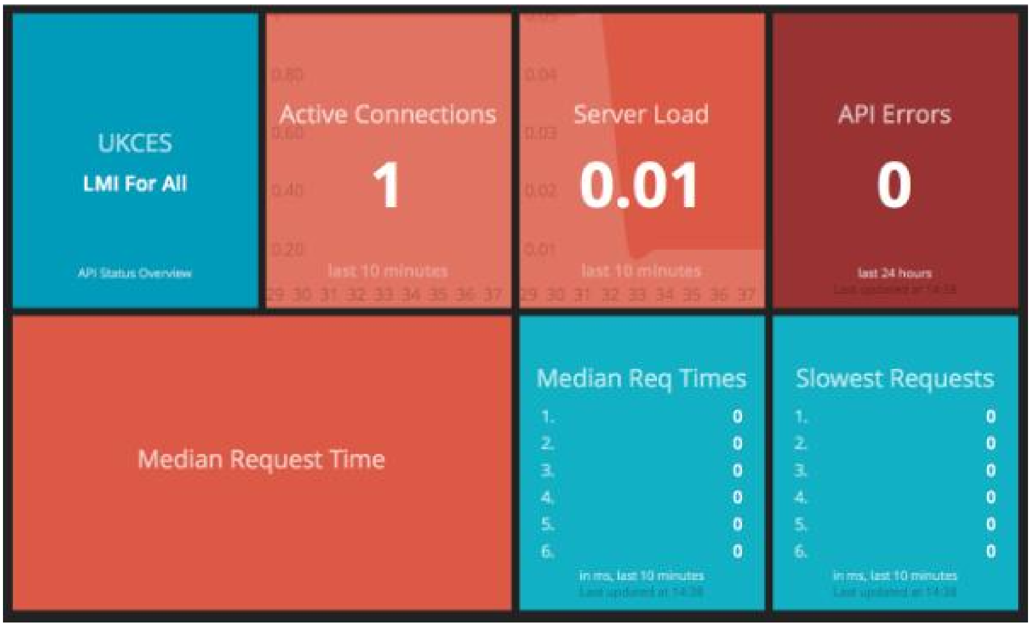
It took time to develop a shared understanding of the standards and form of documentation between members of the data team and the technical team. A template was produced (but can also be iteratively amended) and has been successfully implemented in the database development process. The process and documentation is robust and offers a quality standards approach for project collaboration.

3.6. Improved API reliability

The LMI for All API developed for the project allows external queries to be made to the database. The results of these queries, such as employment or wage rates associated with a particular occupation, can then be displayed in a web page or external application. The LMI for All API has been rewritten from the ground up (compared to the prototype developed in Phase 1) to offer much greater robustness and performance. It is now compliant with a variety of industry standard server containers[12], allowing for more flexibility in deployment and management.

A dashboard (see figure 4, below for a screenshot) at (http://collab.lmiforall.org.uk:3030/l4a) has been developed, which allows the performance of the database to be monitored in real time and thus allow any issues to be identified. It also monitors the performance of the API in returning queries. Automated monitoring and statistics have been deployed to try and detect problems arising out of usage before they become critical.

Figure 4 LMI for All database dashboard



[12] A standard server or web container (also known as a Servlet container) is the component of a web server that interacts is responsible for managing servlets, mapping a URL to a particular servlet and ensuring that the URL requester has the correct access rights.

3.6.1. API reliability issues and solutions

The previous version of the API (developed in the pilot) was written in PHP, which is a widely used general purpose language for web applications. For Phase 2A, key parts of the API were focused upon, such as speed, reliability, robustness and ease of deployment. This was facilitated by changing the programming language and programming framework to Scala and Scalatra[13], which is intended for use in precisely this kind of scenario. The previous version of the API had limited built-in monitoring, which was acceptable as it was a prototype intended mainly for the hack and modding days. The Phase 2A LMI for All API is public facing, so automated monitoring is a necessity to detect critical performance problems early and prevent service outages. System performance is now sampled in real-time and visualised on a system status dashboard. Automated watchers detect critical errors and outages and alert the technical team by email to ensure the service is not disrupted.

3.7. Improved API documentation and Client Code Generation

The new LMI for All API has been designed to be developer friendly as it allows developers to easily explore the functionality implemented in the API, as well as make test calls 'on the fly'. Many APIs are documented by hand. This can be a laborious process and requires considerable experimentation before testing and developing applications. The API automatically generates documentation from the running API code. This means that the documentation and actual functionality of the API are always in sync and up-to-date. Since the documentation is automatically generated by machines, it follows a structure that is also machine-readable. Software may interpret this documentation using automated processes and generate human-readable documents, exploratory applications, or even client code.

Having a machine-discoverable API makes it very simple to automatically generate client code for a variety of programming platforms. These API client libraries significantly reduce the time it takes developers to use the API and implement the various methods it supports. They provide code for different programming languages to communicate correctly with the API layer. API client libraries are automatically generated for the LMI for All API in many programming languages (most notably Python, Ruby, JavaScript, Java/Scala, Objective-C and PHP). An API discovery web app is provided to developers, which automatically explores the API and generates sample requests for data.

3.7.1. API documentation and Client Code Generation issues and solutions

The pilot version of the API also had auto-generated documentation, but it was less detailed and not machine-readable, since the source for the auto-generation was a collection of hand-typed text snippets. The Phase 2A documentation generator actually analyses the structure of the API itself and provides quite strict and detailed documentation (enhanced with human-written comments and explanations) in machine-readable form. This machine-readable form can be explored by developers using the provided API Explorer, which has received a high level of acceptance by developers on the hack and modding days.

[13] Scalatra (using Scala) is a web micro-framework that helps the developer quickly build high-performance websites and APIs.

3.7.2. Tracking and API keys

Functionality has been developed to track API usage across websites, and authorise API access via the use of per-entity access keys. This functionality was disabled for a period of time after the launch of the API, pending server upgrades and discussions with stakeholders about probable usage and potential cost-sharing agreements with high-volume users of the API. This functionality will be installed in Phase 2B.

3.7.3. Tracking and API keys issues and solutions

It is a common practice to limit the rate of access to public APIs to avoid overloading the server. On the hack day, one team of developers downloaded a dataset by bulking queries, which will not be allowed for the public release as many queries such as this would overwhelm the server. The envisaged access model for the API is that queries will be limited to a rate of 30 queries in 60 seconds by IP address (i.e. individual end users). Institutions with higher volumes of data will need to apply for an API key, which raises the rate based on the number of IP addresses/computers the institution has deployed. The present tracking allows us to identify fully the use of the API. However, if in the future we needed to limit the API to a specified number of hits, then we would need the API key. The API could also track the originating websites that display LMI to their end users, and limit rates based on this information. If a single website uses up a great volume of API query capacity (such as a large careers guidance service), this could perhaps be limited until a way forward is agreed, such as a cost sharing plan.

However, none of these methods are currently active. Keeping track of IPs and originating websites requires memory capacity that the servers currently do not have. It is recommended that for Phase 2A and 2B (and after providing enough memory), a position on limits and costs be agreed, depending on the volume of queries the API will receive in practice. It is moot to try and predict the query volume on a fully released public API, so the best way forward is a ‘ramp-up’ model, where system capacity is scaled along with usage.

3.8. LMI for All web portal

A lightweight web portal has been developed at http://www.lmiforall.org.uk. The present portal is designed to provide access to information about the project and data tools for both careers advisers in different organisations and to application and web site developers. The web portal uses a content management system, which is simple-to-use and has a strong permissions environment. This should allow both members of the project team and UKCES to edit the web portal if they wish with minimal training needs.

3.8.1. Web portal issues and solutions

One issue that has been discussed is web portal branding. It was decided that the current web portal would include the UKCES logo, but with light branding. There was also some discussion over the aims for the web portal, target groups and content. It was agreed that the site should be lightweight, but would be revised after the release of the database (see issues for Phase 2B below).

3.9. Improved search

One objective under Phase 2A of the project was to improve the search functionality on the database. Three different approaches were proposed, including:

- Providing a semi-naturalised query interface (similar to a Google search box or Wolfram Alpha);
- Developing a set of pre-made parameterised queries for high-interest topics (such as query macros);
- Improving data cleanliness, such as in the case of vacancy information, through geo-tagging and full-text search.

The three proposals could not be completed with Phase 2A, so further discussion on priorities with UKCES were necessary. The selection of processes was largely objectively determined. Due to the launch of the Universal Jobsmatch service, providing access through an API, but limited by a proprietary occupational coding system, there was little that could be undertaken to improve data cleanliness or geo-tagging from this source of data. Equally, until the release of the API and database, the high-interest topics for users will be unknown. Therefore, the development of a semi-naturalised query interface was undertaken. This provided extra value since many end-users are already familiar with interfaces like Google. The present query interface is based on the index of 28,000 job titles linked to SOC unit groups. However, in general LMI is insufficiently key worded to yield many useful matches, for instance on pay and gender etc. In Phase 2B, we will further explore how a more advanced natural language query interface might be developed.

3.9.1. Improved search issues and solutions

Improvements have been made to searching for SOC codes. Since detailed textual explanations of the requirements, tasks and other descriptions for every SOC code have been included in the database, it is comparatively easy to run fuzzy text searches over these data, yielding better matches than the pilot SOC code search. However, other data are not explicitly described, so breaking down a textual query into its constituent LMI dataset parts is complex. A list of keywords, conjunctions and syntactical positions will have to be developed in Phase 2B that maps query terms more or less accurately to actual queries on the data.

3.10. Terms and Conditions of Service

The process of developing Terms and Conditions of Service for the API was complicated and work was undertaken jointly between the technical development team, UKCES and key stakeholders. For Phase 2A data and all tools were released under an Open Government license.

3.11. FAQ

An initial draft of Frequently Asked Questions has been developed to be included on the web portal. These have been based on questions that have been asked by both technical developers and providers of careers advice. In Phase 2B of the project, a more interactive means of developing the FAQ, including the provision of a forum, may need to be developed.

3.12. Implications for Phase 2B

3.12.1. Server infrastructure

A major issue for Phase 2B is the server and technical infrastructure required. This is problematic for a number of reasons and was underestimated in Phase 2A. Most web-based technical services start with a closed system, before releasing a closed beta and a more open beta, prior to open release. This process, firstly, allows the debugging of systems. Secondly, it allows feedback on features, design and functionality. Critically, it also allows the testing of the system under real user conditions, albeit with a controlled number of users. Hold-ups and delays can be identified and systems optimised for speed and reliability. The time constraints for Phase 2A have not allowed this to take place. Various checks to ensure that the data reported in response to queries are consistent with published information have been undertaken in Phase 2A. This will be continued in Phase 2B, including responses to any queries and concerns raised by users as the database is used more intensively.

The API will be available for use from the start of Phase 2B. It is quite possible that the early adopters will be national career organisations, which will bring considerable numbers of users and a large number of requests to the database. The development of an advanced dashboard is of considerable assistance to the technical team in being able to monitor performance of different components of the system in real time. A method of overcoming potential problems with high user demand from these organisation is to add extra server capacity as it is needed and the use of cloud computing will enable this process. Over demand is seen as good in that such ventures are often running at a loss, but are able to raise more capital if they are seen as attracting a considerable user base.

LMI for All is a public service, supported by limited public funding, so some caution is required. Therefore, a series of scenarios have been presented to UKCES to estimate the infrastructure required and provide costings for different Phases of 2B.

Access to the API is tracked and the number of hits from different applications and organisations can be monitored. This will enable the technical team to build in rate limiting, meaning that the number of queries from any one source over a given time period to ensure the database does not become overloaded and can be limited. In particular, the use of rate limiting allows us to block bulk querying (as discussed in Section 3.7.3, above). Without accurate data on the likely number of queries, it is difficult to advocate any one approach. The estimates provided are based on best estimates. It is also quite likely that server costs will continue to fall over the next two years, easing resources. The most important step is that we closely monitor usage, especially when large careers services come on line and regularly review performance and infrastructure needs.

3.12.2. Linking data and SPARQL

In the current setup, the API Layer acts as a firewall between the data users and the actual database. Data access is restricted to the queries that are run by the API layer. There is no direct data access allowed.

At present the API allows queries through JSON, a text-based open standard designed for human-readable data interchange. It is language-independent, with parsers available for many languages, and is an industry standard familiar to most, if not all, developers. The JSON format is often used for serialising and transmitting structured data over a network connection. It is used primarily to transmit data between a server and web application.

Some developers have asked the technical team to provide enhanced access to the database and in the future the possibility to provide SPARQL or RDF access to a dedicated data mart that only contains guaranteed non-disclosing information will need to be considered. SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language) is an RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format. SPARQL is a format favoured by linked data proponents as it allows advanced queries and the ability to query between different datasets.

It is at least technically feasible to provide most or all of the LMI for All data via the SPARQL query language. However, this is a non-trivial amount of work and will likely require additional developer capacity in the technical team. It seems prudent to first project the actual usage of SPARQL queries on the API before committing to a decision. SPARQL is not widely in use outside the open/linked data communities, and the general impression is that most 'regular' developers are fairly ambivalent about it. However, although LMI for All is largely based on open standards, RDF or SPARQL access form part of the Government's 5 star open data agenda.

It has also to be considered whether most of the early adaptors and users in terms of service will come from the open data movement or from developers offering more standard services for careers web sites. Furthermore, increased functionality for cross querying of different data sets through extending the present API and improving indexing, rather than investing heavily in SPARQL development, may be considered.

3.12.3. Non-standard data

The design of the LMI for All database is based on data integration through the SOC2010 4-digit classification system. However, it became apparent in Phase 2A that some data were not available at SOC 4-digit level.

Discussions have been held with a number of organisations that have their own APIs to data, but do not use the SOC 4-digit classification system. In some cases these data are seen by developers and end-users as extremely important. One example is the Universal Jobmatch service that provides access to employment opportunities and for which there is also an API. However, the classification system used does not match SOC 4-digit and is not a UK proprietary standard. In response, a technical approach called approximate string matching (often colloquially referred to as fuzzy string searching) has been applied. This technique is based on finding strings that match a pattern approximately (rather than exactly).

This approach is obviously not as accurate as matches based on the SOC 4-digit classification system. However, it allows access to a wider range of data than would otherwise be possible and, as long as end developers understand the limitations, this provides opportunities for more innovative applications and has been implemented within the API.

3.12.4. Web portal and app store

The present web portal has been developed to provide ready access to the outcomes of Phase 2A. At present, it targets technical developers interested in building web or mobile applications based on the API. The hack and modding days provided a view of the wide range of potential applications as well as the potential use of LMI for All by a variety of end-user groups. Various issues were identified, including to what extent to:

- Drive application development; and
- Provide assistance to careers professionals and application developers with their developments.

These lead to further questions of how to best develop the web portal (for instance through highlighting instances of effective or innovative use) and how much to encourage and support developers and users in sharing ideas and applications. One approach to achieving this is possibly through an LMI for All application store allowing easy access to highlighted examples of project implementations. Positively, the development and maintenance of a vibrant web portal with support services for users will promote uptake, but this will require resources that may be better expended in further data and technical development.

3.12.5. Database integration and architectures

To a considerable extent, some of the difficulties and limitations that were encountered in the database integration and architectures were anticipated. However, within the timing and resourcing of Phase 2A the technical team were unable to develop some of the more complex and efficient data marts that were considered to represent the best architecture for the project. However, an upgrade to MS SQL Server has been successfully achieved, which provides a professional database environment. Importantly, this allows for the implementation of formal automated ETL processes and ETL processes have been implemented for all current data sources that reduce manual tasks in the import and integration process.

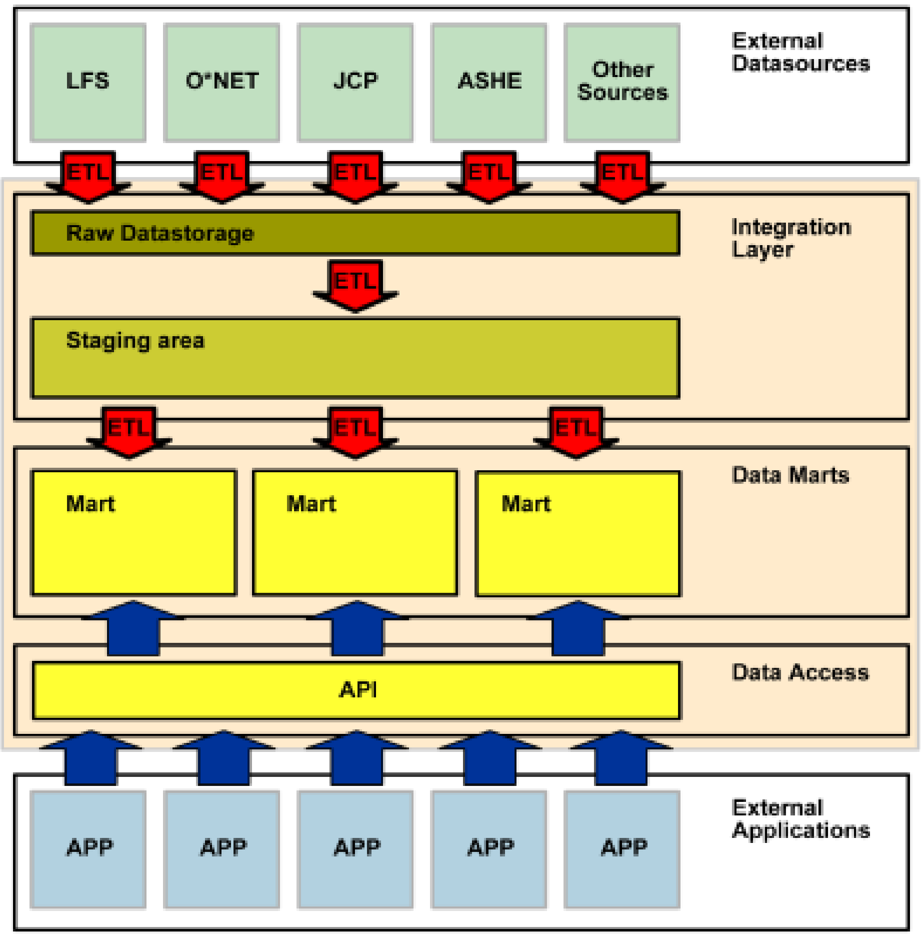
The environment to allow for the introduction of separate datamarts for various functional needs has also been successfully implemented. Currently, a single data mart in the form of the current ‘production’ namespace is maintained.

In Phase 2B, it is proposed to move to a more advanced infrastructure (see figure 5, below):

- **Data Integration Layer** – Data are imported from the external data sources into the raw data storage section in the Data Integration Layer. The ETL processes that are used for this initial input will have basic data cleaning tasks built-in. The resulting data in the raw data storage section will be error free and ready for use. From the raw data storage section, a separate set of ETLs is used to combine the data in a single staging area. In this step the data will be re-indexed (and where necessary recalculated) to a common set of dimensions.

- **Data Marts** – This layer holds a number of multidimensional data cubes or relational databases that hold pre-calculated query results. Depending on the measures and dimensions that are needed to feed the data requested through the API there may be a number of separate marts defined in this layer.
- **Data Access** – This layer holds the API and its interface to the available Data Marts. This would also be the access point for SPARQL or RDF if a decision was taken to develop this. External applications can access the pre-calculated data in the data marts by using strictly defined API calls to the Data Access layer.

Figure 5 Representation of the LMI for All platform and database recommended for Phase 2B



The data marts should greatly improve the efficiency of the database and overcome some of the issues encountered in Phase 2A. As the data marts will contain pre-calculated query results, they will provide fast access and will withstand a heavy volume of simultaneous queries. They will also, potentially, allow us to extend the range of data queries.

However, it should be noted that in the original project application the Open Data White Paper (HM Government, 2012) highlights how data gathered by the public sector is not always readily accessible (see section 1.1, above). Quality of the data, intermittent publication and a lack of common standards are also barriers. A commitment is given to change the culture of organisations, to bring about change: ‘This must change and one of the barriers to change is cultural' (p. 18). As part of this process, the Government intends to adopt the Five Star Scheme as a measure of the usability of its Open Data (p. 24). Most data providers have been keen to collaborate with UKCES and the project team. However, access and formats of the data remain varied and this will impact on the extent to which the ETL processes can be automated.

4. Stakeholder engagement and communication

This section of the report outlines the processes followed for the hack and modding days, details the applications (apps) and interfaces (or 'hacks') developed, and summarises feedback collected from the hackers and careers stakeholders. Issues, resolutions and recommendations for Phase 2B are also identified.

The main purpose of the stakeholder and communication activities for Phase 2A was testing the LMI for All database and API before it was released through both a 'hack day' and 'modding day'. The events were held at a central London location in March and April 2013, with Rewired State leading on the organisation of events. The primary audience were developers (or hackers), with the main purpose to ensure that the LMI for All database and API are sufficiently accessible and in a format where they can be easily used by software developers, at this stage of the project. The secondary audience were career stakeholders. Their involvement was crucial to gauge their responses to the relevance and useability of the applications produced from the days.

4.1. Testing the database API

Hack days (also known as Hackathons or Appathons) bring together experts and developers to collaborate or work alone rapidly prototyping software or hardware, building mobile and web apps or quick models for new ideas and features. The aims of a hack day are to: solve problems; test new data; test and launch new APIs; come up with new ideas or apps; or to highlight issues and areas of improvement. This is called a hack day, because developers are hacking projects together in a short space of time, experimenting, improvising, creating and playing. The modding day follows a hack day. Its aim is to take forward the developments of the hack day and to produce a more useable and defined product.

The LMI for All 'hack day' was organised for 12 March 2013. The objectives were to:

- Test the functionality of the LMI for All API;
- Develop apps that used the LMI for All API to demonstrate the potential; and
- Present the apps developed during the day to key stakeholders working in careers to get feedback for suitability and relevance for practice.

4.2. The developers

Ten experienced developers were recruited, competitively, from a strong field of 26 expressions of interest. The developers, of which nine are male, ranged (where age was recorded) from 15-41 years. Developers are variously involved in: accessibility and open data; front-end and back-end development; product management; IOS developments; social and mobile apps development; and API development. Skills included: HTML5; CSS; Photoshop; wireframing and semantic web technologies; 3D visualisations; Python, Perl, PHP; Java; Javascript; GIMP; OpenGL; JQuery; and Graphics Programming. One developer has particular skills as a designer and in social media. Overall, the developers represented a wide-range of skills and knowledge.

4.3. The careers stakeholders

Ten leading experts on the use of data in careers guidance were also invited to attend the hack day, to validate the source data and the applicability of such data for career decision-making in a parallel meeting. Ten accepted the invitation with alacrity, but on the day, one was not able to attend. The purposes of the involvement of the careers stakeholders were to:

- Find out more about the LMI for All project and its future;
- Shape the development of the data tool by feeding into the future thinking about the LMI needs of potential data users; and
- View and comment on the working examples of career apps, produced by developers on the day.

Prior to the hack day, the careers stakeholders were invited to complete a pre-event questionnaire (a summary of the responses are included in Annex D). Careers stakeholders represented a range of sectors, including: education; charity; and private. They comprised: a freelancer and independent trader; employers; managers; and employee/organisational representatives. Their roles varied from LMI/Information specialists, careers guidance professionals, managers to website/media developers. The majority of stakeholders noted that they are not restricted to one geographical area, some covering national and international contexts (such as Canada and the USA). Others were restricted to London, the South East and the Midlands. Overall, they represented a range of client situations. A summary of information on the careers stakeholders was provided to the developers to provide some background and context to the hacks they would be designing and developing.

4.4. The hack day

One week before the hack day the developers were given access to the web portal with information about the data and how to access the LMI for All API. The developers were also provided with a set of 11 use cases to provide some context to the database and potential uses (see Annex D for the use cases). Additionally, they were provided with a summary of the careers stakeholders pre-event questionnaire (as explained above, Section 4.3 and Annex E).

At the start of the day the developers chose to work together as a team to outline ideas and possibilities for apps and websites before dividing into teams or working individually in order to develop and present their ideas. During the day, they produced early stage prototype applications and interfaces, thus validating the web portal. The project team who had built the web portal were available throughout the day to respond to queries and fix any errors identified by the developers. Seven applications and interfaces were developed[14]. At the end of the hack day, the seven applications/interfaces were presented and judged by the careers stakeholders. The following table describes the seven apps and interfaces (or ‘hacks’) as well as summarising the feedback provided by the careers stakeholders.

[14] Screenshots of the apps and interfaces can be viewed at http://hacks.rewiredstate.org/events/lmiforall.

Table 1 Descriptions of apps and interfaces, plus stakeholder feedback

Apps and Interfaces – Descriptions	Feedback from careers stakeholders
On Demand	
Users search by region and are shown jobs that are hardest to fill and most in demand by employers. Forecast and regional comparative data are available to show forecast employment in the occupational area.	❖ Using hard-to-fill job data was considered very innovative ❖ Interesting perspective ❖ Great potential to add in more data – such as vacancies, skill sets etc. ❖ Great app to support advisers' work ❖ Particularly liked the regional focus ❖ Felt the forecasting element useful/innovative
JobBungee	
An IOS app that presents LMI facts about a job, displaying live job listings and typical CVs. It aims to help people with career plans by providing smart access to LMI data. It could display: relevant LMI data; typical education requirements; whether vacancies are hard-to-fill; average salary (breakdown by region); live data from Jobcentre Plus and other job sites; and real sample CV data from LinkedIn and other CV sites.	❖ Really liked the mash-up element of app ❖ Style and accessibility good ❖ CV element liked – particularly the way it provided access to CV depository ❖ Great potential to link to other datasets and careers videos ❖ Simplicity appealing
Job Quest	
A 90s Role Playing Video Games style character generator where users rate their skills and attributes resulting in a suggested job. The user could build a virtual version of themselves, increasing personal attributes and skills such as programming or mathematics by undertaking various activities or quests (such as reading suggested texts, undertaking skills development courses). When altering their skills, the user is presented with the most appropriate job the application can find in the database.	❖ Great combination of data and potential to expand and look at future possibilities ❖ Liked the role playing idea and thought this would be appealing ❖ Liked exploration element of app ❖ Needed something more visual than just borders ❖ Gaming element where users awarded points was appealing ❖ Liked the way it used skills and traits (not one or the other)
Job Trumps/Job Recipe	
This used core employment data and visualised it in a fun and approachable way for use by young people, key points of the card would be interactive to provide the user with further information on how to develop key skills for specific job roles.	❖ Design element and style liked – visually appealing ❖ Viewed as fun and accessible ❖ Thought this was very accessible for younger audiences ❖ Addressed equality and diversity issues

Table 1, continued…

Apps and Interfaces – Descriptions	Feedback from careers stakeholders
Linda	
This was designed as a digital, data backed, careers adviser using LMI and other data to help students select careers. She can be expanded with additional datasets and questions, and can use APIs to change her advice when circumstances change.	❖ Not considered new in what it offered ❖ Good use of combining data ❖ Simple step-by-step approach appealing ❖ Possibility of organisations personalising 'Linda' attractive ❖ Concerns that this would be seen as replacing careers advisers in the future
LMI Everywhere	
This is a reusable widget to distribute labour market predictions by providing a simple interface, which allows search for forecasts about the future labour market in specific jobs in the different UK regions. It can detect which region is relevant by geolocation or by configuration and comes ready to be installed in any website through a simple script tag.	❖ Widget that could be embedded in other sites was considered very appealing ❖ Seen as having most potential ❖ Liked that it contextualised with individual location ❖ Future predictions of jobs particularly appealing
Quick Stats	
This used data from the O*NET to represent the relative importance of each subject enabling easy interpretation of key information and comparison between job opportunities.	❖ Liked that it highlighted skills and comparative element ❖ Unsure that this would be understandable to wider audience ❖ Liked the design element and how it worked on a smart phone ❖ Visually accessible ❖ Use of O*Net categories (location/age) positive

Four hacks were selected to be taken forward for further development during the modding day, with two of these merged to create three applications overall. The hacks selected address different profiles of potential users:

- JobQuest;
- JobBungee/On Demand;
- LMI everywhere widget.

Overall the feedback from the hack day was positive and the future potential of the LMI for All web portal and API was proved. It was considered that good use had been made of the available data and those hacks that 'mashed up' or included a range of data from a range of sources were found most appealing. The range of apps and interfaces were thought to address the differing needs of users, to be appealing to wide ranging audiences, and address their learning styles and data needs. The potential to use these apps and interfaces on different devices was also found to be appealing.

4.5. The modding day

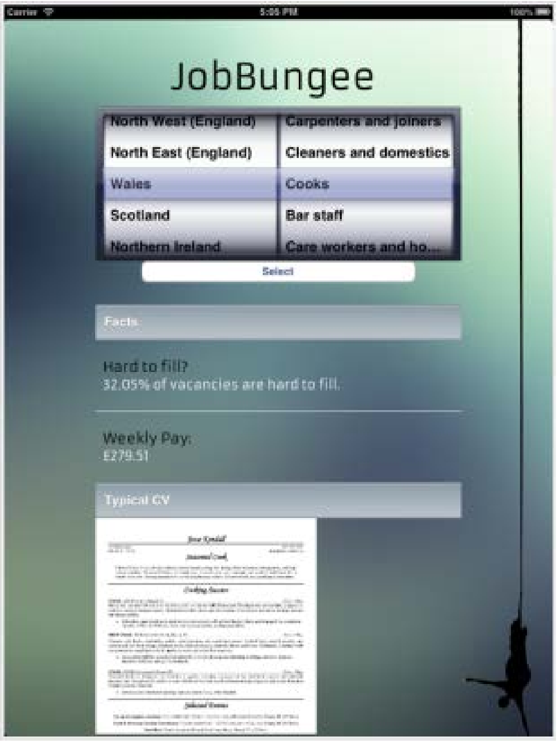
The LMI for All modding day was held on 24 April 2013, about six weeks after the hack day. It was designed to take the winning projects through another day of intensive development, so that they were nearer completion. The further day meant that the developers could take advantage of additional data added to the database (to expose via the LMI for All API) and combine this with the insights gained from the original hack day. Nine of the original developers worked together to improve the applications produced from the hack day and modify them to work with the newer version of the LMI for All API and database. Careers stakeholders who had attended the hack day, together with others representing their organisations, attended the end of the modding day in order to judge the presentations of the hacks that had been taken forward. The stakeholders were asked to complete a feedback form and vote on their preferred and second choice hacks. The feedback and results are summarised in the next section.

JobBungee/On Demand

JobBungee (IOS app) and On Demand (online interface) apps were combined for the modding day. The On Demand app enabled users to search for the hardest to fill and most in demand jobs by employers by region. Forecast and regional comparative data were available to show job prospects. The JobBungee app presented a range of key data (such as typical education requirements, whether vacancies are hard-to-fill and average salary) for a specific occupation combining it with live job listings and real sample CV data from LinkedIn and other CV sites (not occupational specific).

During the modding day the apps were integrated and a new navigation was developed for both the IOS and online interface. Workflow between the systems was also improved and various bug fixes undertaken. The app and online version were also improved to include: more LMI (salary and qualification data); searches for typical CVs (i.e. occupational specific rather than generic); videos from icould and youtube; plus graphs of jobs in demand by gender and employment status.

Figure 6 JobBungee/On Demand screenshots



JobBungee

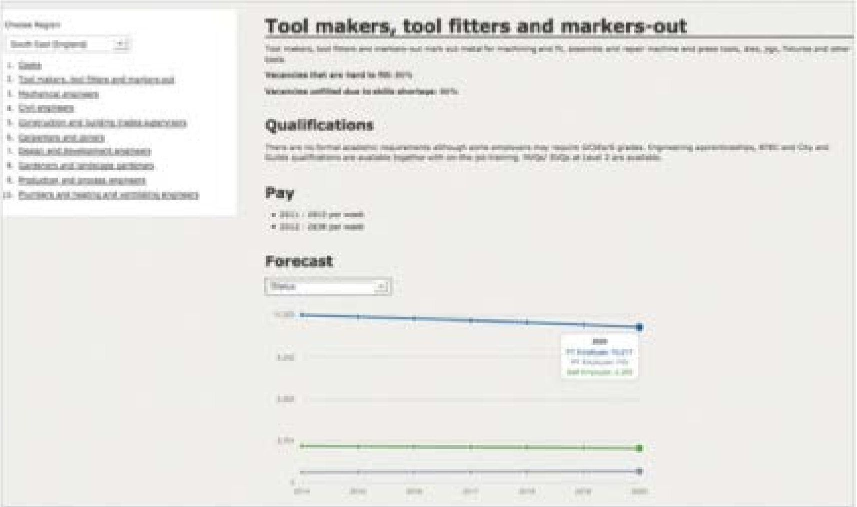
North West (England)	Carpenters and joiners
North East (England)	Cleaners and domestics
Wales	Cooks
Scotland	Bar staff
Northern Ireland	Care workers and ho...

Facts

Hard to fill?
32.05% of vacancies are hard to fill.

Weekly Pay:
£279.51

Typical CV


Choose Region

South East (England)

1. Cooks
2. Tool makers, tool fitters and markers-out
3. Mechanical engineers
4. Civil engineers
5. Construction and building trades supervisors
6. Carpenters and joiners
7. Design and development engineers
8. Gardeners and landscape gardeners
9. Production and process engineers
10. Plumbers and heating and ventilating engineers

Tool makers, tool fitters and markers-out

Tool makers, tool fitters and markers-out mark out metal for machining and fit, assemble and repair machine and press tools, dies, jigs, fixtures and other tools.

Vacancies that are hard to fill: 80%

Vacancies unfilled due to skills shortage: 80%

Qualifications

There are no formal academic requirements although some employers may require GCSEs/S grades. Engineering apprenticeships, BTEC and City and Guilds qualifications are available together with on-the-job training. NVQs/ SVQs at Level 2 are available.

Pay

- 2011 : £613 per week
- 2012 : £636 per week

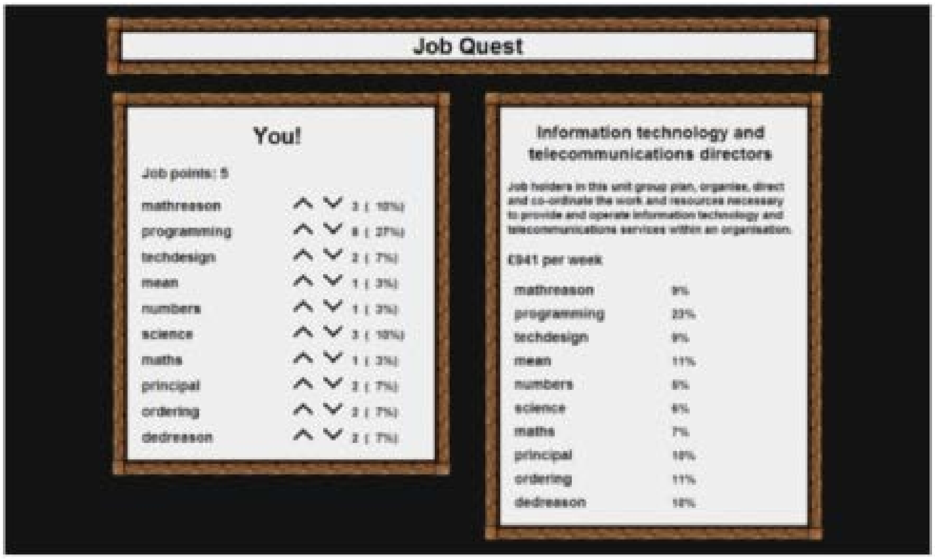
Forecast

JobQuest

This development was founded upon the idea of Role Playing Video Games where users rate their skills and attributes resulting in a suggested job. The idea is that users can build a virtual version of themselves, increasing personal attributes and skills by undertaking various activities or quests (such as reading suggested texts, undertaking skills development courses). When altering their skills, the user is presented with the most appropriate job the application can find in the database.

During the modding day, a mobile version of JobQuest was created. The app was integrated with the Google Books API in order to returns a list of advised readings for the skills in which the user has identified as in need of development. This was created via an API that accepts the percentage values for the O*NET skills composition of both the description and the occupation (using Standard Occupational Code) found.

Figure 7 JobQuest screenshot



Job Quest

You!

Job points: 5

mathreason	3 (10%)
programming	8 (27%)
techdesign	2 (7%)
mean	1 (3%)
numbers	1 (3%)
science	3 (10%)
maths	1 (3%)
principal	2 (7%)
ordering	2 (7%)
dedreason	2 (7%)

Information technology and telecommunications directors

Job holders in this unit group plan, organise, direct and co-ordinate the work and resources necessary to provide and operate information technology and telecommunications services within an organisation.

£941 per week

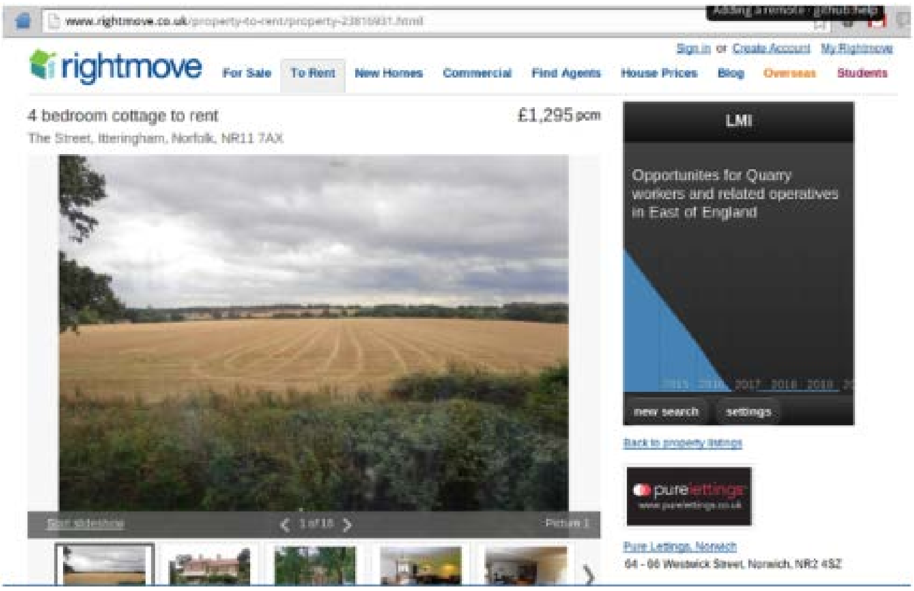
mathreason	9%
programming	23%
techdesign	9%
mean	11%
numbers	8%
science	6%
maths	7%
principal	10%
ordering	11%
dedreason	10%

LMI everywhere widget

The LMI everywhere widget was designed as a reusable widget to distribute labour market forecast data by providing a simple interface, which allows searches for forecasts about the future labour market in specific jobs by region to be visualised. It was designed to detect region by geolocation or by configuration. At the end of the hack day it was ready to be installed on any website through a simple script tag.

Figure 8 LMI everywhere widget screenshot

Note The LMI widget has been embedded in the Right Move website.



www.rightmove.co.uk/property-to-rent/property-23816931.html

Sign in or Create Account My Rightmove

4 bedroom cottage to rent £1,295 pcm

The Street, Itteringham, Norfolk, NR11 7AX

LMI

Opportunites for Quarry workers and related operatives in East of England

Back to property listings

Pure Lettings, Norwich

64 - 66 Westwick Street, Norwich, NR2 4SZ

4.5.1. Feedback from careers stakeholders

Of the 13 voting and comments cards received, nine careers stakeholders reported that the integrated JobBungee/On Demand app was their preferred choice. The LMI everywhere widget was voted second. Feedback from the stakeholders on each of the apps included:

- Feedback on **JobBungee/On Demand** from the careers stakeholders was focused on the range of information available through one interface in a readable and accessible format. Stakeholders felt that this interface contextualised the data very well and that it was an excellent approach to displaying LMI data. They also considered the focus on jobs in demand most interesting and appealing. The addition of careers videos was positively received. Some suggested that greater regional disaggregation would be useful. Many stakeholders liked the addition of CVs in the interface. However, a few raised concerns about the CVs selected and felt that these required some kind of moderation or quality check. Overall, JobBungee/On Demand was preferred because of the interface and its design, the approach to displaying data and the focus on jobs in demand or hardest to fill. The mash-up of data was considered very powerful as this included all relevant domains for a range of audiences.
- The **LMI everywhere widget** was regarded as being a very good idea by many with the visualisation of complex data most appealing. Many noted the potential use of the widget, as it was considered flexible in its application. Others felt that the data could be misleading and that greater explanation would be needed in order for users to interpret the data. The contextualised and regional aspect of the data was most popular amongst the stakeholders. One suggested that data needed to be more local, whilst another suggested that data be displayed in different formats (i.e. a graph, bar-chart or map).
- The careers stakeholders thought that **JobQuest** would be most appealing to young adults. Although this was considered the least developed app, it was regarded as having the most potential for further development and application. JobQuest was evaluated as not being very user-friendly, with one stakeholder reporting that the gaming element had got confused with producing serious outcomes. Another stakeholder suggested that users would need help identifying skills and understanding how to fill gaps. It was noted that individuals are generally not competent in assessing their skills. Also, understanding types of skills could be complex for some. Others suggested that more LMI was needed. Generally, the inclusion of access to books was disliked. Despite these criticisms, the idea of getting individuals to assess their skills and abilities to be matched to relevant jobs with the gaming element was considered very promising.

The feedback from the careers stakeholders supports evidence from Phase 1 that the LMI for All database and API is not only effective in inspiring the development of apps and interfaces to support those making career decisions, but also in enabling access to wide ranging information on the labour market. To sum up, the hack and modding days have proven that there is great potential for the development of careers relevant apps and interfaces using a range of data for a variety of audiences and purposes.

4.5.2. Feedback from the developers

The developers were clear that the careers stakeholders participation and involvement in the hack and modding days had provided valuable input into the hack day development process. The case studies and pre-event questionnaire information had provided good background and contextual information for the design and functionality of the hacks. This process had ensured that hacks were of interest and, more importantly, considered useful.

The developers were unanimous in regarding the LMI for All API as 'amazing' in terms of its functionality and ability to deal with requests. Having the technical and data teams available throughout the hack day had also proved useful in order to respond to queries about the data and respond to errors in the database. Advanced notice of the data available in the LMI for All database through the API was also noted to be useful.

4.6. Next steps for stakeholder engagement and communication

For Phase 2B, the stakeholder engagement and communication will be undertaken through two major sets of activities. The first will be testing the detail and technical aspects of the database and API with developers to ensure that it is accessible and useful. This will be undertaken by organising a second hack day and data modding day in a similar process to the work undertaken in Phase 2A. In Phase 2B, it is proposed to extend the hack day to a two day process in order for developers to make more progress. It has been agreed that the same developers who participated in first hack and modding days will be invited to participate. This will ensure some continuity as they will be familiar with the database, API and the context in which apps and interfaces will be developed in the future. All developers invited were keen to continue their involvement in the project.

The second set of activities will be through a series of events with stakeholders that will: increase awareness of the database and API; gain feedback that can inform the final development of the database and API; and explore the potential for linking the database using the API from other websites. It is suggested that this would include the organisation of a conference with not more than 100 participants, presenting work to date and focusing on how useful the data are, and how it might be used by stakeholder organisations. In addition, a series of three stakeholder engagements involving not more than 30 participants each will be held. These will provide more focused opportunities to gain feedback that can inform the final developments of the database and explore the potential for linking from other websites with targeted stakeholder groups (for example, career practitioners and their managers).

To date, dissemination activities have been successful with, for example, members of the project team presenting at a careers conference and an open data conference. This activity has been successful in generating interest in the project from careers stakeholders and the IT and technical communities. Further activities (e.g. conference presentations, workshops and publications in a range of professional and academic journals) can be undertaken to maintain the momentum of interest. For example, the project team has been invited by the British Journal of Guidance & Counselling to submit an article to a forthcoming symposium edition on ICT in careers guidance. This represents one type of audience. The editor of the journal for the Careers Development Institute (CDI) has also approached the team with a request to submit an article to inform the practitioner community of this development.

5. Summary of Phase 2A progress and recommendations

5.1. Issues and resolutions

The original idea of simply tapping in to existing sources via standard APIs has proved to be problematic as the data sources are not designed to provide data at the level of detail required. This is the result of a combination of factors (including avoidance of disclosure, confidentiality and the need for statistical robustness). In many cases sample sizes in the surveys are simply inadequate to provide statistically robust estimates at the level of detail required.

However, what is needed in a careers guidance context is general information (especially on Employment and Pay). This can be developed from the official sources in a manner that can paint a useful picture without falling foul of problems of disclosure, braches of confidentiality or statistical imprecision. This involves developing estimates or predictions of the key indicators rather than relying on the raw sample data.

Based on that solution, adequate data are available to populate the LMI for All database, but this requires considerable processing to create a complete dataset for key indicators such as Employment, Pay and Hours.

In the longer-term, it would be better if the predicted estimates used for these three key indicators could be replaced by “raw” survey data, which could be updated automatically as they are published without the need for intermediate processing. In practice, this might pose some problems as discussed in the Annex (section A.6).

A second key issue relates to classification of occupations in particular. Some data are not classified in a manner suitable for inclusion in the database (notably vacancies reported by DWP and course information). Detailed occupational categories are at the heart of the database since occupation is one of the most important characteristics of jobs from a careers guidance perspective. However, not all data are classified in a consistent fashion. In this is a problem of changes over time, but there are some key datasets (notably that for current vacancies) where there appears to be a problem in some instances. Overall, it should be noted that an innovative approach has been taken and key data requirements have been successfully fulfilled.

5.2. Implications for costing and timescales

5.2.1. Data processing

The amount of data preparation undertaken during Phase 2A has greatly exceeded initial expectations and costings. However, many generic and specific issues have been raised and solved which will place us in a good position for Phase 2B.

5.2.2. Server costs

For Phase 2A, server costs were underestimated and costs increased due to space requirements and the higher SQL costs. As a consequence server costs for Phase 2B have not be adequately costed. Although it is difficult to estimate actual server costs for Phase 2B, as the size of server space required is dependent on the number of requests to the database. Three costing scenarios have been presented to UKCES.

5.3. Recommendations

- ❖ This exercise has demonstrated the practical feasibility of developing a data portal to serve the needs of the careers guidance and advice community. LMI for All should be further developed to meet the LMI needs of these groups (as well as other potential users in the longer term).
- ❖ The main indicators in the LMI for All database in Phase 2A (October 2012 – May 2013) should continue to be used in the next phase of the project (Phase 2B, June 2013 – March 2015), including:
    - Employment and employment forecasts based on *Working Futures* (these include information on qualifications and replacement demands);
    - Unemployment rates (using the International Labour Organization definition of unemployment[15]) based on the LFS;
    - Pay (estimates based on a combination of ASHE and LFS data);
    - Hours worked (ASHE);
    - Vacancy estimates (based on ESS and Universal Jobmatch);
    - Vacancies (based on a fuzzy search from Universal Jobmatch);
    - Occupational descriptions (ONS).
- ❖ Various refinements to the way these estimates are generated are proposed, some of which can be implemented in Phase 2B (e.g. focusing on medians/ deciles, rather than means). Others involve work outside the project (e.g. refining the projections of employment at the 4-digit occupational level, which will require an extension to the current *Working Futures* database).
- ❖ Further consideration of use of “raw” survey data as opposed to estimates/predictions.
- ❖ The full, revised O*Net dataset, including Skills and Abilities, as well as a number of other skill related indicators, should be implemented in Phase 2B of the project.
- ❖ Other possible indicators and enhancements to the LMI for All should be considered, including:
    - Further work to integrate UJM vacancy data into the database more fully, once mapping to occupational categories has been resolved.
    - Making greater use of data from higher education, such as HESA information on the destination of graduates (however, this will require detailed negotiation with

[15] The ILO definition of unemployment covers people who are: out of work; want a job, have actively sought work in the previous four weeks and are available to start work within the next fortnight; or out of work and have accepted a job that they are waiting to start in the next fortnight.

data owners).

  - Course information - a great deal of information is available about courses of study and links to different career paths, but this is not well coordinated or consistent. More work needs to be done to bring this into the database.
  - The UK Census of Population, especially local labour market information (since there is limited other sub-regional information), including some commuting and workplace data);
  - NOMIS, using the API to include workforce jobs data at regional level, the unemployment claimant count and data from the APS;
  - Cedefop pan-European employment database – equivalent to UK *Working Futures*, (but only available at 2-digit occupational level), move to the revised ISCO08 data as soon as they are available (early 2014) and exploit more detailed information (if and when it is published);

- ❖ The following should not be included in the database in Phase 2B: ONS Vacancy Survey (no occupational detail); Annual Population Survey (does not add much to LFS); Jobcentre Plus vacancies (historical data only – series discontinued); and EULFS (problems with availability and detail).
- ❖ Early discussions need to take place in Phase 2B regarding technical priorities and server capacity. The development and maintenance of a vibrant web portal with support services for users and developers will promote uptake. Consideration needs to be given to the amount of resources this will require, not only in technical terms, but in design, moderation and intervention to respond to and support developers and users. Such resources have to be balanced with priorities for further data and technical development.
- ❖ Continuous encouragement and support should be given to organisations with an interest in using the early release of the web portal and API, which is part of the approach to testing, evaluating and improving the pilot tool, as well as demonstrating the benefits to a wider audience. The nature and level of this support should be discussed.
- ❖ A more strategic use of social media and dissemination at key events should be adopted throughout Phase 2B, to ensure the web portal and API are promoted to create a market for the product and to maintain the momentum of interest. This will complement the planned dissemination events (three workshops and a conference), to be delivered throughout the Phase 2B.
- ❖ The successful format of the hack and modding days should be adopted in Phase 2B, since these were successful in not only proving the viability of the database, but also enabled career stakeholders to contribute to the development process.
- ❖ Active participation of key stakeholder representatives throughout the project should be carefully designed to ensure stakeholder engagement. This will be achieved through key stakeholder participation in future hack and modding days, alongside a conference and seminar events for the different stakeholders groups (namely careers representatives, policy makers and developers). Throughout Phase 2B, there will be an on-going dialogue with organisations that have expressed an interest in using the API and their feedback gathered in order to inform further refinements and amendments to the database and

API.

- Communication of the web portal concept should go beyond traditional dissemination methods (e.g. newsletters, professional publications, presentations at various events, etc.). Visual representations of potential applications should be made available to various audiences, in response to advice on priority target groups and their career needs collected from key stakeholders (e.g. the National Careers Service; Careers Wales; the National Apprenticeship Service; TAEN, etc.).

Annex A: Main data sources included in Phase 2A

A.1 Introduction

LMI for All aims to provide detailed data on a range of ley labour market indicators to those interested in careers prospects and progression (Bimrose, 2012). These include Pay and employment, plus a range of other labour market information.

The original design was to access various official datasets directly. However, concerns about breaching confidentiality and releasing disclosive data into the public domain severely limit the level of detail that can be published. Therefore, an alternative approach has been proposed for a number of the core indicators. This uses the official data to generate the detailed information required, but does not release the original survey data into the public domain (Bimrose and Wilson, 2013).

Annex A comprises:

- This section sets out the rationale for this approach and describes the information placed into the public domain.
- Section A.2 summarises the case for making detailed data on Pay and employment available as part of the LMI for All database.
- Sections A.3 and A.4 then set out in general terms how this has been accomplished, while at the same time ensuring this is non-disclosive (and not in breach of confidentiality restrictions recommended by ONS). Section A.3 deals with Pay and weekly Hours worked and Section A.4 with employment.
- Section A.5 briefly describes the other datasets and indicators included in Phase 2A.
- Section A.6 goes on to discuss some longer-term issues, including how official survey estimates might be improved to replace the predicted figures for the key indicators (employment, Pay and Hours).
- Section A.7 provides technical details of the regression analysis undertaken for pay predictions.
- Section A.8 provides technical details of the algorithms used to ensure that the predicted estimates for employment, Pay and Hours are consistent with the official published data.

A.2 The case for detailed data in the LMI for All database

The LMI for All database requires detailed data if it is to be useful for careers guidance and advice. Individuals and their advisers have a professional interest in knowing which jobs are available, distinguishing sector, occupation and typical qualifications required, as well the typical pay associated with those jobs.

Ideally, the full set of detail required is as follows:

- Occupation (up to the 4-digit level of SOC 2010, 369 Categories);[16]
- Sector (up to the 2-digit level of SIC2007, about 80 categories);
- Geographical area (12 English regions and constituent countries of the UK);[17][18]
- Gender and employment status (full-time, part-time employees and self-employed).

The main official data sources for such data are:

- the Business Register and Employment Survey (BRES);
- the Labour Force Survey (LFS); and
- the Annual Survey of Hours and Earnings (ASHE).

These sources collect data on individual organisations and individual people, but such detail cannot be published because of concerns about disclosure and confidentiality.

It is important to emphasise that the specific individual observations on Pay or employment from these official surveys are not necessarily required. What is needed is general information on 'typical' pay or general employment opportunities in particular areas for people with selected characteristics. The official data are a means to this end rather than being required for their own sake.

The level of detail required in the LMI for All database can be obtained by replacing the official 'raw' data by ***estimates*** or ***predictions***.

- For ***pay*** – these are based on an earnings function approach.
- For ***employment*** – the *Working Futures* employment database has been used.

Estimates of Pay and employment (by the detailed categories as described above), and based on these methods, form the core of the LMI for All database.

[16] Some have argued for an even more detailed breakdown to the 5-digit level of SOC, but this is not feasible given data currently available.

[17] Plus for some purposes additional information on: Age; Gender; Status; and Qualification (highest held).

[18] It should be noted that to enhance usability for careers professionals there would be merit in presenting sub-regional data where possible.

A.3 Providing detail without being disclosive – Pay and Hours worked

Pay: In the case of ***Pay***, an earnings function can be estimated using the original detailed individual data under secure conditions.[19] Such a function can then be used to generate ***estimates*** of pay (including confidence intervals) that are not disclosive.

A typical earnings function takes the form:

$$\text{Ln (E)} = a + b*A + c*A^2 + D*X + u$$

Where:

- Ln (E) is the log of earnings or pay;
- A is age;
- X is a vector of other explanatory variables which will include (inter alia) all the key dimensions as set out in Annex A.2;
- D is a vector of parameters associated with the vector X;
- a, b and c are also parameters to be estimated;
- u is the standard regression error term.

X includes:

- Gender (default is Male (0), a 1 indicates Female);
- Region (default is London, 11 other 0/1 dummies one for each other region);
- Sector (default is currently Agriculture, plus 78 other 2-digit SIC2007 categories as used in *Working Futures*)[20];
- Occupation (default is Chief executives and senior officials, plus 368 other 4-digit SOC 2010 categories);
- Qualification (default is a degree or equivalent and 5 other qualification categories[21] (highest held)).

Using the estimated parameters, point estimates of the typical pay of individuals in a range of different situations and with a range of different characteristics can be generated. In principle, these estimates could also include other indicators (such as the *median* or *quartiles*), as well as *confidence intervals* around these point estimates. For Phase 2A, the focus has been on mean pay only

The parameters have been estimated using the full and most detailed sets of raw individual data in ASHE or the LFS available (under the secure conditions imposed by the ONS Secure

[19] Other estimation methods than a standard earnings function might also be used. These might have some advantages, but for the present a simple standard earnings function is proposed.

[20] The regression using LFS data currently adopts the full set of SIC2007 2-digit categories, but it is proposed to replace those by the *Working Futures* 79 industry categories in the final version.

[21] Including ‘none’ and ‘don’t know’.

Data Service (SDS)). These parameters are then used to generate the estimates for the careers database. Table A.1 shows some typical regression results based on the publically available LFS dataset.

Note that data on pay could also be potentially ***disclosive*** if it were to identify a particular employer. It is necessary to treat pay as for employment in terms of addressing queries to the database, so that potentially disclosive information is not placed into the public domain. Effectively this requires some censoring (as described in Section A.4 in Annex A).

Some 'common sense' rules are imposed in dealing with queries to the database so that nothing unreliable is revealed. These rules are based on general ONS guidelines for dealing with LFS data (e.g. anything involving fewer than 10,000 observations (grossed up) will be flagged up as potentially unreliable. Anything involving fewer than 1,000 observations (grossed up) will result in a query defaulting to a higher level of aggregation and return a 'not available' message. This avoids generating estimates of pay where there are tiny (or even zero) numbers of people involved.

ONS were requested to confirm that the process described is in line with current rules regarding access to ASHE and LFS data via the SDS. This confirmation was achieved implicitly by the process of formal application to use the ASHE and LFS data via the SDS, and the checks imposed on the extraction of the relevant parameters from the SDS.

Hours: Information on weekly ***hours*** worked is also required. This has been obtained from ASHE. There is no obvious analogous approach that can be adopted using a simple earnings function type, as described above for pay. Due to technical problems of simultaneity, as well as the need to include external variables relating to economic cycle, etc., estimating an hours equation is not a straightforward option.

Nevertheless, this possibility has been explored using LFS data. If the focus was on predicting hours worked at an individual level, these issues would pose more serious concerns, but given that the focus is on average hours for broad groups it is less of a concern. A regression with hours of working being the dependent variable, and including all the other dimensions and interactive terms as independent variables as for the earnings equation other than age seems to deliver reasonable results. This could be repeated in Phase 2B using ASHE data once these are made available in the SDS classified using SOC 2010. In any event, variations in hours worked are much less significant than those for pay across occupations. Therefore, the focus is on providing broad-brush indicators across occupations and other key dimensions. In the current version of the database information on hours is not derived from an equation but is extrapolated from published ASHE data.

Indicators of part-time working can also be based in part on the *Working Futures* employment database described in Annex A.4. This provides, for example, information on the percentage of jobs that are part time. In the longer-term (Phase 2B) the database could also include other indicators taken from ASHE (or the LFS) such as banded hours figures for broad occupations (<.20; 20-40 and >40 hours per week). Such information would not breach confidentiality nor be disclosive even if carried out at a more detailed 4-digit level of SOC. However, there are many gaps in such data at the 4-digit level cross classified by other important dimensions, so this has not been attempted in Phase 2A.

Table A.1 Typical Earning Function Results

Variable	Coefficient
Age (continuous variable)	0.06
Age squared	0.00
(default =male)	
female	-0.15
(default =London)	
North East	-0.18
North West	-0.19
Yorkshire & Humberside	-0.19
East Midlands	-0.15
West Midlands	-0.18
Eastern	-0.10
South East	-0.09
South West	-0.18
Wales	-0.21
Scotland	-0.16
Northern Ireland	-0.21
(default =degree or equivalent)	
Higher education	-0.10
GCE A Level or equivalent	-0.17
GCSE grades A-C or equivalent	-0.23
Other qualifications	-0.28
No qualification	-0.34
(default=Agriculture, etc.)	
02 Coal, oil & gas	0.77
03 Other mining and quarrying	0.44
04 Mining support	0.69
05 Food products	0.21
06 Beverages & tobacco	0.33
07 Textiles	-0.15
08-79……………….etc, etc	*
(default =Chief executives and senior officials)	
1120 'Elected officers and representatives'	-1.24
1121 'Production managers and directors in manufacturing'	-0.52
1122 'Production managers and directors in construction'	-0.57
1123 'Production managers and directors in mining and energ	-0.40
1131 'Financial managers and directors'	-0.45
1132 'Marketing and sales directors'	-0.11
1133 'Purchasing managers and directors'	-0.36
1134-9274……………etc, etc	*
9275 'Leisure and theme park attendants'	-0.58
9279 'Other elementary services occupations n.e.c.'	-0.99
constant	5.82

*Notes: * The highlighted rows are missing from the table.*

A.4 Providing detail without being disclosive – Employment

A.4.1 Data sources and the problems of disclosure and confidentiality

There are two main official data sources for time series information on employment. These are the Business Register and Employment Survey (BRES) and the Labour Force Survey (LFS). Together with some other data they can be combined to provide a very detailed picture of employment patterns.

The BRES dataset is based on a survey of employers. It provides detailed information on employment (employees only) by detailed sector (up to 5 digits) and by detailed geographical location (down to Local Authority Districts). The key issue is whether or not the data are disclosive (i.e. can individual companies/ units be identified).

In fact the BRES data are collected for workplaces or establishments (units) rather than companies or enterprises. Nevertheless, the potential for identifying the information as pertaining to a particular company is obvious. For some sectors where there are only one or two companies operating, this may be a problem even at a UK level (for example there is only one manufacturer of Nuclear Submarines). Therefore, if the sectoral level of disaggregation is detailed enough such a company will inevitably be identifiable. If sector is cross classified by geographical area, there are many more companies that can potentially be identified (for example, there is only one company that produces cars in Derbyshire).

The LFS dataset is a survey of households and individuals. It provides information on occupation and qualification as well as industry and region. In principle, it can be used to identify individual respondents. Given enough dimensions (age, gender, location of employment, sector, occupation, qualifications, etc.) it is possible (in principle) to identify the individual that has responded to the survey. Revealing this information, and any associated survey data, would breach confidentiality.

Providing detailed ***estimates*** for employment analogous to those described for pay is much more complex. There is no simple analogy to the earnings equation which can be used to produce econometric estimates of employment as an alternative to publishing the raw survey estimates. However, there is an alternative set of very detailed employment estimates available that has been developed by IER on behalf of UKCES. It covers all the main dimensions needed (although currently only up to the 2-digit level of SOC). It is constructed using various official datasets, available either in the public domain or through NOMIS (subject to a Chancellor of the Exchequer's Notice (CEN)). This is the *Working Futures* database.

The sectoral aspect (which at its heart is based on BRES data) is potentially problematic because of concerns about ***disclosure***. Although the data in the *Working Futures* database are not the raw BRES numbers,[22] for some sectors there may be only a handful of organisations involved, especially at a sub-UK level, so potentially these cases could be identified from the ***Working Futures*** data. The key question is how to deal with this problem (of not being disclosive) while providing as much detail as possible?

[22] In practice, the *Working Futures* database does not use the BRES data as such, but makes use of the various sectoral employment times series ONS publish based on BRES and made available via NOMIS under the terms of a CEN.

A.4.2 The *Working Futures* database

The numbers within the *Working Futures* database are *estimates*, just as the pay figures from an earnings function are.[23] The *Working Futures* database is the result of a complex combination of datasets, models and assumptions (including various iterative procedures).[24,25]

The *Working Futures* database does not include any of the original raw survey data upon which it is based. Given all the adjustments, assumptions, and amendments made to the data, the final *Working Futures* estimates of employment numbers are far removed from the original source data (BRES and LFS).[26]

Nevertheless to avoid any risk of disclosure, and to comply with the terms under which the data used to construct the database have been obtained via NOMIS (under the auspices of a Chancellor of the Exchequer's Notice (CEN)) the *Working Futures* numbers are currently made available only to those who have signed up to a Secondary CEN). This has been done to ensure that there is no breach of the rules on confidentiality or disclosure (as some of the underlying data were obtained via NOMIS under the terms of a CEN).

Where sector is not involved, there is no danger of disclosure since identification of a company or unit depends on sector. However, sector is an important aspect from a careers guidance perspective, so it is not possible to simply remove it from the LMI for All database.

A.4.3 BRES information on number of establishments/units

ONS publish information that can be used to assess the sample size (number of units) on which the *Working Futures* employment dataset is based. This enables the risk of disclosure to be assessed. The data source for this information is the Inter Departmental Business Register (IDBR), which is the sampling frame for the BRES and ABI surveys (which in turn underlie the *Working Futures* employment estimates).

[23] Effectively the generation of the *Working Futures* database can be regarded as equivalent to estimating the probability of employment in a certain category defined by: industry (79 categories); occupation (currently there are only 25 2-digit SOC categories, but this can in principle be extended to the 369 4-digit categories); gender; status (3 groups full-time, part-time employees and self-employment); 'region' (12 countries and English Regions within the UK (or many more areas if we were to include LEPs)); and qualifications. These probabilities sum to 1 when added up across all these dimensions. Applied to an estimate of total UK employment they generate an employment estimate analogous to the pay estimates from the earning function.

[24] For full details of how the *Working Futures* database is constructed see Wilson and Homenidou (2012b).

[25] The main iterative procedure used is called RAS. This is a well-established technique for generating a matrix A which is consistent with target row and column totals (R and S respectively). Assuming consistent totals, the process involves summing the matrix across rows and columns in turn, comparing the totals with the targets, and then scaling to meet the targets. Typically, a solution is reached in just a few iterations. This simple two dimensional technique can be extended to cover multiple dimensions.

[26] BRES data are used by ONS to produce their published employment figures. The latter are used to constrain the *Working Futures* estimates. ONS revised their published estimates in the light of other information, so that figures used may gradually diverge from the original BRES estimates as official data are revised.

Analysis of these data suggest that only a handful of the industries in the *Working Futures* database are problematic. If the smaller industries are further aggregated to make just 75 industries rather than the 79 in the original *Working Futures* database, then no case (industry by region cell) would have fewer than 10 units. Such data should, therefore, not be disclosive.

This suggests that aggregation of those few industries into slightly broader categories mean that ***NONE*** of the *Working Futures* data would be disclosive.

Regarding confidentiality, since the *Working Futures* estimates are based on publically available data, there is no danger of the data breaching confidentiality from a LFS perspective. The data on individuals are not used directly. There are so many adjustments and process involved that none of the original data are in fact released into the public domain.

ONS were requested to confirm these interpretations:

- That employment estimates by aggregated sectoral categories by region (by combining them with other categories) would NOT be disclosive; and
- Combining this information with data from the publically available LFS dataset in order to generate breaks by occupation and qualification will not breach rules regarding confidentiality.

A.4.5 Case for ONS to place more detailed data into the public domain

At present, many of the more detailed data used to construct the *Working Futures* database are only available via NOMIS.[27] It would be helpful if ONS could place most of the information currently collected in order to construct the *Working Futures* database via NOMIS into the public domain. That would mean that the *Working Futures* database (possibly excluding sub-regional analysis) could be based solely on publically available data and would not, therefore, be disclosive.

If the *Working Futures* database were redesigned to be dependent only upon data in the public domain this would remove the need to impose any restrictions.

In the short-term, this can be achieved by censoring the data so that no disclosive information is revealed (by aggregating the sectors highlighted).

In the longer-term, it is proposed that ONS release more detailed data into the public domain, so that information currently only available through NOIMIS, and subject to a CEN, would be publically available.

ONS are requested to consider the case for releasing data at a more detailed level into the public domain (at the level of the 75 industries aggregated up from 79). This only requires a modest increase in the level of detail currently made available.

[27] These data are therefore obtained subject to possession of a CEN and which cannot be passed on to a third party.

A.5 Other indicators included in the LMI for all database

This section briefly describes the other datasets/indicators included in Phase 2A.

A.5.1 Unemployment (LFS)

The unemployment rate in an occupation is a key indicator from a careers guidance perspective providing information on the likelihood of securing employment. Various sources provide information on unemployment including the Census of Population and the official series on claimant unemployment made available on NOMIS.

Only one source offers the possibility of developing a consistent time series on the unemployment rate by detailed occupation classified using SOC2010. This is the LFS. This adopts the standard ILO definition for unemployment rate (those unemployed and actively searching for work expressed as a percentage of the economically active workforce).

The data available are only classified on a SOC2010 basis from 2011 onwards, but data on the old SOC200 basis are available for earlier years.

A.5.2 Vacancies (UKCES ESS)

The detailed UKCES Employer Skills Survey (ESS) collects information on skill deficiencies, including vacancies. It is a sample survey covering some 80,000 establishments. The information is normally published up to the 2-digit level of SOC2010, but the survey company have made more detailed information available at a 4-digit level

The survey does not cover all vacancies at this level of detail. Information is collected for up to six occupations per establishment. Unfortunately, the survey does not collect data on the numbers employed in each occupation. Therefore, the indicators that are possible to generate are limited to the number of vacancies, hard-to-fill and skill shortage vacancies, plus the percentage of total vacancies which are hard-to-fill and skill shortage within each occupation.

The survey is intended to produce estimates of the total number of vacancies, hard-to-fill vacancies and skill shortage vacancies in the UK from this large sample of establishments. This is achieved by multiplying the results of a survey by a weight derived from the ratio of the number of establishments in the survey to the total number of establishments in the UK. The dataset includes the weighted and unweighted number of establishments upon which each value in the dataset is based. Vacancy counts from the survey have been multiplied by the survey's employment weight in order to provide an estimate of the total number of vacancies of this type in the UK or region. The most detailed geographical breakdown available is to regions in England and the other nations of the UK: Wales; Scotland; and Northern Ireland. The time period covered is 2011. The ESS has been conducted on a similar basis roughly every two years. Results from the 2013 survey are expected to be available at the end of 2013. This is the first ESS to cover the entire UK and the first to use the SOC2010 classification.

The dataset can be queried on the occupation or industry code, and returns a set of the vacancies for this occupation, and how many of those vacancies are hard to fill or have skills shortages.

The Employer Skills Survey is a sample survey. Estimates based on an unweighted cell count of less than 50 should not be reported.

A.5.3 Occupational descriptions (ONS)

ONS have prepared detailed job description for each occupation distinguished in SOC2010. These go to the 4-digit level. This textual information has been added to the LMI for All database.

The following three text boxes provide examples of the kind of information available for sub major group 1.1 (2-digit level) with information for a selection of two 4-digit level categories (1115 and 1116, referred to as unit groups here). Similar information is available for each of the other unit groups (4-digit categories).

SUB-MAJOR GROUP 11

CORPORATE MANAGERS AND DIRECTORS

Job holders in this sub-major group formulate government policy; direct the operations of major organisations, local government, government departments and special interest organisations; organise and direct production, processing, maintenance and construction operations in industry; formulate, implement and advise on specialist functional activities within organisations; direct the operations of branches of financial institutions; organise and co-ordinate the transportation of passengers, the storage and distribution of freight, and the sale of goods; direct the operations of the emergency services, revenue and customs, the prison service and the armed forces; and co-ordinate the provision of health and social services.

MINOR GROUP 111
CHIEF EXECUTIVES AND SENIOR OFFICIALS

Jobholders in this minor group plan, organise and direct the operations of large companies and organisations and of special interest organisations; direct government departments and local authorities; and formulate national and local government policy.

Occupations in this minor group are classified into the following unit groups:

1115	CHIEF EXECUTIVES AND SENIOR OFFICIALS
1116	ELECTED OFFICERS AND REPRESENTATIVES

1115 CHIEF EXECUTIVES AND SENIOR OFFICIALS

This unit group includes those who head large enterprises and organisations. They plan, direct and co-ordinate, with directors and managers, the resources necessary for the various functions and specialist activities of these enterprises and organisations. The chief executives of hospitals will be classified in this unit group. Senior officials in national government direct the operations of government departments. Senior officials in local government participate in the implementation of local government policies and ensure that legal, statutory and other provisions concerning the running of a local authority are observed. Senior officials of special interest organisations ensure that legal, statutory and other regulations concerning the running of trade associations, employers' associations, learned societies, trades unions, charitable organisations and similar bodies are observed. Chief executives and senior officials also act as representatives of the organisations concerned for the purposes of high level consultation and negotiation.

TYPICAL ENTRY ROUTES AND ASSOCIATED QUALIFICATIONS

Entry may be by appointment or internal promotion, as appropriate, and is usually based on relevant experience although candidates may also require academic qualifications for some posts.

TASKS

- analyses economic, social, legal and other data, and plans, formulates and directs at strategic level the operation of a company or organisation;
- consults with subordinates to formulate, implement and review company/organisation policy, authorises funding for policy implementation programmes and institutes reporting, auditing and control systems;
- prepares, or arranges for the preparation of, reports, budgets, forecasts or other information;
- plans and controls the allocation of resources and the selection of senior staff;
- evaluates government/local authority departmental activities, discusses problems with government/local authority officials and administrators and formulates departmental policy;
- negotiates and monitors contracted out services provided to the local authority by the private sector;
- studies and acts upon any legislation that may affect the local authority;
- stimulates public interest by providing publicity, giving lectures and interviews and organising appeals for a variety of causes;
- directs or undertakes the preparation, publication and dissemination of reports and other information of interest to members and other interested parties.

RELATED JOB TITLES

Chief executive
Chief medical officer
Civil servant (grade 5 & above)
Vice President

1116 ELECTED OFFICERS AND REPRESENTATIVES

Elected representatives in national government formulate and ratify legislation and government policy, act as elected representatives in Parliament, European Parliament, Regional Parliaments or Assemblies, and as representatives of the government and its executive. Elected officers in local government act as representatives in the local authority and participate in the formulation, ratification and implementation of local government policies.

TYPICAL ENTRY ROUTES AND ASSOCIATED QUALIFICATIONS

Entry is by election.

TASKS

- represents constituency within the legislature and advises and assists constituents on a variety of issues;
- acts as a Party representative within the constituency;
- participates in debates and votes on legislative and other matters;
- holds positions on parliamentary or local government committees;
- tables questions to ministers and introduces proposals for government action;
- recommends or reviews potential policy or legislative change, and offers advice and opinions on current policy;
- advises on the interpretation and implementation of policy decisions, acts and regulations;
- studies and acts upon any legislation that may affect the local authority.

RELATED JOB TITLES

Councillor (local government)
Member of Parliament

A.5.4 Other possible indicators

- **Skills and abilities (O*NET Skills data)**

  This source is covered in a separate annex (Annex C).

- **Vacancies (DWP/Monster)**

  In principle, this is a key dataset from a careers guidance perspective. Detailed information on the number of jobs available classified by occupation is a crucial element. This used to be available via DWP as Jobcentre Plus vacancies (see discussion in Annex C.4).

  The Monster contract with DWP does include a specification for LMI, which ‘needs to be displayed in an intuitive and logical way so the general public can understand what is happening to the labour market nationally, regionally and locally’. This includes use of SIC and SOC codes and geography, though Universal Jobmatch does not seem to follow standard statistical definitions at present; (and this lack of standardisation has been the subject of debate on the Labour Market Statistics User Group). This lack of standardisation also applies to other dimensions such as geography. Regional options in England that Universal Jobmatch offers to employers posting jobs include ‘Anglia/ Home Counties/ Midlands/ North West/ London/ South East & Southern/ South West/ Tyne-Tees/ Yorkshire’. These do not match statistical regions.

The technical team can attempt a fuzzy matching for this dataset as a stopgap solution in part of Phase 2B, by which time DWP/Monster may have resolved these matters.

A.6 Longer-term issues relating to employment, pay and hours

In the longer-term, it would be better if the predicted estimates used for the three key indicators in the database, employment, pay and hours, could be replaced by survey data, which could be updated automatically as they are published. This raises two questions:

- ❖ If and when it will ever be possible to replace at least some of the predicted / estimated values for some indicators by 'real' survey values; and
- ❖ Checks on the reliability robustness of some of the more detailed predictions/estimates.

In principle, it is possible to use 'real' survey values where these are statistically robust and non-disclosive and to only use predicted values to fill in the many gaps. In practice, this might pose some problems, if and when the predicted values and real values show significant divergence. This is something that can be explored in more detail in Phase 2B. This will require further detailed consultation with ONS and the development of an agreed methodology for merging 'real' and predicted values in a seamless fashion.

In the short-term, the database uses predicted values throughout. However, a checking algorithm is built-in to the API to avoid 'publishing' estimates that might be regarded as unreliable. This algorithm checks whether or not the employment numbers would be disclosive or not statistically robust. The use of the slightly more aggregate 75 industry categories avoids the immediate issue of disclosure, since ONS have agreed that data at that level are not disclosive.

However, some of the numbers could still be unreliable because they are based on small sample numbers. In the *Working Futures* database, this is dealt with by adopting some simple rules of thumb and the same applies in the LMI for All database.

The rules of thumb used are:

1. If the numbers employed in a particular category / cell (defined by the 12 regions, gender, status, occupation, qualification and industry (75 categories)) are below 1,000 then a query should return 'no reliable data available' and offer to go up a level of aggregation across one or more of the main dimensions (e.g. UK rather than region, some aggregation of industries rather than the 75 level, or SOC 2-digit rather than 4-digit).
2. If the numbers employed in a particular category / cell (defined as in 1.) are between 1,000 and 10,000 then a query should return the number, but with a flag to say that this estimate is based on a relatively small sample size and if the user requires more robust estimates they should go up a level of aggregation across one or more of the main dimensions (as in 1).

This is done not only for any queries about Employment (including Replacement Demand calculations), but also for Pay and Hours.

In the case of Pay and Hours, the API interrogates the part of the database holding the employment numbers to do the checks, as in points 1 and 2 above, but then reports the corresponding Pay or Hours values as appropriate.

Currently, data are provided at the most detailed level possible for all three indicators. More aggregate estimates are obtained by simple summation (for employment) or by creating weighted averages (using the employment numbers as weights).

A.7 Details of the regression analysis for pay predictions

A.7.1 Introduction

This section provides a general description on issues that need to be clarified for the specification and estimation of wage functions using ASHE and LFS data on Pay. It also introduces the data sources, definition of the variables included and methods used in the estimation. Details about how the estimation results are used to predict wages and caveats that need to be borne in mind when using and interpreting the outputs are also provided.

The data used for the current wage analysis is from the UK Labour Force Survey (LFS). This is due to the unavailability of the 2012 Annual Survey of Hours and Earnings (ASHE) data to researchers using the 2010 Standard Occupational Classification (SOC2010). All the estimation so far has been carried out based on UK LFS using SOC2010. Once the 2012 ASHE data becomes available in the UK Secure Data Service (SDS), the same approach as used in LFS (and described in detail below) will be applied to the ASHE data

The discussion here does not attempt a detailed explanation of the estimation outcomes, but aims to provide some notes to help the reader understand how the analysis has been conducted and when care is needed in using or interpreting some of the results. It is structured in 6 sections. Section A.7.1 gives a brief introduction to the study. Section A.7.2 explains how the LFS database is constructed and introduces the definition of wages and other variables used in the analysis. Section A.7.3 discusses the specification of the wage functions and how the estimated results are used for predicting wages. Section A.7.4 explains some supplementary equations focusing on mean pay which are used to generate prediction by age ‘on the fly’ in the LMI for ALL API. Section A.7.5 introduces ASHE and outlines its advantages and limitations compared to UK LFS. Section A.7.6 concludes.

A.7.2 Data and definitions

The dataset used for the current wage analysis is a pooled sample from the UK Labour Force Survey (LFS). The LFS is a quarterly survey which collects information from households living at private addresses and is representative of the entire population of the UK. Each quarterly sample is made up of five waves with approximately the same sizes. Each wave is interviewed in five successive quarters. The sample is designed in a way that over the period of any four consecutive quarters, wave one and five will never contain the same households. Thus, for the construction of an annually representative sample of the population, wave one and wave five of each quarter in 2011 and 2012 are pooled together to form an aggregated sample of 288,937 different individuals covering two years. For the purpose of this wage study, the pooled sample is further constrained to full-time employees aged 16 and over, leaving 87,830 individuals in the core research sample.

Gross weekly pay is used in all equations and it is defined as: usual gross weekly pay before any deductions, which includes overtime and any other possible income sources. Information on components of gross wage and the contribution of each component is not available in LFS.

The current study focuses on mean (average) pay as opposed to the median or deciles (these may be introduced in a later version of the LMI for All database). The way the current estimates are constrained to match published headline numbers relies on an algorithm that scales the sum of detailed categories employment numbers multiplied by the corresponding mean wage to match the published aggregate figures. The same relationship does not apply for medians (nor for deciles). Some alternative approach will be necessary to generate the measures of pay distribution. One possibility would be to apply an analogous approach to that used for variations by age (as set out in Section A.7.4 below). This would involve running a secondary equation that relates pay for selected percentiles to the mean value for the category concerned. This can be explored further in Phase 2B.

A.7.3 Earnings function

Again, this section does not attempt a detailed interpretation of the regression results, but explains what has been included in the wage equation and what/how things have been done.

The earnings function has been run using the log of gross weekly wage as the dependent variable. The independent variables included, and their definitions, are as follows:

- Age: a continuous variable ranged 16 to 84;
- Age squared: continuous variable;
- Gender: male and female, 1 dummy variable for male (base category: female)
- Region: 12 government official regions of England or devolved countries within the UK, 11 dummy variables in the regression (base category: London)
- Highest qualification: two variables with different qualification classifications have been created and separate earning functions have been run with each of the categorisations. One has 9 categories (QCF1-8 and no qualification) and 8 dummy variables in the regression (base category: QCF8); and the other one has 6 categories (NQF1-5 and no qualification) and 5 dummy variables in the regression (base category: NQF5) ;
- Industry: standard 75 categories as used in *Working Futures*, 74 dummy variables in the regression (base category: Agriculture, etc.);
- Occupation: 4-digit SOC2010, 369 categories and 368 dummy variables in the regression (base category: 115 Chief executives and senior officials).

Interactive terms have also been included to detect heterogeneity across different groups:

- Gender by occupation: gender is interacted with 4-digit occupation categories to control wage differences between male and female within each occupation. The base group is female Chief executives and senior officials.
- Industry by time trend: a time trend variable is created for 2011 and 2012. It is interacted with industries to control time trend differences within each industry. The base groups are industries in 2011.
- Occupation by time trend: the time trend is also interacted with occupations to control time trend differences within each occupation. The base groups are occupations in 2011.

A linear earnings function with a quadratic term for age indicating changes of age effect on wage is estimated using the ordinary least square method. The estimated coefficients of the independent variables and the constant term can be used to derive the expected wage for an individual with certain characteristics (as defined by the variables included). For the earnings function specified in this study, the default reference group is female workers living in London with highest qualification QCF8 (or NQF5) working in the Agriculture sector and are Chief executives or senior officials in 2011. The log expected wage for an individual with these default characteristics at a certain age can be calculated by adding the following parts together: coefficient on age times age; coefficient on age square times age square; plus the coefficient for the constant term. The calculation of log expected wage for people with other characteristics can simply be made by adding coefficients for relevant dummy variables and interaction terms to this default log expected wage. For example, for a male worker with all the other same characteristics as default, his log expected wage is the default log expected wage plus the estimated coefficient of the male dummy. To obtain the expected wage, the log numbers need to be converted back to wage following: EXP(log expected wage).

Given a regression function like this, it still leaves the question of how to provide the information for individuals whose combination of characteristics are not reflected in the dataset. This is because the expected wages derived from the estimated coefficients in the regression package are based on taking the fitted values for each individual in the regression, so it is not possible to produce expected wage where there is no sample numbers in a particular cell. This is, therefore, done outside the Stata regression package used to estimate the parameters.

A.7.4 Supplementary age equations

In order to produce predictions of pay by age in the LMI for All database a supplementary equation is used which enables calculation of variations of pay by age 'on the fly'.[28] These supplementary age equations reflect how age affects the deviation of pay from the mean pay in groups with different combination of characteristics. The factors used in the supplementary equations only include age, age squared, and log mean pay of a certain combination (or log wage of a person at mean age in a certain combination). The coefficients vary across occupation (at the 4-digit level).[29]

The combinations are defined by a set of dimensions including: gender (male and female), 12 regions, 75 industries, 4-digit occupations, 9 or 6 categories of highest qualification, and two time periods. In the case of 9 highest qualification categories, there are in total 62,844

[28] This was to avoid too large a data file of predicted pay being used in the API which caused some problems of access speed, as well as allowing the mean pay predictions to be constrained to match published pay totals using an iterative process. The latter requires information on the numbers of people in each category which was not available for individual age categories.

[29] The supplementary age equation only includes interactions between age, age squared and 4-digit occupations. Gender is not interacted with age and age squared. In order to run the age equation, mean pay of each combination (defined by gender, region, industry, occupation and year) is identified first, and takes into account the gender aspect. So by interacting age and age squared with occupation, the focus is on what the age profile looks like for people with all the other same characteristics (including gender, region, industry, occupation and year).

different combinations, and there are 59,701 combinations for 6 highest qualification categories (these two numbers both exclude combinations, which do not have any observations and only counts combinations with observations).

To generate the mean pay of each combination, mean age is first produced for all the combinations and then plugged back into the main earning equations (as described in Annex A.7.3) to calculate the expected mean pay of each combination using the estimated coefficients produced previously. Supplementary equations for 9 and 6 qualification categories are conducted separately by regressing the same log gross weekly pay as used in the earning equations on age, age squared, log mean pay and interaction terms for age and 4-digit occupations and age squared and 4-digit occupations. These differentiate the effects of age and age squared on wages in different occupations.

The estimated coefficients from the supplementary equations enable the estimation of pay by age 'on the fly' from mean values for all ages. The coefficients on the interaction terms reflect how the age parameters differ across occupations. These supplementary age equations reflect how age affects the deviation of pay from the mean pay in groups with different combination of characteristics. Predicted pay by age can be generated from these new coefficients and the mean pay for the group concerned. The default group in the supplementary age regressions includes workers in the occupation of Chief executives and senior officials. Log expected wage for any individual in this occupation group can be derived by adding together the coefficient of log mean pay times mean pay, the coefficient of age times age, the coefficient of age squared times age squared and the coefficient on the constant term. Again, to get the expected wage, the log of the expected wage needs to be unlogged. To obtain the log expected wage for individuals in other occupations, the coefficient of the relevant interaction terms need be added to the default log expected wage.

In the original version of the dataset supplied to the Technical Team, pay predictions are made for each single year age category, resulting in a huge data input file. In the revised version information on variation by age for different occupational groups for the database is be based on parametric methods and computed in the API 'on the fly'. The cut down version, with just pay for the mean age in the database is the preferred option.

The factors used in the supplementary equations only include age, age squared, and log mean pay of a certain combination (or log wage of a person at mean age in a certain combination). The coefficients vary across occupation (at the 4-digit level).

The functional form for the age equation is:

$$\text{Log (pay)} = a + b^*\text{Log (mean pay)} + c\ {}^*\text{Age} + d^*\ \text{Age}^2 + e^*\text{Age}^*\text{occupations} + f^*\text{Age}^2{}^*\text{occupations} + u$$

Where pay is the individual gross weekly pay as used in the wage equation, mean pay is the average pay in each combination defined by gender, region, industry, occupation and year.
To generate the mean pay of each combination, mean age is first produced for all the combinations and then plugged back into the main earning equations estimated before. The expected mean pay of each combination can thus be calculated using the estimated

coefficients produced in the earning equations. Supplementary equations for 9 and 6 qualification categories are conducted separately by regressing the same log gross weekly pay as used in the earning equations on age, age squared, log mean pay and interaction terms for age and 4-digit occupations and age squared and 4-digit occupations. The two interactive terms can differentiate the effects of age and age squared on wages in different occupations.

u is the error term and the estimated coefficients of a, b, c, d, e, and f can be subsequently used to estimate gross weekly pay by age on the fly from mean values for all ages considering differences between 4-digit occupations.

The 2-digit and 4-digit occupation codes have been experimented with for the purpose of finding the most suitable occupational variable. For the reasons that it does not make much difference in the explanation power of the age equations and for the consistency with the earning equations, 4-digit occupation codes have been adopted to interact with age in the supplementary analysis.

A.7.5 ASHE and comparison with LFS

When the data for ASHE 2012 becomes available, the same earning functions and supplementary equations will be conducted using ASHE. ASHE has many advantages compared to LFS. The most recognised advantage in ASHE is that it provides administrative data with wage information collected directly from company records. It is regarded as more reliable and accurate compared to wage information in LFS. The following sections introduce ASHE and provide a more detailed comparison with LFS.

ASHE originated from the New Earnings Survey (NES) which was started in 1970 and carried out each year subsequently. It is the most comprehensive sources of earnings information on the structure and distribution of earnings in the UK. It collects data on level of wages, wage components, paid hours of work, pension arrangements and other job characteristics from all employee jobs (self-employed workers are not included in ASHE) in all industries and occupations across the whole of the UK. The samples are designed to select all employees whose National Insurance Number ends in a particular pair of digits and has a sample size around 180,000 employees in the UK. The selected sample covers about one per cent of the whole working population in the UK. However, the ASHE data are only available to researchers at Great Britain level (data for Northern Ireland have not been released by the Department of Enterprise Trade and Investment Northern Ireland). Therefore, the estimates from ASHE will only apply to Great Britain. The industry and occupation classifications have changed over the ASHE waves. From 1997 to 2001, SOC1990 was used to classify occupations, from 2002 to 2011, SOC2000 was used, and from 2012 onwards, SOC2010. For the standard industry classification, SIC2003 was used for the period of 1997-2007, and SIC2007 for 2008-2011.

Compared to LFS, which is the other major source of information for labour market and wage studies in the UK, ASHE has the advantage of having a larger sample size. It is also considered to have more accurate wage and working hour measures, since these are based on data provided by employers based on pay records. The actual amounts of wages that have been paid to employees and the number of hours that the employees have been paid for are collected. In contrast, for LFS, information is collected from households. Either employees or other representatives from the employee's household provide the wage information. This can lead to proxy and estimated responses. Moreover, the LFS collects information on hours worked rather than hours paid which can increase the deviation between estimates and real hourly wages. ASHE also provides industry and occupation classification from a business perspective which is regarded as more accurate than the classification provided by individuals in LFS (self-assessed).

However, ASHE is not without its own limitations. The biggest disadvantage of ASHE, comparing to LFS, is its limitation in providing information on individual characteristics of employees. Due to the fact that ASHE collects information from employers, information which is not held by the employer, such as education, ethnicity, household composition, health, etc., is not available. Thus it is not possible to control for education/qualification when explaining wage differentials in ASHE. Variables that are available in ASHE include breakdowns of wages by age, gender, full time/part time, region, industry and occupation classifications. Another limitation is linked with the fact that ASHE is conducted in April, thus it does not fully cover seasonal work (which is only undertaken in other seasons, such as summer or winter work).

In the absence of the ASHE 2012 data, LFS regressions without the qualification variable have been conducted to match the regressions in ASHE. Parameters have been set out in the same way as if we were using ASHE so that the numbers can be revised quite quickly once the ASHE 2012 dataset using SOC2010 becomes available in the SDS.

A.7.6 Concluding remarks on pay predictions

The section has set out various issues that need to be borne in mind when using the estimated results from the wage functions and supplementary age regressions. Details on how the research sample has been generated, what variables have been included and how they are defined are explained. The current results are based on the UK LFS. The same methods and analysis will be applied using suitable ASHE data if and when it becomes available. ASHE has a number of advantages compared to LFS. However, ASHE does not provide any information on education, thus it will not be possible to include the same highest qualification variable as in LFS. The estimated coefficients derived for other variables using ASHE will very likely be overestimated because they are taking account of education effects (omitted variable bias). The estimates from ASHE will therefore not be fully comparable with those from the LFS. This could be seen as an argument for just relying upon the LFS for the regression analysis. However, the larger sample size in ASHE, and the more reliable data from employer records, outweighs such considerations.

A.8 Technical details of the algorithms used to constrain the data to match official estimates of pay and hours

A.8.1 Introduction

Key elements of the data requirement set out in the project plan are pay, hours and employment. Ideally the aim is to break these down into as much detail as possible by:

The full set of detail ideally required is therefore as follows:

- Occupation (up to the 4-digit level of SOC 2010, 369 Categories);
- Sector (up to the 2-digit level of SIC2007, 75 categories); and
- Geographical area (12 English regions and constituent countries of the UK);

Plus

- Age;
- Gender;
- Status; and
- Qualification (where available).

The original idea was to access these data directly from the original survey sources, but this poses various problems of confidentiality and disclosure if information is to be made available at the levels of detail that would be really useful for a careers database. These problems are exacerbated when the additional dimensions such as gender, employment status (full-time, part time, self-employment), age and qualification are added or when additional granularity is demanded in key dimensions such as sector or occupation. The data have, therefore, been estimated using data from *Working Futures* and using econometric analysis (earning functions).

This section sets out details for the algorithm used to constrain the data to match official totals. This is based on the well-established RAS process.[30] RAS procedures have been developed to generate detailed data on Pay, Employment and Hours consistent with published data from official sources.

A.8.2 RAS processes

There are three main elements to the database that require RASing to make sure the data agree with published figures. These relate to employment, pay and hours

[30] RAS is an iterative procedure where the rows and columns of preliminary estimates of a two dimensional array are iteratively changed using proportions that are based on the 'target' row and column totals. The basic RAS technique relates to a two dimensional matrix, but can be extended in to n dimensional arrays. For some references see: McMenamin and Haring (2006); Miller and Blair (2009); and Toh (1998).

Employment

Employment data already exist at the 2-digit level in the published *Working Futures* (WF) database (See Wilson and Homenidou, 2012a, 2012b). This dataset has been expanded from the current 25 2-digit occupations in the WF dataset to 369 4-digit categories for LMI for All. In the first instance, this is done using a simple assumption of fixed and constant shares of employment of the 369 categories within each of the 25 digit ones, based on LFS data. The focus is on *25 sets of shares* (each summing to 100 per cent) showing the proportions of employment in 4-digit categories within each 2-digit category. In principle, this analysis could be extended to allow these shares to vary by other dimensions, such as industry, but this will need to be undertaken in Phase 2B.

In the longer-term, it is also necessary to think about how these patterns change over time and how to extend the projections to 2020 and beyond, but for the moment these shares are constant, based on 2011/2012 LFS data (for further discussion see Annex C.6).

The main steps are as follows:

1. Interrogate the LFS and extract the sets of shares of 4-digit occupations within 2-digit categories:
    a. Across the whole of the UK;
    b. Showing variations by ‘region’ (12 countries and English Regions);
    c. Variations by Type (FT, PT, SE) and gender;
    d. Variations by Sector (*Working Futures* 6 broad sectors.

    There are just two years of LFS data available classified using SOC 2010. These have been combined for this purpose, avoiding double counting of individual cases in the standard manner.

    To begin with the data are extracted in the form of *numbers* in employment at the most detailed level required (369 occupations, 75 industries, 12 countries/regions and 6 types). This information is then aggregated to create the sub-totals in (a) - (d) above by simple summation. The shares of occupational employment in 4-digit categories within 2-digit categories can then be computed.

2. Using this information a full and consistent set of shares that covers the full WF database is then developed:
    a. Occupation (25);
    b. Region (12);
    c. Industry (79);
    d. Type (6);
    e. Qualification (9).

3. The final set of shares are applied to all years of the WF database.

This requires a RAS process to ensure the overall UK patterns at the 369 level are still satisfied, and some of the subtotals too, as well as maintaining all the existing WF employment structure.

For many of the cells in the data array created there will be only tiny numbers of people involved (many are empty). The information in such cells cannot be regarded as statistically robust but it is not possible to quantify this by estimating precise confidence intervals. Instead "rules of thumb" based on ONS general guidelines for use of LFS data are adopted.

1. If the numbers employed in a particular category/cell (defined by the countries/regions, gender, status, occupation, qualification and industry) are below 1,000, then a query returns 'no reliable data available' and offers to go up a level of aggregation across one or more of the main dimensions (e.g. UK rather than region, some aggregation of industries rather than the most detailed level, or SOC 2-digit rather than 4-digit).
2. If the numbers employed in a particular category/cell (defined as in (1)) are between 1,000 and 10,000 then a query returns the number but with a flag to say that this estimate is based on a relatively small sample size and if the user requires more robust estimates they should go up a level of aggregation across one or more of the main dimensions (as in 1).

Pay

The second element is the corresponding pay database. This is based on a combination of ASHE and LFS data. Various checks and adjustments are made to ensure it is consistent with published data. This involves the following steps:

1. Published ASHE pay data are extracted from the ONS website, using common definitions (including overtime). These relate to the main dimensions of the database:
    a. Occupation (SOC2010 2-digit (25) and 4-digit (369) categories, summed over all other dimensions);
    b. Industry (standard *Working Futures* 6/22/75 categories, see Annex A.9);

    In each case these are selected by Type (4 of the 6 (SE not available) and Region (the 12 countries and English regions that make up the UK).

    In the long-term, the aim is create a consistent time series, but changes in occupational classification and also industry classification limit how far back it is possible to go. At present the focus is on just a single year (2012).
2. These data form 'targets' to be used to constrain the much more detailed data generated from the regression analysis
3. The data generated in 1 - 2 are used to create a wages and salaries database (PAY *EMPLOYMENT) = that can be used as a suitable set of constraints/ targets for the RAS process
4. The ASHE dataset does not include information on Qualification (6/9). This is

obtained from the LFS, constrained to be consistent with the ASHE data.

5. A new custom written programme has been developed to constrain the existing LMI for All database created from the regression analysis to match this set of targets using RAS methods.
6. Note that the initial set of pay predictions from the earnings equations also vary by age, whereas this dimension is NOT available in the employment database (the LFS and other data sources are simply not large enough to supply a detailed age breakdown as well as all the other dimensions of interest).
7. The constraints are, therefore, imposed across all ages (summing up across all age groups).
8. Age is an important dimension for the LMI for All database, so this is dealt with using a supplementary procedure which recognises how pay varies by age across occupational categories (see Annex A.7).

Hours

The third element that requires a RAS process is Hours. This is currently based on ASHE data (although the LFS could also be used).

Published ASHE data on hours are extracted from the official sources, covering all the main dimensions of the database to form ‘targets’:

a. Occupation (SOC2010 2-digit (25) and 4-digit (369) categories, summed over all other dimensions.
b. Industry (*Working Futures* 6/22/75 categories where available).
c. Country/region (12)
d. Type (4 of the 6 (SE not available)

As for pay, the long term aim is to produce time series. Because of changes in classification, etc., this is difficult. The current focus is just on 2012.

The aim is to have values of typical weekly hours for the fully detailed dimensions of the database, but with repetitions (defaults to higher levels of aggregation) where the data are weak (especially at the SOC 4-digit level). In part, this depends on how much variation in hours there is within the SOC 2-digit categories.

The detailed data are generated using the non-parametric procedure described in Annex A.8.3. There is no obvious equivalent to the earnings equation for pay, although a simple equation can be estimated that shows how hours vary across all the main dimensions. In principle, it is possible to replace the current estimates by data based on an equation analogous to that used for pay. As discussed in Section A.3 above, it has not been possible to estimate such an equation on ASHE data classified using SOC2010, as these data are not yet available in the SDS. Therefore, the non-parametric approach set out in Section A.8.3 below has been retained

The detailed data are generated by using multiplicative ratios of the differentials applied successively covering all the dimensions – region, gender, status, industry and occupation.

1. The starting point is an average weekly hours figure from the ASHE dataset.
2. The differential factors for a particular dimension are also based on average hours worked per week from ASHE (aggregated across all other dimensions).
3. The process starts with the average hours for occupations and multiplies each 'cell' by appropriate industry (and other) differentials in turn to ‘fill in the gaps’.
4. These data are the converted to total hours worked by multiplying by employment and then RAS’d to get a consistent set of total hour figures.
5. Average hours are then calculated by dividing total hours by employment.

This is roughly equivalent to running a regression similar to the one for earnings but:

- Linear rather than log-linear;
- No age variable (age or age squared) is included.

Although it is not possible to run this type of equation on ASHE data for SOC2010 categories at present (because the data have yet to be set up in the SDS), it can be done using the LFS.

A working hours equation

The results of regressions (using LFS data) of an analogous form to that used for Pay, with a full set of dummies and interactive terms as for pay, but EXCLUDING age and age squared suggest that such a methodology could deliver robust estimates once the data are available in SDS. A linear regression for working hours was estimated using LFS data to explore how various factors influence an employee’s hours worked per week.

The sample includes all working people – including full-time and part-time workers, but is constrained to employees only to match with ASHE. The dependent variable is actual hours of work per week for the main job, including overtime. The independent variables are the same as the ones in the wage equation excluding age and age squared.

The goodness of fit measures R squared for the working hours regressions are around 20 per cent compared to 50 per cent for the wage equations. This is probably due to that the more constrained sample to full-time workers only in the wage equations reduces the possible biases arising from inaccurate measures of weekly wage from part-time workers.

In the working hours equation, gender does not have any significant effect indicating men and women tend to work same hours per week given other characteristics the same. While in the wage equation, men are significantly earning more gross weekly wage than women. Regions and qualifications continued to be significant in the working hours equation with people living in London and people with higher qualification significantly working more hours per week than others. Differences in working hours between industries and occupations are mixed.

A.8.3 RAS processes – Generating pay and hours ‘targets’ from official data

The basic data are taken form the ONS website which publishes headline figures for all the main totals (by region, industry, occupation, etc.).

An initial step is needed to ensure the targets for industry, occupation and qualification are consistent. This requires a scaling of the three sets of wage bill targets (average wage * employment), by industry, by 2 or 4-digit occupations and by qualifications, to match the overall wage bill for all categories (and analogously for total hours worked (hours * employment). For this purpose employment is based on the *Working Futures* estimates.

It also is necessary to fill some of the gaps in the more detailed breaks used as ‘targets’.

These targets (for wage bills or total hours) can be generated as follows (the example shown is for wage bill and wage/ pay rates):

Wage ***rate*** for the industry (or occupation) * regional differential * gender/status differential)* relevant employment number, where:

- Regional (r) differential = wage rate (r) / wage rate for all regions;
- Gender (g) / status (s) differential = wage rate (g/s) / wage rate for all gender status categories.

Note that if employment is zero the wage bill (or total hours) will therefore be zero.

Note also that in the main LMI for All database pay is generated only for full-time employees at present. It also covers just a single year (2012).

Within the database, *weekly pay (excluding overtime)* * *employment* for the following dimensions (focusing on industry) are included:

1. Overall total (all gender-status, all industries, UK);
2. Totals for 4 gender-status (all industries, UK);
3. Totals for males and females separately (all industries, UK);
4. Totals by 12 regions (all gender-status, all industries);
5. Totals by 75 industries (all gender-status, UK);
6. Industry (75) by region(11) by gender-status (4).

The same output is repeated for hours worked.

For occupations, there are the following outputs for total hours worked:

1. Overall total (all gender-status, all occupations, UK);
2. Totals for 4 gender-status (all occupations, UK);
3. Totals for males and females separately (occupations, UK);

4. Totals by 12 regions (all gender-status, all occupations);
5. Totals by 25 SMG occupations (all gender-status, UK);
6. Totals by 369 4-digit occupations (all gender-status, UK);
7. 25 SMG occupations by region(11) by gender-status(4);
8. 369 4-digit occupations by region(11) by gender-status(4).

A.8.4 Main Steps (for Pay)

1. The starting point is:
    a. A detailed database (based on the econometric analysis of LFS data[31]) showing predicted pay across the various dimensions (sector, occupation, gender, region, qualification, status and age).
    b. Summary data on pay 'targets' for the main dimensions, based on published statistics from ASHE (and the LFS for qualifications).
2. These are combined together to ensure that 1a is consistent with 1b.
3. This is achieved by using a RAS process to ensure that wages & salaries (pay*employment) are consistent in the same way as is already done for employment in the *Working Futures* employment database (in this case i.e. the RAS targets are set in terms of wages & salaries rather than employment).
4. The process is more complex than that for employment for a number of reasons:
    a. The pay/wages & salaries dataset covers age, which is not a feature for employment;
    b. There are two alternative sources of data on pay (ASHE and LFS) that need to be reconciled. UKCES were very keen that the data should be consistent with ASHE, so this has been built in as a key feature. However, ASHE excludes qualification, so a second stage involving LFS data are also needed.
5. The main stages are as follows:
    a. Estimate the earnings equation (excluding qualifications) using ASHE data in the SDS;
    b. Using the extracted parameters from 5a. generate predicted pay for step 1.a. (note that this EXCLUDES the qualification dimension);
    c. Based on the data assembled in 1b, generate a set of consistent 'RAS targets' to constrain the dataset in 5b. These are based on mean pay data multiplied by relevant *Working Futures* employment numbers.

[31] In the long-term, this will be replaced (in the main) by analysis based on ASHE data carried out in the SDS. However, the SDS does not currently have ASHE data reclassified to SOC 2010, so LFS data have been used instead. Because ASHE does not include qualifications some part of the regression analysis will always be based on the LFS.

d. An adjustment to these wages & salaries estimates is needed to ensure that they add up consistently to the same totals when summed across different dimensions set out in 1a (excluding qualifications).

e. Using these targets the data in 5b can be RAS'd to be consistent with those from 5c. This required the development of a new programme analogous to the employment one (excluding qualifications).

f. A key issue is how to deal with age, which is not a dimension in the *Working Futures* database. In principle, this requires the *Working Futures* database to be extended to cover one year age groups. However, given that there is no reliable source of data to provide such information cross-classified by all the other dimensions simultaneously, and that even if such data were available the limits of the current Python programme have been reached it was decided to focus on 'all ages' for this step.

g. The alternative method developed ignores age in the RAS process and imposes a 'sensible' age profile for each category *Ex post.* (based on the use of a supplementary equation which allows parameters to vary by occupation). This allows a prediction pay 'on the fly' in the API based on the predicted mean pay from the main earnings equation and age.

h. The predicted pay at step 5 b is made for the average age for that particular combination of sector, occupation, gender, region, qualification and status.

6. These steps generate a pay database EXCLUDING qualifications. Further steps are needed to expand the dataset to include the qualification dimension. This relies on LFS data. The main stages are as follows:

   a. Estimate an extended version of the earnings equation (including qualifications) using the SDS version of the LFS;

   b. Using the extracted parameters from 6.a. generate an extended version of 5b and 5g (now including the qualification dimension);

   c. Based on analogous data to that assembled using ASHE, generate a set of 'RAS targets' based on LFS data to constrain the dataset in 6.b, including qualifications. These are based on mean pay data multiplied by relevant *Working Futures* employment numbers. As in step 5 it is necessary do some work on these wages & salaries estimates to ensure that they add up consistently to the same totals when summed across different dimensions set out in 1a (this time including qualifications). This requires a separate initial RAS process to ensure that the 'target' data remain consistent with these from step 5.

   d. Using these targets the data in 6.b can be RAS'd to be consistent with those from 6c. This requires a further extension to the new programme in step 5 (now including qualifications).

   e. As in step 5 a key issue is how to deal with age which is not a dimension in the *Working Futures* database. The solution adopted is analogous to that in steps 5f -h.

A.8.5 Technical details of data sources and methods for the RAS process (Pay and hours targets)

Files are downloaded from the ONS website and details of the programs created to read them and descriptions of the workbooks themselves are set out in the Notes sheets. The workbooks and associated programs read in the headline 'Hours' and 'Pay' data and then write this information out in a suitable form to act as the constraints for the RAS procedures, including generating "wage bills" (average wage x employment) and "total hours" (average hours x employment). The workbooks also include procedures for filling any gaps if required. The workbooks have a 'Notes' sheet giving an overview of the procedures adopted and relevant further information.

Creation of pay and hours targets for the LMI for All database

The immediate aim is to arrive at a set of 'targets' for a RAS process. The final aim is to ensure the LMI for All database is consistent with published data. Published ASHE data on pay was downloaded from the ONS website (weekly pay including overtime has been used). The main dimensions needed are:

a. Occupation (SOC2010 2-digit (25) and 3-digit (369) categories, summed over all other dimensions;
b. The same by Region (12);
c. The same by Industry (*Working Futures* 6/22/79 categories where available) – all levels are needed;
d. The same by Type (4 of the 6 (SE not available)).

The most suitable tables available for download contained pay by 369 (4-digit) occupations with 25 (2-digit) occupations interspersed as sub-totals.

UK and regional data are in the same table, but type (4 gender-status) are in separate tables also by occupation by region. Summary values for the UK and region appear at the head of each geographical area.

There is no industry by occupation breakdown, but pay by 88 (2-digit) industries with 21 industry levels interspersed as sub-totals is available. UK and regional data are in the same table, but type (4 gender-status) are in separate tables also by industry by region. Unlike the occupational data table, summary values for the UK and region appear at the head of each table grouped together.

Initially, 2012 year data have been used, but eventually the aim is to create a time series. However, changes in occupational classification and industry classification limit how far back it is possible to go. Using SOC2010 classification, only 2011 and 2012 are available. Before that SOC 2000 is used by ASHE. More years are available for SIC2007, which is available from 2008 onwards.

Data on hours worked was also downloaded from ASHE. These tables are laid out in the same way as for pay. Note that for Northern Ireland only overall values by type are present in the tables. There is no breakdown by industry or occupation. The following web link can be used to access downloads from ASHE: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-280149 Two Visual Basic (VB) based Excel programs were developed to read the tables to produce output by industry by region and type and similarly for occupation by region by type.

Each program can read either pay or hours as required. Complicated table layout made it necessary to search for the occupation or industry required rather than basing it on a fixed pattern layout. The advantage, however, is that the programs should cope with different years and minor table variations.

One complication addressed by the programs is the aggregation from 88 (2-digit) industries to 75 industries used by the database. This is performed in the program that deals with industries. The method is to multiply the mean pay (or hours) by the number of jobs surveyed, aggregate the results and then divide by the total number of jobs. This gives a new mean pay (or hours) for the aggregated industries. This only works if all required data are present in the table.

Creation of pay targets for 75 Industries, for both 2-digit and 4-digit occupations and for 6 and 9 levels of qualification

1a. Average pay levels in tables downloaded from ASHE are read by separate programs and the pay levels re-written to this workbook in a suitably rearranged way.

Pay levels by 6 and 9 qualifications levels from the LFS are produced in a similar format (6 and 9 levels by 4 gender-status by 12 regions).

Average pay by industry by gender-status by region and average pay by 2-digit occupation by gender-status by region are identified.

4-digit occupation by gender-status by region is also identified alongside average pay by qualification level by gender-status by region.

1b Where no values are given in the original ASHE tables the entry is set to zero.

2a The above programs also read employment levels by industry by gender-status by region (from the *Working Futures* database) and employment by 2-digit and 4-digit occupations for gender-status and region.

Employment levels with the LFS qualification data are also taken from *Working Futures* data.

2b Employment levels are written to the workbook.

2c Average pay and employment are multiplied together and written alongside previous output.

2d Summary pay averages and employment totals are also written. They are:

- ❖ Overall - All gender-status, all industries (or occupations), UK;
- ❖ Averages and totals for each gender-status category;
- ❖ Averages and totals for each gender;
- ❖ Averages and totals for full time and for part-time;
- ❖ Averages and totals for each region;
- ❖ Averages and totals for each industry (or occupation).

The summary values appear at the top of each worksheet.

3. At this stage some gaps remain where the tables contain no data. It might be to avoid disclosure or because the levels are either nil or negligible. An additional step has been added here to fill the gaps and the calculations are performed in the worksheets themselves. This is done as follows:

a. Differentials for each of the gender-status categories (wage rate for the category / wage rate for all categories) are calculated in the worksheets.

b. In the same way differentials for each of the regions are calculated as wage rate for the region / wage rate for all regions.

c. If there is no gap at the detailed 75 industry, 25 occupational or 369 occupational level then values are left unchanged. However, if there are gaps

Then, they are filled by using the formula:

Estimated wage bill = Wage rate for the industry (or occupation) * regional differential * gender/status differential * relevant employment number

The same method is used for industry and for occupations and qualifications. Note that if there is zero employment then this step returns a zero.

4. The resulting arrays from Step 3c are next scaled in two ways:

a. so that the sum of all industries (or occupations or qualifications), all types, all regions, agrees with the overall UK wage bill, (i.e. Using a single scaling ratio overall UK wage bill / sum of all types, all regions and all industries (or occupations),

b. so that the sum of all industries (or occupations or qualifications) and all types for the regions agrees with the each regional wage bill, (i.e. Using twelve scaling ratios of a regional wage bill / sum of all types, all industries (or occupations or qualifications) for the same region.

Ratios of 4a:4b were added temporarily for checking whether repetitive scaling, effectively a RAS process is necessary. Scaling of 4a and 4b have been done separately for 75 industries, 2-digit occupations, 4-digit occupations, 9 and 6 levels of qualifications.

Creation of targets of average hours worked for 75 Industries, for both 2-digit and 4-digit occupations.

1. Initially there are two sets of targets for 'Hours', one relating to Industries and the other Occupations.

   As for 'Pay' initial values for the 'Hours' sheets are written.

   Further 'Hours' calculations/adjustments and scaling are performed by links in the worksheets in a manner analogous to that described above for Pay so we have arrays for Hours for both 75 Industries and 2-digit and 4-digit Occupations. Both also have 12 region and 4 gender-status dimensions.

2. A Visual Basic macro to combine Hours for Occupations and Industries into one larger array has been written. The aim of this routine is to read the initial HOURS estimates for Industry and Occupations (for both 25 2-digit and 369 4-digit occupations separately) and combine them to produce 2 arrays of gender-status by region by occupation by industry. The process starts with the readings from ASHE that have been filled and scaled in the 48 regions by occupations arrays and multiplies each cell' of the array by the industry differential of which there are 75:
   Industry differential = Overall total for a particular industry / total for all industries

3. Then multiply by employment to calculate Man-Hours and finally scale these levels to match the overall regional all g-s, all occupations or all industry totals. The results of this calculation are written for the 4-digit occupations by Industry and for 2-digit occupations by Industry.

A.9 Details of the data on employment, pay and hours provided in the LMI for All database

A.9.1 Data provided

1. ***Employment***: The *Working Futures* employment data may be refined in the future. Any refinements will be discussed as part of possible developments for Phase 2B. Estimates of replacement demand will be generated 'on the fly' in the API. These will be based on data and instructions provided in the Wiki.
2. ***Pay***: the pay data (based on a combination of ASHE and LFS) are supplemented by a second file which provides the parameters necessary to generate estimates of pay by age. These will be created 'on the fly' within the API.
3. ***Hours***: hours database on ASHE contains information on average weekly hours. It does not cover variations by qualification (which is not available in ASHE).

A.9.2 Employment

The *Working Futures* employment data are supplied at a very detailed level without any sub totals. Data for an N-dimensional data array, with the following dimensions

- year - 2000-2020;
- gender – 2;
- status – 3;
- industry – 75;
- occupation – 369;
- geography – 12;
- qualification – 9.

The first column is the 'year', which runs from 2000 to 2020. The second to seventh columns, from 'gender' to 'qualification', indicate the characteristics of people covered by the dataset.The last column, 'weight', represents the number of people in the year specified in the first column and with the characteristics in columns two to seven. 'Weight' is simply the number of people (or fractions of a person). Most of the cells in this data array will have fewer than 10,000 people employed and many fewer than 1,000. In these cases the API needs to flag this up or suppress the numbers replacing them by more aggregate information as set out below. There are two main possibilities here:

1. Replacement by a *sub-total* across one (and/or if necessary more) of the main dimensions;
2. Replacement by categories at a higher level of aggregation (e.g. 2-digit or 3-digit rather than 4-digit SOC).

Sub-totals

Ignoring the time dimension, dealing with 1 requires the creation of the following sub-totals:

- both genders;
- all status types;
- all industries;
- all occupations;
- all countries /English regions;
- all qualifications.

In each case, all of the other dimensions can still be provided in full detail. This is done at the stage of preparing the data for the API by the Technical Team.

Aggregate categories

Dealing with 2 is in some respects more complicated as there are various possible aggregations.

No alternatives are possible for gender and status.

For ***industries*** the industries can be aggregated to various levels such as the 22 used in the *Working Futures* reports or 6 broad sectors also used there (see Tables A.2 and A.3).

For ***occupations*** aggregation could be made to the 3 or 2-digit level of SOC 2010.

Table A.4 shows the 1 and 2-digit levels only.

Countries /English regions – a possible aggregation here would be to the whole of England and the rest of the UK. These are not standard.

Finally, for qualifications, the nine fold classification based on the new NQF categories could be aggregated to the old six fold one in which the higher levels are combined. This can be aggregated to six broad categories of the old NQF as shown in Table A.5.

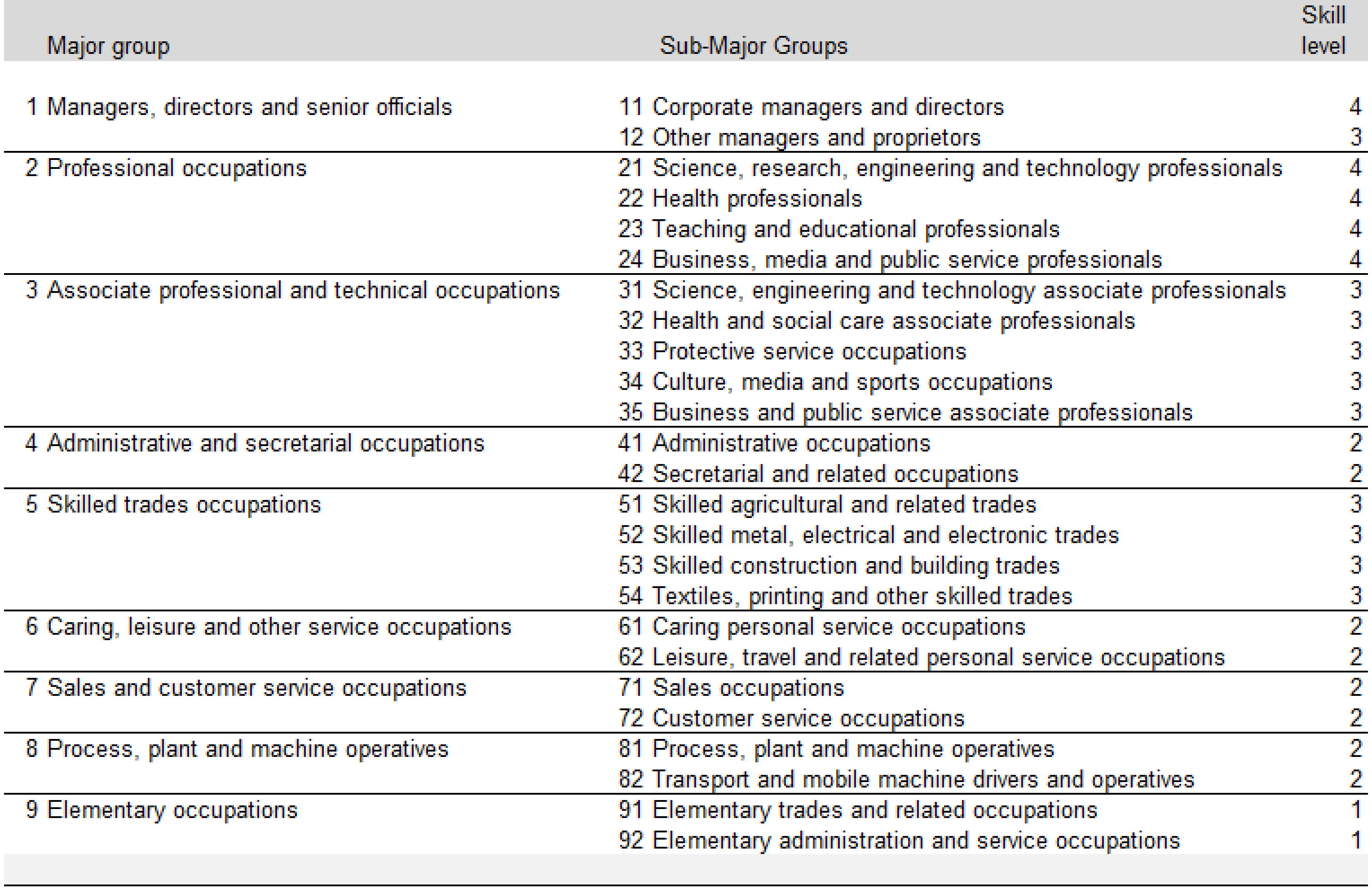
Table A.2 Broad Sectors (SIC2007)

Broad Sector	SIC2007 Section	SIC 2007 Division	Industry full name	Ind 22	Ind 79
1. Primary sector & utilities	A	01-03	Agriculture, forestry and fishing	1, 2, 6, 7	1-4, 28-31
	B	05-09	Mining and quarrying		
	D	35	Electricity, gas, steam and air conditioning		
	E	36-39	Water supply, sewerage, waste management		
2. Manufacturing	C	10-33	Manufacturing	3-5	5-27
3. Construction	F	41-43	Construction	8	32-34
4. Trade, accomod. & transport	G	45-47	Wholesale and retail trade; repair of motor vehicles	9-11	35-44
	H	49-53	Transport and storage		
	I	55-56	Accommodation and food activities		
5. Business & other services	J	58-63	Information and communication	12-17, 21-22	45-67, 73-79
	K	64-66	Financial and insurance activities		
	L	68	Real estate activities		
	M	69-75	Professional, scientific and technical activities		
	N	77-82	Administrative and support service activities		
	R	90-93	Arts, entertainment and recreation; other services		
	S	94-96	Other service activities		
6. Non-market services	O	84	Public administration and defence etc	18-20	68-72
	P	85	Education		
	Q	86-88	Human health and social work		

Table A.3 Industry Groups (SIC2007)

Ind22	Ind22 name	SIC2007 Section	SIC2007 Division	Industry full name	Industry 79
1	Agriculture	A	01-03	Agriculture, forestry and fishing	1
2	Mining & quarrying	B	05-09	Mining and quarrying	2-4
	Manufacturing	C	10-33	Manufacturing	5-27
3	Food drink & tobacco		10-12	Food drink and tobacco	5-6
4	Engineering		26-28	Engineering	20-22
5	Rest of manufacturing		13-25, 29-33	Rest of manufacturing	7-19
6	Electricity & gas	D	35	Electricity, gas, steam and air conditioning	28
7	Water & sewerage	E	36-39	Water supply; sewerage, waste management	29-31
8	Construction	F	41-43	Construction	32-34
9	Whol. & retail trade	G	45-47	Wholesale and retail trade; repair of motor vehicles etc	35-37
10	Transport & storage	H	49-53	Transport and storage	38-42
11	Accommod. & food	I	55-56	Accommodation and food activities	43-44
	Information & comm.	J	58-63	Information and communication	45-50
12	Media		58-60, 63	Media and communication	45-47, 50
13	IT		61, 62	Information technology	48-49
14	Finance & insurance	K	64-66	Finance and insurance activities	51-53
15	Real estate	L	68	Real estate activities	54
16	Professional services	M	69-75	Professional, scientific and technical activities	55-61
17	Support services	N	77-82	Administration and support service activities	62-67
18	Public admin. & defence	O	84	Public administration and defence etc	68
19	Education	P	85	Education	69
20	Health & social work	Q	86-88	Human health and social work	70-72
21	Arts & entertainment	R	90-93	Arts, entertainment and recreation; other services	73-76
22	Other services	S	94-96	Other service activities	77-79

Table A.4: SOC2010 Major Groups and Sub-major Groups



Major group	Sub-Major Groups	Skill level
1 Managers, directors and senior officials	11 Corporate managers and directors	4
	12 Other managers and proprietors	3
2 Professional occupations	21 Science, research, engineering and technology professionals	4
	22 Health professionals	4
	23 Teaching and educational professionals	4
	24 Business, media and public service professionals	4
3 Associate professional and technical occupations	31 Science, engineering and technology associate professionals	3
	32 Health and social care associate professionals	3
	33 Protective service occupations	3
	34 Culture, media and sports occupations	3
	35 Business and public service associate professionals	3
4 Administrative and secretarial occupations	41 Administrative occupations	2
	42 Secretarial and related occupations	2
5 Skilled trades occupations	51 Skilled agricultural and related trades	3
	52 Skilled metal, electrical and electronic trades	3
	53 Skilled construction and building trades	3
	54 Textiles, printing and other skilled trades	3
6 Caring, leisure and other service occupations	61 Caring personal service occupations	2
	62 Leisure, travel and related personal service occupations	2
7 Sales and customer service occupations	71 Sales occupations	2
	72 Customer service occupations	2
8 Process, plant and machine operatives	81 Process, plant and machine operatives	2
	82 Transport and mobile machine drivers and operatives	2
9 Elementary occupations	91 Elementary trades and related occupations	1
	92 Elementary administration and service occupations	1

Source: SOC2010: Volume 1: Structure and Description of Unit Groups

Table A.5 Qualifications

id	NQF	QCF	NQF (old)		qualification
1	NQF 8	QCF8 Doctorate	NQF 5		Higher degree or equivalent
2	NQF 7	QCF7 Other higher degree			
3	NQF 6	QCF6 First degree	NQF 4		Higher education
4	NQF 5	QCF5 Foundation degree; Nursing; Teaching			
5	NQF 4	QCF4 HE below degree level			
6	NQF 3	QCF3 A level & equivalent	NQF 3		GCE, A-level or equivalent
7	NQF 2	QCF2 GCSE(A-C) & equivalent	NQF 2		GCSE grades A*-C or equivalent
8	NQF 1	QCF1 GCSE(below grade C) & equivalent	NQF 1		Other qualifications
9	No Qualification	No Qualification	NQF 0		No qualification

Rules for answering queries – Employment

The proposed solution for employment is for the Technical team to:

1. generate the sub-totals from the file; and also
2. the following aggregations:
   a. industries - to the 22 industry level as in Table A.3;
   b. occupations - the 2-digit level as in Table A.4;
   c. qualifications - to the 6-fold level as in Table A.5

Note that any query relating to Replacement Demands needs to be dealt with analogously to employment:

1. If the numbers employed in a particular category/cell (defined by the 12 regions, gender, status, occupation, qualification and industry (75 categories)) are below 1,000 then a query should return 'no reliable data available' and offer to go up a level of aggregation across one or more of the main dimensions (e.g. UK rather than region, some aggregation of industries rather than the 75 level, or SOC 2-digit rather than 4-digit). The API is designed to default to fine data, but if that returns no reliable data available it offers the option of searching on coarse data. It is, of course, possible to pre-set the query to coarse data.
2. If the numbers employed in a particular category / cell (defined as in 1.) are between 1,000 and 10,000 then a query should return the number but with a flag to say that this estimate is based on a relatively small sample size and if the user requires more robust estimates they should go up a level of aggregation across one or more of the main dimensions (as in 1)

A.9.3 Pay and Hours

Analogous data are provided for pay and hours. Note that the pay and hours data currently relates to just a single year (2012). The file for an N-dimensional data array, includes the following dimensions:

- gender – 2;
- status 1 (full-time employees only at present);
- industry – 7;
- occupation – 369;
- geography – 12;
- qualification – 9.

The first column is the 'year', currently just for 2012. The second to seventh columns, from 'gender' to 'qualification', indicate the characteristics of people covered by the dataset. The last column, represents the average weekly pay or average weekly hours for the year specified in the first column and with the characteristics in columns two to seven.

In order to assess reliability, the API needs to check back on the corresponding 'weight' from the *Working Futures* Employment dataset (which gives the number of people (or fractions of a person) employed in the relevant category). Most of the cells in this data array will have fewer than 10,000 people employed and many fewer than 1,000. In these cases the API needs to flag this up or supress the pay or hours estimates replacing them by more aggregate information as set out below. As for employment there are two main possibilities here:

1. Replacement by a *sub-total* across one (and /or if necessary more) of the main dimensions;
2. Replacement by categories at a higher level of aggregation (e.g. 2-digit or 3-digit rather than 4-digit SOC).

The proposed solution for hours and pay is to default to:

Create sub-totals; plus the following aggregations:

a. industries – to the 22 industry level as in Table X.2;
b. occupations – the 2-digit level as in Table X.3;
c. qualifications – to the 6-fold level as in Table X.5.

Note that in the case of pay additional supplementary information is provided to enable the generation of estimates of pay by age form the mean value and the selected age. See Annex A.7 for details

Annex B: O*NET

B.1 Introduction

As mentioned in Annex A, skills data from the US O*NET database was one of the sets of indicators used in the pilot phase (focussing in particular on STEM skills). In Phase 2A, the potential of this source has been further explored. This Annex describes that process in more detail, including recommendations for Phase 2B.

B.2 Initial Approach

The initial approach followed the one developed by Dickerson and Wilson (2012). Dickerson and Wilson (2012) established the general feasibility of exploiting the huge investment made in the US O*NET system by mapping this to categories defined using the UK Standard Occupational Classification (SOC). The LMI for All project has moved a step closer to fully operationalising this process. This has involved sorting out various 'teething problems' identified in the initial feasibility study and extending the exploitation of the O*NET database to include other domains.

The US-based Occupational Information Network (O*NET) system provides almost 250 measures of skills, abilities, work activities, training, work context and job characteristics for each of around 1,000 different US occupations (based on a modified version of the US Standard Occupational Classification), with information gathered from both job incumbents through standardised survey questionnaires, as well as assessments by professional job analysts.

The first area identified for improvement relates to improving the CASCOT Matching Process.[32] Dickerson and Wilson concentrate in their report on what they refer to as 'Variant 3' matching, which matches the 56,634 job titles in the O*NET-SOC2009 lay job title file titles into SOC2010, using the SOC2010 classification dictionary and rules in CASCOT.[33] It was proposed to refine and extend this process in order to get a better match.

In the original matching process the distribution of CASCOT scores that measure the strength of the match indicated some problems. It was decided that to better increase the chances of job titles to be matched, it would be of some benefit to match the sub groups first. Using a mixture of Cascot and Excel the 1103 ONET sub-categories were matched to the most relevant 369 SOC categories, by hand (and using some web search for any difficult decisions). Matching the sub-categories in this way then makes CASCOT's job simpler in the next steps. A job title will only be searched within its designated sub-category, which means it should not be matched to a completely irrelevant job.

[32] CASCOT (Computer Assisted Structured COding Tool) is a piece of specialist software, originally developed by Warwick IER, designed to classify occupational title into Standard Occupational Classification (SOC) categories.

[33] For a detailed description of O*NET see Tippins and Hilton (2010).

The method involved the following steps:

- Initially a joint index of the sub categories was created, this was to enhance the CASCOT scoring so that when the ONET (56,000) jobs are put through CASCOT, each job title would be more likely to allocated to the correct SOC sub-category (369).
- All the job-titles (ONET and SOC – Almost 80,000) were then placed into Excel, as the SOC job titles already have a relevant sub-category code, at this stage the ONET sub-category codes needed to be matched to the relevant SOC sub-category code. This was done using the index function in excel (as by hand, this would have been an extremely laborious task).
- The joint SOC+ONET classification was then created with the CASCOT editor by opening the CASCOT bundled SOC2010 classification. The index which was created in steps 2-3 was imported and then saved as a classification file.
- Within CASCOT, all 56,000 job titles were run through, using automatic matching and the output created was saved.
- The output file was then converted to an Excel workbook for viewing purposes.

After assessing the initial output using this method, and consulting with Professor Peter Elias the designer of CASCOT, it was decided that a different approach would be more satisfactory.

B.3 Alternative approach for LMI for All

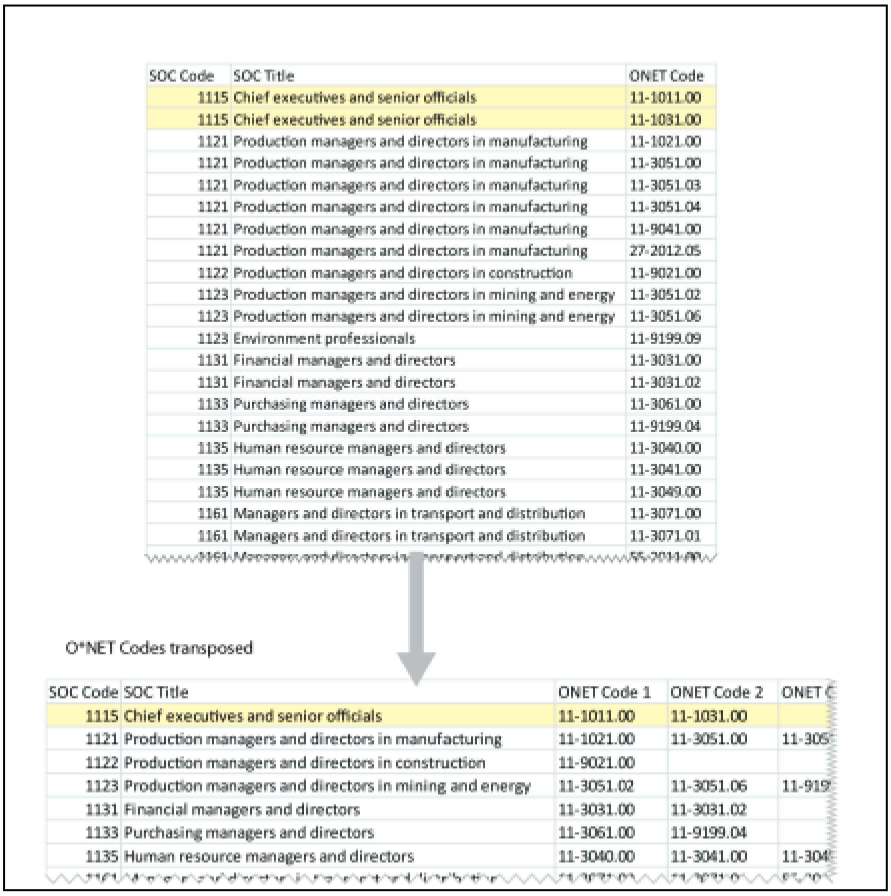
In the course of this analysis it became clear that there was considerable difficulty in getting unambiguous matches and finding unique one to one mapping, as well as on developing a suitable weighting scheme for combining occupations together.

Rather than creating a SOC 4-digit O*Net database (369 categories) the emphasis therefore shifted to one of linking directly from the 369 SOC 4-digit categories (and the underlying 28,000 SOC occupational titles (effectively SOC 6 digit)), recognising that there may be no unique mapping, but links to more than one O*NET group.

In the course of Phase 2A of LMI for All, it was decided to explore this alternative approach which involved using CASCOT to match from the 28,000 (or so) occupational titles used in SOC2010 directly to the ONET categories (and thereby to the skills database).

As the ultimate aim of the task is to link skills information available from O*NET for each UK occupational title or category, it was recognised that the previous method (above) would not necessarily bring up the correct skills associated. Therefore it was decided that it would be best to match O*NET US SOC (1,103) to UK SOC (369) at a unit-group level. This is a more straightforward approach, and (as it is done using human judgement) produces much better results. However, it also means that the scores are of less interest. Using CASCOT to manually go through every entry, each US unit-group was individually considered and matched to a corresponding UK unit-group and checked (with the help of the O*NET website search facility). The unit-group mapping was then exported from CASCOT as a CSV file, the list (1,103 rows) of US unit-group codes mapped to an O*NET counterpart. This CSV file was imported into Excel and saved. To get the data into a more useful form, any multiple O*NET codes were then transposed to multiple columns (see diagram below). The rows of the revised table correspond to the 369 UK SOC2010 digit categories. For instance: UK SOC2010 #1115 maps to both US SOC2009 #11-1011.00 & 11-1031.00.

Table B.1 Mapping from SOC 4-digit categories directly to O*NET



SOC Code	SOC Title	ONET Code
1115	Chief executives and senior officials	11-1011.00
1115	Chief executives and senior officials	11-1031.00
1121	Production managers and directors in manufacturing	11-1021.00
1121	Production managers and directors in manufacturing	11-3051.00
1121	Production managers and directors in manufacturing	11-3051.03
1121	Production managers and directors in manufacturing	11-3051.04
1121	Production managers and directors in manufacturing	11-9041.00
1121	Production managers and directors in manufacturing	27-2012.05
1122	Production managers and directors in construction	11-9021.00
1123	Production managers and directors in mining and energy	11-3051.02
1123	Production managers and directors in mining and energy	11-3051.06
1123	Environment professionals	11-9199.09
1131	Financial managers and directors	11-3031.00
1131	Financial managers and directors	11-3031.02
1133	Purchasing managers and directors	11-3061.00
1133	Purchasing managers and directors	11-9199.04
1135	Human resource managers and directors	11-3040.00
1135	Human resource managers and directors	11-3041.00
1135	Human resource managers and directors	11-3049.00
1161	Managers and directors in transport and distribution	11-3071.00
1161	Managers and directors in transport and distribution	11-3071.01
1161	Managers and directors in transport and distribution	55-2011.00

O*NET Codes transposed

SOC Code	SOC Title	ONET Code 1	ONET Code 2	ONET C
1115	Chief executives and senior officials	11-1011.00	11-1031.00	
1121	Production managers and directors in manufacturing	11-1021.00	11-3051.00	11-305
1122	Production managers and directors in construction	11-9021.00		
1123	Production managers and directors in mining and energy	11-3051.02	11-3051.06	11-919
1131	Financial managers and directors	11-3031.00	11-3031.02	
1133	Purchasing managers and directors	11-3061.00	11-9199.04	
1135	Human resource managers and directors	11-3040.00	11-3041.00	11-304

Once the data had been transposed, any missing UK SOC sub-group codes were then filled in and corresponding O*NET codes were chosen (with the help of the O*NET website) and placed into the ONET Code columns. At this stage a complete set of SOC unit-group codes had been filled with potentially multiple corresponding O*NET unit-group codes. Each one was checked and any further suitable additions or adjustments were made to complete the unit-group mapping. The data file was then further extended to the full set of UK SOC occupational titles (27,739). To expand to the full list of UK occupations, a file with just the complete list of job titles was imported into Excel. Using the 'INDEX()' and 'MATCH()' function it was possible to match each job title into the corresponding unit-group mapping. Note, although the data is extended to the full UK SOC 27,739 job titles it is not necessarily a unique map from a UK code to just one US SOC category and this gives the same mapping as the aggregate one described above.

Converting UK SOC to US SOC

UK SOC identifies 27,739 occupations, these have been mapped to multiple US O*NET occupations by matching the sub-groups to one another using CASCOT (Computer Assisted

Structured COding Tool) software (see below), as well as Excel.

- ❖ Match 1,103 US SOC sub-groups to 369 UK SOC sub-groups using CASCOT software and refining choices within Excel;
- ❖ Extend to detailed UK SOC occupational level (27,739) (currently all titles within a SOC 4-digit code are allocated to the same O*NET category).

Skills and abilities data

There are potentially many new data, including the Skills and Abilities, which can be matched. For instance, 'Abilities.txt' and 'Skills.txt' (see Tables B.2 to B.4), which both come from the O*NET website and contains ability or skills scores for O*NET SOC codes (occupations). The information shows both the levels of skills or abilities required and the importance of these skills/abilities for the occupation concerned. See Table B.2.

Table B.2 Data layout of 'Skills.txt' and 'Abilities.txt'

Variable Name	**Variable definition**
Scale ID	Scale used as the basis for rating, IM (Importance) or LV (Level) (see below)
Data Value	Rating associated with the O*NET-SOC occupation (Importance 1-5) (Level 0-7) (see below). These are included as two separate rows for each occupation, one for IM and one for LV.
N	Sample Size *
Standard Error	Indication of each estimate's precision
Lower CI Bound	Lower 95% Confidence Interval Bound (see below)
Upper CI Bound	Upper 95% Confidence Interval Bound (see below)
Recommend Suppress	Low Precision Indicator (Y=yes, N=no)
Not Relevant	Not Relevant for the Occupation (Y=yes, N=no) (see below)
Date	Date when data was updated *
Domain Source	Source of the data *

* These items are probably not very relevant for the LMI forAll database and could be omitted. For the moment the O*NET file is included in its entirety.

Table B.3 show details from the Abilties.txt file. The O*NET-SOC Code is linked by its 8-digit unique occupation identifier to the Element ID (Ability Outline Position in the Content Model Structure) and to the Element Name (Names of the 52 abilities included).

Table B.3 Abilties.txt

Arm-Hand Steadiness	Facility
Auditory Attention	Comprehension
Flexibility	Expression
Precision	Originality
Reasoning	Speed
Perception	Vision
Flexibility	Sensitivity
Strength	Control
Strength	Time
Flexibility	Orientation
Vision	Attention
Dexterity	Localization
of Closure	Orientation
of Ideas	Clarity
Sensitivity	Recognition
Body Coordination	of Closure
Body Equilibrium	of Limb Movement
Sensitivity	Stamina
Reasoning	Strength
Ordering	Sharing
Dexterity	Strength
Reasoning	Color Discrimination
Memorization	Visualization
Coordination	Speed
Vision	Comprehension
Vision	Expression

Similarly, Table B.4 shows how details from the Skills.txt. Again the O*NET-SOC Code (with its 8-digit unique occupation identifier links to Element ID (the Skill Outline Position in the O*NET Content Model Structure and the Element Name (the names of the 36 skills identified).

Table B.4 Skills.txt

1. Active Learning
2. Active Listening
3. Complex Problem Solving
4. Coordination
5. Critical Thinking
6. Equipment Maintenance
7. Equipment Selection
8. Installation
9. Instructing
10. Judgment and Decision Making
11. Learning Strategies
12. Management of Financial Resources
13. Management of Material Resources
14. Management of Personnel Resources
15. Mathematics
16. Monitoring
17. Negotiation
18. Operation and Control
19. Operation Monitoring
20. Operations Analysis
21. Persuasion
22. Programming
23. Quality Control Analysis
24. Reading Comprehension
25. Repairing
26. Science
27. Service Orientation
28. Social Perceptiveness
29. Speaking
30. Systems Analysis
31. Systems Evaluation
32. Technology Design
33. Time Management
34. Troubleshooting
35. Writing

Table B.5 Alternative steps to improving the matching

The matching of O*NET occupational titles to SOC was refined as follows:

1. Each US O*NET sub-group was matched to a preferred UK sub-group. Using CASCOT, each O*NET sub-group entry (1,103) was chosen using Google and a search on the O*NET site as help, to the best UK SOC match.

   - The US index (sub-group) was used as the input file in CASCOT.

   - The UK SOC index was used as the classification.

   - The output file would be a 1 to many file (for instance, there are 1,103 US sub-groups, therefore the UK sub-groups in some cases will arise multiple times.

2. The choices were further refined within Excel.

3. The suggested refinements were then combined into a single column which gave a complete set of O*NET sub-groups to UK sub-groups.

4. As this process gave a 1 to many output (for instance O*NET codes 11-1011.00 and 11-1031.00 both map to UK 1115, the data were placed into a pivot table to show duplicates and make the data more easily transposable for the next step of the process.

5. The pivot table data was then copied and pasted into another worksheet and duplicate UK SOC sub-groups were transposed into the columns.

6. As is it possible for more than one UK sub-group to be allocated to a US sub-group, there were 14 missing SOC codes, which needed to be further filled by hand. This was done and further refinements were made. This step removed any codes which were deemed unsuitable and adding in anything else which may help the skills search process.

7. The worksheet is a cleaned up version.

Other developments

These will be largely completed in Phase 2B.

O*NET has updated to US SOC2010

Dickerson and Wilson (2012) used the version of the US-SOC available when the work was undertaken. This has now been updated. For US-SOC and O*NET-SOC, Dickerson and Wilson used the 2009 classifications but since then, the O*NET system has updated its SOC classification to a new O*NET-SOC2010 version. This new taxonomy is used with release Version 15.1 of the O*NET database. The O*NET-SOC2010 taxonomy is designed to be compatible with changes made to the US SOC2010 and to align the two classification systems. This modification to the O*NET SOC will not cause any immediate problems for the project, but will have implications for potential future revisions. The O*NET-SOC2010 taxonomy has 1,110 occupational titles, 974 of which will have data within the O*NET system. Much of the information for O*NET-SOC2009 will carry over, but the matching of job titles should be updated to the O*NET-SOC2010.

A decision was taken to focus on ‘ONET 15’ in Phase 2A to maintain consistency with what had been done previously. The data incorporated in the current database are therefore based on ONET 15.0 (US SOC2009) rather than the latest ONET 17.0 (US SOC2010).

Weighted database using employment weights and scores

Dickerson and Wilson constructed weights based on both the CASCOT scores and also the importance of the occupation in the US (using employment weights derived from BLS 2008). This has not been carried out in the revised process so a database has not been constructed. Further work could be done to carry this task through but this is not central to what is needed for LMI for All and is probably best seen as a possible separate project.

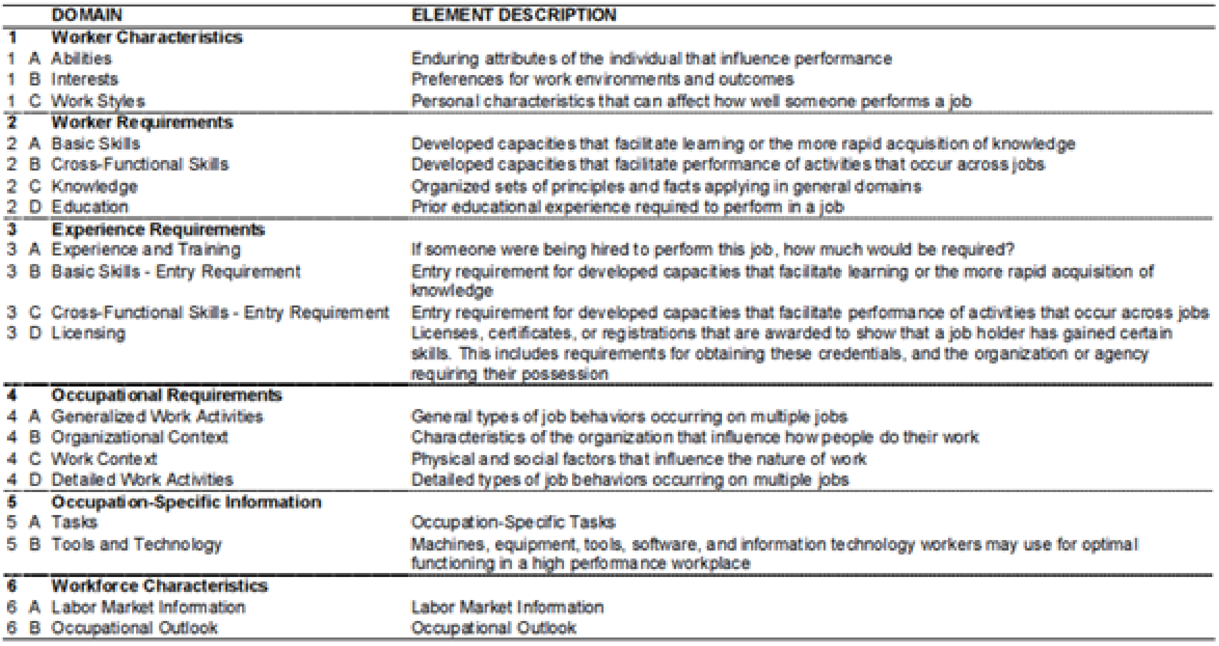
Extending to other O*NET domains

Dickerson and Wilson (2012) focussed on the skills and abilities domains in their report in order to demonstrate feasibility. They gave examples focussing on STEM occupations. This can now be extended to cover the full range as shown in Tables B.2 and B.3.

There are also many other domains in O*NET that Dickerson and Wilson did not examine. Some of these are not really relevant since they are US-specific - but others could potentially provide useful information that could be relevant for the database (e.g. required training, etc.).

More time needs to be spent interrogating all of the 250 or so items that are available, and considering which are likely to provide useful information for the UK occupations profiles. The full set of indicators available is summarised in Table B.5 below. This will require careful judgements about, which of the many dimensions would be most useful.

Table B.6 The O*NET content model



		DOMAIN	ELEMENT DESCRIPTION
1		**Worker Characteristics**	
1	A	Abilities	Enduring attributes of the individual that influence performance
1	B	Interests	Preferences for work environments and outcomes
1	C	Work Styles	Personal characteristics that can affect how well someone performs a job
2		**Worker Requirements**	
2	A	Basic Skills	Developed capacities that facilitate learning or the more rapid acquisition of knowledge
2	B	Cross-Functional Skills	Developed capacities that facilitate performance of activities that occur across jobs
2	C	Knowledge	Organized sets of principles and facts applying in general domains
2	D	Education	Prior educational experience required to perform in a job
3		**Experience Requirements**	
3	A	Experience and Training	If someone were being hired to perform this job, how much would be required?
3	B	Basic Skills - Entry Requirement	Entry requirement for developed capacities that facilitate learning or the more rapid acquisition of knowledge
3	C	Cross-Functional Skills - Entry Requirement	Entry requirement for developed capacities that facilitate performance of activities that occur across jobs
3	D	Licensing	Licenses, certificates, or registrations that are awarded to show that a job holder has gained certain skills. This includes requirements for obtaining these credentials, and the organization or agency requiring their possession
4		**Occupational Requirements**	
4	A	Generalized Work Activities	General types of job behaviors occurring on multiple jobs
4	B	Organizational Context	Characteristics of the organization that influence how people do their work
4	C	Work Context	Physical and social factors that influence the nature of work
4	D	Detailed Work Activities	Detailed types of job behaviors occurring on multiple jobs
5		**Occupation-Specific Information**	
5	A	Tasks	Occupation-Specific Tasks
5	B	Tools and Technology	Machines, equipment, tools, software, and information technology workers may use for optimal functioning in a high performance workplace
6		**Workforce Characteristics**	
6	A	Labor Market Information	Labor Market Information
6	B	Occupational Outlook	Occupational Outlook

Source: *O*NET Content Model*

Annex C: Review of potential of other data to be considered for Phase 2B

C.1 Introduction

This annex covers a number of potential sources which might enrich the LMI for All database. These include:

C.2 ONS Vacancy Survey
C.3 Annual Population Survey (APS)
C.4 NOMIS:

- Employment (at local level)
- Claimant unemployment rate
- Job Centre Plus vacancies (historical series)

C.5 Census of Population
C.6 *Working Futures* – refinement of employment projections (SOC 4-digit level)
C.7 Cedefop – pan-European employment projections
C.8 Other European datasets
C.9 Course information (including HESA destination data)

C.2 ONS Vacancy Survey

The ONS Vacancy Survey is intended to be an accurate count of the total stock of vacancies, addressing the problem that the administrative count is thought to only capture around a third of all vacancies. It is a regular survey of vacancies across all businesses with employment greater than 1. Designed to minimise the administrative burden on businesses, it asks only one question; 'how many vacancies an organisation had on a set date for which external applicants are actively being sought'. The survey commenced in November 2000, covering only the Production, Construction and Public Administration industrial sectors, and was extended to cover all industry sectors except agriculture, forestry and fishing in April 2001. Employment agencies are excluded in order to avoid the risk of double counting vacancies. The survey is sampled from the Interdepartmental Business Register (IDBR), with around 6000 telephone interviews per month, 1,300 of which are to large enterprises included each time. The remaining 4,700 smaller enterprises are randomly sampled on a quarterly basis. The quarterly sample size is approximately 15,400 separate enterprises and the annual sample size is 57,700 separate enterprises. The sampling error is around 3 per cent for monthly estimates, 1.5 per cent for the 3-monthly rolling averages and 10 per cent for three-month average vacancy counts for a typical industry sector.

The survey yields UK estimates of the total number of vacancies by firm size and industry for rolling quarters from 2001 onwards. It yields no information by occupation. Data is published on the ONS website (http://www.ons.gov.uk/ons/rel/lms/labour-market-statistics/april-2013/index-of-data-tables.html#tab-Vacancies-tables), and for this there are no issues regarding access or confidentiality.

Indicators available (for the UK as a whole only):

- Total vacancies;
- Number of unemployed persons per vacancy (the U/V ratio);
- Vacancies by size of enterprise;
- Vacancies by SIC 2007 industry section (and selected 2-digit industries);
- Vacancies per 100 employee jobs by SIC 2007 industry section (and selected 2-digit industries).

Concluding remarks

This source is probably the most accurate measure of the total number of vacancies in the economy. The ONS datasets based upon this source present the trend over time in the number of vacancies and the unemployment/vacancy ratio (an indicator of how hard it is to obtain a job and whether it is becoming harder or easier). However, the survey yields no information by occupation.

It could be used in an introductory page to indicate the general state of the job market and how complete other sources are. The main focus of LMI for All is on helping people seeking careers guidance and advice. It is not clear how much they need information on the general state of the labour market although such information is useful for supporting general labour market analysis. It is therefore of lower priority than other datasets discussed in this document.

Given that no occupational detail is possible it is recommended NOT to include this source as a priority.

C.3 Annual Population Survey

The Annual Population Survey (APS) is a boosted version of the Labour Force Survey (LFS). The aim of the boost is to achieve a large enough sample in each local authority district for statistically reliable labour market measures to be derived. First conducted in 2004, it combines results from the LFS and the English, Welsh and Scottish LFS boosts. Datasets are produced quarterly, with each dataset containing 12 months of data. The sample size is 155,000 households and 360,000 persons per dataset. The sample size is largest in Unitary Authorities (including all Welsh and Scottish local authorities), followed by London Boroughs. In most lower tier local authorities in England, the sample size is a few hundred, and is smallest in rural areas.

In principle, APS microdata can yield a large amount of information on the employment characteristics of residents of a geographical area. The survey collects the most detailed industry and occupation codes for current and previous job. It is available from the UK Data Archive (UKDA) in three versions - End User Licence, Special Licence and Secure Data Service. The level of geographical detail increases in these from statistical region and nation (in the End User Licence version) to local authority district (in the Special Licence version) to Output Area (in the Secure Data Service version). Restrictions on access become greater as the level of detail increases and limit the ability of analysts to distribute data from the APS to third parties.

The same variables which have been defined for the LFS could be created for Unitary Local Authority Districts using the Special Licence APS. However, the sample size may be too small for reliable estimates to be made for many areas, although it may be large enough for some local areas. The sample size can be increased by combining data for a series of years aggregated, but this reduces the topicality of the data.

Unlike the LFS, the region where an individual works is only available in the Special Licence version of the APS. Thus data on the occupational breakdown of employment by workplace can only be generated using a version of the dataset for which access is more restricted.

The End User Licence version of the APS has least restrictions placed on its use. Variables which can be generated using this version of the APS describe the characteristics of workers living in an area, rather than those of people working in an area. These include:

- Qualifications of workers;
- Occupational profile of workers;
- Prevalence of self-employment by occupation;
- Unemployment rates by age or qualification level.

The APS is an important source of labour market intelligence for government statisticians and some departments of central government and devolved administrations have the capacity to generate tables from the APS. If such tables could be provided by the ONS or another department, these might be used in the database as an alternative to generating tables from APS microdata (which are subject to restrictions on their wider distribution).

Concluding remarks

The APS dataset provides access to the same range of variables available from the LFS, at a more detailed geographical scale, and thus (in principle) is a better choice for the database than the LFS. Though the sample size is too small for the APS to yield information for all local authority districts, it can provide information for cities and most London Boroughs. The data covers successive 12-month periods. Data could either be presented for the most recent calendar year or for the most recent 12-month period for which data was available.

The current use of the LFS in the LMI for All database has been narrowed down to providing unemployment rates (see Section 2 of the main report). The extra value of the APS in this regard is probably quite limited although the more aggregate 'headline ' figures could be recomputed using the APS since the latter contains the same variables for a larger sample size and offers the potential of more detailed geographical breakdowns.

However, the implications of ONS licence restrictions upon the data which can be generated from the LFS and APS needs to be further investigated in order to decide whether additional LFS/APS data can be included. If data generated from microdata accessed via the UK Data Archive cannot be used, then extracts of the data (from which the list of variables above could be generated) could be requested from ONS or other government statisticians. A more limited range of APS-derived variables are accessible via NOMIS (discussed in the next section).

There are restrictions on access to information derived from LFS and APS microdata via the UKDA (i.e. even for access to End User Licence data it is necessary to be a registered user, to describe the purpose the data are to be used for (which should be broadly academic, and for which the period of access is limited). It is not clear whether this would allow the freedom to distribute such data publicly by including it in the database. This will require further negotiations with the UKDA and ONS. The marginal benefits are therefore probably modest and the marginal costs quite high.

C.4 NOMIS

- Employment data
- Unemployment claimant count
- Job Centre Plus vacancies (historical series)

NOMIS holds the full range of labour market related ONS and DWP statistical outputs available at the sub-regional scale. It provides extremely easy access to a time series of data going back to 1982 for the majority of datasets and 1971 for a few others (e.g. employment and June unemployment). Most NOMIS datasets cover either Great Britain or the whole of the UK. The NOMIS datasets are accessible via a ‘Restful’ API interface.

Employment data

Employment data in NOMIS derives from a number of official sources: the ONS annual surveys of employment, ONS estimates of workforce jobs and the Annual Population Survey. The first of these encompasses data aggregated to geographical areas from the (Annual) Census of Employment, the Annual Business Inquiry (ABI) and the Business Register and Employment Survey (BRES). This provides an (almost) annual time-series of employment located in a geographical area from 1971 onwards. The only variables contained in the dataset are the industry (to the lowest level of the relevant version of the Standard Industrial Classification) and employees broken down into full and part-time working. Until the BRES was introduced in 2008, there was also a breakdown of employees by gender. The BRES presents a count of employees by full- and part-time status and total employment (including working proprietors). The current geographical breakdown of employment is for Census Output Areas (small areas containing on average 200 households) and for all larger areas, which these nest into. A flag is attached to each data item indicating whether the data is statistically robust. All numbers must be rounded to the nearest 100.

The indicators that could be derived include:

- Location of jobs by industry;
- Industrial profile of employment in an area;
- Location of part-time jobs.

However, access to the data is problematical, because these surveys are collected under the Statistics of Trade Act 1947 which promises to maintain the confidentiality of data provided by survey respondents. Hence all users have to apply for and purchase a ‘Notice’ from the Department for Business Innovation and Skills in order to use the data.

The ONS estimates of workforce jobs provide quarterly information on employment by SIC 2007 industry section for English regions and the other nations of the UK from December 1992 onwards. This source includes estimates of total workforce jobs, together with its breakdown into employee jobs, self-employment, government-supported trainees and HM forces. Employment numbers are disaggregated by gender and full-time/part-time status. There are no restrictions upon access to this dataset.

The indicators that could be derived for each region include:

- Industrial breakdown of employees;
- Industrial breakdown of workforce jobs;
- Percentage of workforce jobs accounted for by the self-employed by industry;
- Percentage of employee jobs full-time by industry;
- Percentage of employee jobs female by industry.

The Annual Population Survey was described above. NOMIS includes a number of standard tables and variables created from the APS for a range of geographical scales. These include the occupational breakdown of employment and the qualifications of workers. The occupational breakdown is limited to SOC 2010 major and sub-major groups. Most tables and variables represent the characteristics of workers resident in an area. A smaller range of tables present the occupational and industrial profile of jobs located in an area. For individual cells of a table and individual variables a flag is provided which indicates the degree of statistical reliability of the value. Where the sample size is too small and standard error too great the data value is suppressed.

The indicators available from the APS via NOMIS include:

- Occupational profile of employment;
- Qualification profile of employment;
- Labour market participation by age group, gender, ethnicity and nationality.

The advantages of using APS data from NOMIS is that there are no problems of access, it is available for different levels of aggregation and the statistical flags attach identify which data is reliable and indicate the limits of data disaggregation.

Unemployment claimant count

NOMIS provides access to monthly ONS claimant count statistics from June 1971 onwards. The official definition of the unemployment count changes occasionally and is currently the number of people claiming Job Seekers Allowance and National Insurance Credits. The method of collection changed from manual to computerised processing in 1982. Since 1982 monthly or quarterly data on stocks and flows of people claiming unemployment benefit have been produced, disaggregated by age, gender and duration of claim. Since 2005, these statistics have been disaggregated by previous occupation (SOC 2000, coded to 4-digit level) and ethnic group. The unemployment series includes marked seasonal fluctuations, which can be adjusted for. Following the introduction of Universal Credit (being introduced from April 2013), the claimant count will include: people claiming contribution-based JSA (which is not affected by the introduction of Universal Credit), people claiming means-tested JSA during the transition period while this benefit is being gradually phased out, and people claiming Universal Credit who are not earning and who are subject to a full set of labour market jobseeker requirements (i.e. required to be actively seeking work and available to start work). The impact of Universal Credit upon the count is currently very small and confined to the pilot areas in Greater Manchester.

Since June 1982, the data has been produced for electoral wards and the geographical hierarchy of administrative and statistical areas. While there is thus comprehensive information on the number of unemployment claimants, the incidence of unemployment is measured less well. The unemployment rate is the number of unemployed people as a percentage of the economically active population. Until the late 1990s, unemployment rates were calculated for Travel-to-Work Areas (TTWA), which represent relatively self-contained local labour market areas. The economically active population was estimated as the sum of unemployed people plus the total number of jobs located in the TTWA. Since then the unemployment rate denominator at the regional scale and above has been derived from estimated workplace jobs (the sum of employee jobs, self-employment jobs, HM Forces and government-supported trainees) and unemployment. For smaller area, an unemployment proportion has been published, which is the ratio of the claimant count to the number of people aged 16 to 64 (taken from the annual population estimates).

Variables relevant to an appreciation of the labour market which can be defined using the claimant count (most can be disaggregated by gender):

- Unemployment rate (for regions);
- Unemployment proportion;
- Likelihood of becoming unemployed;
- Likelihood of leaving unemployment;
- Proportion of the unemployed in each occupational category.

Jobcentre Plus (JCP) vacancies

NOMIS holds a time series of vacancy data from 1978, with data derived from automated processing since June 1982. The datasets encompass notified and unfilled vacancy stocks and flows for industry (SIC 92 and SIC 2003, 2-digit level) and occupation (SOC 2000, to 4-digit level) and by duration. This is now a historical series, because data collection ended in October 2012 when Monster.co.uk took over from Jobcentre Plus.

Vacancy data is available for the statistical hierarchy of geographical areas from electoral wards to counties, regions and nations and for Jobcentre office areas. Data for the former are generated from the true location of the vacancies, but Jobcentre areas provide information about the location of the Jobcentre Plus office that is designated as owning the vacancy.

Possible indicators:

- Unfilled (live) vacancies by occupation and gender;
- Duration of vacancy by occupation and gender.

Concluding remarks

NOMIS provides access to rich data on employment and the labour market. It is regularly updated and the NOMIS team solves many of the problems associated with changing statistical geographies and variable definitions. Data can be directly read via a Restful API interface and items selected from the database for varying geographies and time periods. Though not covered above, NOMIS also provides very easy access to data from the 2011 Census of Population via this interface and via simple bulk downloads. It is a source that is invaluable for any general labour market analysis application.

Recommendations relating to the individual datasets described above:

- Employment – the main problem for the inclusion of employment data is the legal conditions which apply to access. Detailed data from the government surveys of employment cannot be included because of this. NOMIS can provide API access to the ONS regional workforce jobs estimates by industry section, and estimates of employment by (SOC2010) occupation and qualification from the APS These are regularly updated and not subject to restrictions on their use. It could therefore be worth including workplace job estimates as an alternative employment measure (which is in the public domain) as well as the APS occupational employment data. However, before doing this it would be wise to do some detailed comparisons with the existing *Working Futures* estimates. Many of these data have become available since the *Working Futures* database was created. They would be used to update it, as and when a new round of *Working Futures* is commissioned. However, the marginal value of adding these data sources is relatively modest compared with what is already available via *Working Futures*.

- Unemployment – these data are valuable as a source of information on the state of the labour market. It is possible to calculate unemployment rates and measures of the probability of leaving unemployment for small geographical areas using these data sources. However, the breakdown of unemployment by occupation uses the SOC2000 classification and thus is not very useful given the focus on SOC2010 categories. Any recommendation to include measures of unemployment incidence and dynamics from NOMIS, would need to be based on the judgment that information on unemployment trends adds value to the database from a general labour market analysis perspective.
- Vacancies – Though a wealth of data on the stocks and flows of vacancies is available via NOMIS, this dataset is no longer live and the occupational classification used is SOC2000, not SOC2010. Therefore, it is recommended not to include NOMIS historical JCP vacancy data (although again it could add value from a more general labour market analysis perspective).

C.5 The UK Census of Population

Possible indicators for the LMI for All database from the UK Census of Population:

- Labour market and employment data from the Census
- The publication sequence
- Commuting and workplace data

General background

The Census of Population is the most comprehensive survey of the socio-economic characteristics of the population. It represents a snapshot of the population on the 'Census Day' – most recently March 27th 2011. The Census is undertaken by three statistical offices: the Office for National Statistics in England and Wales, by the General Register Office for Scotland and the Northern Ireland Statistics and Research Agency. The questionnaire distributed by each is very similar, but there are differences in question content and wording to represent national differences (e.g. to collect information on use of the national languages: Welsh in Wales, Gaelic in Scotland and the Irish language in Northern Ireland).

The strengths of the Census are that it has a response rate of well over 90 per cent and that it yields statistically robust information for very small geographical areas. The smallest areas for which Census data is released ('Output Areas') have populations of around 200 people. A follow-up survey conducted soon after the census (the Census Coverage Survey) is used to calculate response rates and provides input data for the 'One Number Census' process, which adjusts the Census data to represent 100 per cent of the population. During this process, the Census results are also validated against other data sources.

From the 2001 Census onwards, all published data is based on processing 100 per cent of Census responses. Hence the published results of the Census (even for small areas) are not subject to sampling error, but detailed tabulations have the potential to disclose information about identifiable individuals. To protect against this possibility, a small amount of uncertainty is introduced into the data (by swapping the locations of a small number of responses). The population base for most Census tables is the population resident (or planning to be resident) in the UK for 12 months or more.

The Census is only collected once every ten years and it takes at least two years before the results are available (because of the amount of work involved in processing the data), and hence can be criticised for being almost immediately out-of-date. Outputs from the Census are mainly in the form of pre-designed tables, intended to meet the needs of the various stake-holders for information in a standardised form at different geographical scales. The amount of detail disclosed is usually limited by the need to preserve confidentiality in small populations, especially where detailed cross-tabulations are presented.

The first results (a simple count of the number of persons and households present in each local authority district) from the 2011 Census of Population were published in July 2012. Increasingly detailed results are published over a period of 18 months. The first results on the characteristics of the population are published in univariate tabulations for geographical areas. The most basic set are the ‘Key Statistics’, which are accompanied by ‘Quick Statistics’ which provide more detailed breakdowns for each variable. These were published between December and February in England and Wales, in January in Northern Ireland and are scheduled for March in Scotland.

More detailed two and three-dimensional tabulations are published in the form of Local Characteristics and Detailed Characteristics. Local characteristics tables are mainly based upon those produced for the 2001 Census. Not all Detailed Characteristics tables have been specified at present. Unusually, the publication schedule for the 2011 Census has experienced numerous changes and there have been revisions to a number of tables, which have had to be re-released

In England and Wales, detailed Characteristics tables will be published at local authority level in the third release of Census data starting in May 2013, with data for Middle Super Output Areas and electoral wards to follow. Local Characteristics tables will be published in August-September 2013.

The first Census results for Northern Ireland were published a little later than in England and Wales. The publication of Detailed Characteristics tables is scheduled to start in May 2013, with Local Characteristics published during the summer. The publication schedule for Scotland is later. The first population counts for local authority districts were published in March 2013. The release of Key & Quick Statistics tables will commence in Summer 2013, while Local Characteristics tables will be released from Autumn 2013, and Detailed Characteristics tables published in Winter 2013.

Once the main publication effort is completed (in late 2013), the Census Offices will publish flow matrices for journey-to-work and migration, microdata and UK-wide tables. The flow tables provide considerable geographical detail, but are usually limited to industry sections and SOC major groups. The microdata datasets are based on a small sample of Census returns, but include answers to the original Census questions, recoded to the range of classifications used for publication. It is possible to cross-tabulate any variable by any other variable. However, the detail of some variables (notably geography) is limited. The level of detail reduces as the dataset becomes more easily available. The most detailed version of Census microdata is only accessible via the Secure Data Service, all outputs from which must be checked by ONS to ensure that they do not disclose information about identifiable individuals. Since it is based on a sample of the data (typically 3 to 5 per cent), tables generated from Census microdata are also subject to sampling error. The Census Offices will also produce bespoke tables commissioned by users of the Census. There is a charge for this service.

Labour market and employment data from the Census

The labour market-related data available from the Census is derived from questions 26-31, 33-38 and 40 in England and Wales (see Table 1). There are also two questions (40 and 41) on travel-to-work (see Table 2). Question 26 asks about economic activity in the week before the Census. The response rate to this question was 94.9 per cent – the missing 5.1 per cent of responses were imputed. Industry is derived from question 37 on the ‘main activity of your employer or business’. Occupation is derived from questions 34 and 35.

Three types of information on employment will be published:

- Employment characteristics of people resident in an area;
- Characteristics of people working in an area;
- Information on travel patterns of people in work. It will be possible to identify where jobs located in a particular location draw workers from and identify where people living in a particular place work (in each case by industry or occupation).

The Census also yields a large amount of information on the labour force and general labour market conditions. This includes:

- Labour market participation and participation rates. The number of people in each labour market state as a percentage of the population. This can be disaggregated by age and gender.
- Unemployment rates. This can be calculated by age and gender. The question on previous occupation and industry can be combined with current employment by industry and occupation to yield unemployment rates by occupation and industry.
- Long-term unemployment rates by age and gender.

The majority of labour market tables are produced for the population aged 16 to 74. However, Local and Detailed Characteristics tables include breakdowns by age and gender.

This information is available for electoral wards (above a specified population size threshold) and larger geographical areas.

Tables currently available for England and Wales and Northern Ireland and soon to be available for Scotland are:

Key Statistics tables include:

- Economic activity KS601 (persons) KS602 (males) KS603 (females)
- SIC 2007 industry section KS605 (persons) KS606 (males) KS607 (females)
- SOC 2010 major group KS608 (persons) KS609 (males) KS610 (females)
- Highest qualifications KS501 (persons)

Quick Statistics tables:

- SIC 2007 industry section (with more detail for manufacturing) QS605
- SOC 2010 minor group QS606
- Highest level of qualification QS501 (levels 1 to 4, apprenticeships and other)
- Qualifications gained QS502 (more detailed breakdown)
- Economic activity QS601
- Hours worked QS604
- Year last worked QS612
- Method of travel to work QS701

The *indicators* which can be calculated relate to the labour market characteristics of the population and employment indicators measure the characteristics of people living in an area who are working.

These include:

- Economic activity rate by gender
- Employment rate by gender
- Full-time/part-time working by gender
- Unemployment rate by gender
- Percentage of working age population qualified to level 3 or higher
- Percentage of working age population with poor or no qualifications
- Occupational profile of employment by gender
- Industry profile of employment by gender
- Percentage of people using public transport to commute

Forthcoming tables from the ***Local Characteristics and Detailed Characteristics*** releases include cross-tabulations of the variables listed above by age, gender and ethnic group. There is also a cross-tabulation of occupation (sub-major group) by industry section by gender by residence of worker and a cross-tabulation of SOC major group by industry section by location of workplace.

The tables available in ***Detailed Characteristics*** include:
Sex and age by economic activity
Sex and age by employment last week and hours worked
Sex and economic activity by living arrangements
Sex and Age and Highest Level of Qualifications by Economic Activity
Sex and occupation by age
Sex and former occupation by age
Sex and occupation by employment status and hours worked
Sex and industry by age
Sex and former industry by age
Sex and industry by employment status and hours worked
Occupation by industry
Sex and occupation by hours worked
Sex and economic activity and year last worked by age
Economic activity and age of full-time students by household type and tenure
Sex and age by highest level of qualification
Sex and age and economic activity by ethnic group
Sex and occupation by ethnic group
Sex and industry by ethnic group
Sex and occupation by highest level of qualification
Count of qualifications by sex
Age and highest level of qualification by ethnic group
Number of employed people and method of travel to work by number of cars or vans in

household
Sex and age by method of travel to work
Sex and NS-SeC by method of travel to work
Sex and occupation by knowledge of Welsh/Gaelic/Irish
Welsh/Gaelic/Irish speakers and economic activity and year last worked by age
Sex and industry by knowledge of Welsh/Gaelic/Irish
Age and highest level of qualification by knowledge of Welsh/Gaelic/Irish
Sex and age and economic activity by religion
Sex and occupation by religion
Sex and industry by religion
Age and highest level of qualification by religion

Local Characteristics tables include:
Sex and age by economic activity
Sex and age by hours worked
Sex and age and highest level of qualification by economic activity
Sex and occupation by age
Former occupation by age
Sex and age and occupation by employment status and hours worked
Sex and industry by age
Former industry by age
Sex and industry by employment status and hours worked
Occupation by industry
Sex and occupation by hours worked
Economic activity and time since last worked by age
Economic activity and age of full-time students by household type
Age by highest level of qualification
Occupation by highest level of qualification
Sex and age by method of travel to work
Sex and distance travelled to work by method of travel to work
Economic activity by number of cars and vans
Employment status by number of cars and vans
Occupation by economic activity

Possible *indicators*:

- Economic activity rate by gender and age (and by gender/ethnic group and gender/religion)
- Employment rate by gender and age (and by gender/ethnic group and gender/religion)
- Full-time/part-time working by gender and age (and by gender/ethnic group and gender/religion)
- Unemployment rate by gender and age (and by gender/ethnic group and gender/religion)

- Percentage of working age population qualified to level 3 or higher
- Percentage of working age population with poor or no qualifications
- Occupational profile of employment by gender
- Industry profile of employment by gender
- Percentage of people using public transport to commute

A few of these tables present the characteristics of jobs located in a particular area. In particular, there is a cross-tabulation of industry section by occupation sub-major group for people working in an area. This table is available for local authority districts and larger geographical areas.

Possible *indicators*:

- Occupational profile (SOC major group) of employment located in an area
- Industrial profile (section) of employment located in an area

Commuting and workplace data

The bulk of data on the employment characteristics of workplaces will become available from the journey-to-work tables, which are produced toward the end of the publication process (probably late 2013 or early 2014). A set of tables will document the characteristics of workers resident in a location, working in a location and involved in each flow between pairs of locations. These tables use the standard Census geography for the residence of workers, and a new 'workplace geography' for the characteristics of people working in an area. This geographical framework is more detailed than the standard Census geography in areas where employment is concentrated. Thus, it may be possible to use Census data to identify the types of jobs located in particular industrial estates or retail/office centres and the travel behaviour of their workers. This data would allow job seekers to identify what kinds of job were located within their travel horizon (which other sources of employment data cannot do).

The tables available will detail:

- Employment by age and gender
- Employment by occupation (SOC sub-major group)
- Employment by industry (section)
- Employment by NS-SEC
- Employment by highest qualification
- Employment by ethnic group
- Employment by disability
- Employment by mode of travel

Possible *indicators*:

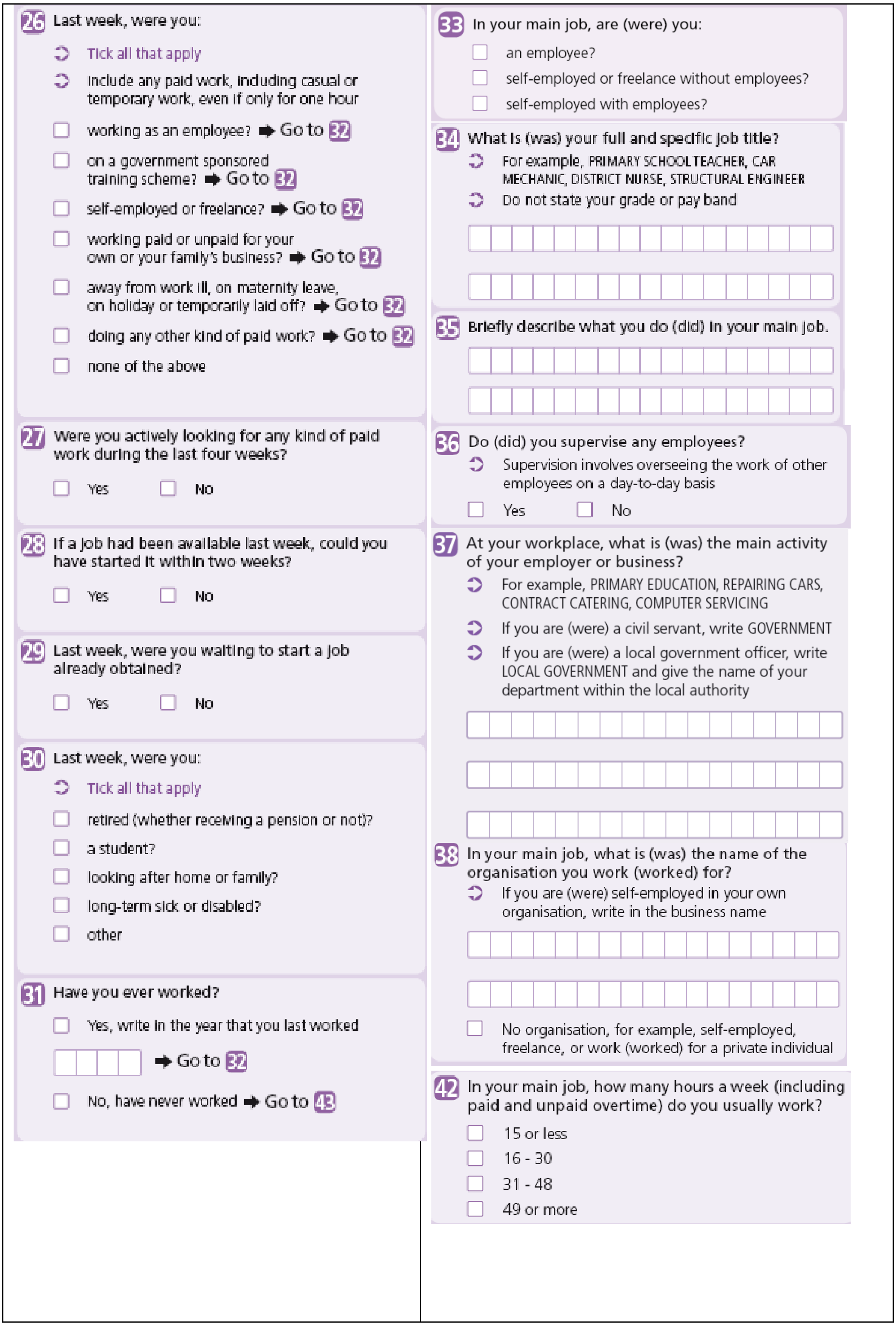
At the time of writing (May 2013), full detail of the Census data on employment by workplace is not yet available. On previous experience, the Census Offices will release 'trip-end' and 'flow' tables. The former detail the characteristics of workers resident in and working in geographical areas. The flow tables present the breakdown of workers involve in each commuting flow. It is possible to calculate the distance travelled from the geographical centroid of each geographical area involved in a commuting flow and hence detailed statistics about the distance travelled to work by workers of different types living in an area or the distances travelled to jobs located in a particular area can be calculated. Trip-end tables published from the 2001 Census presented the characteristics of employees and students by age, gender, family status, mode of travel, SOC major group, NS-SEC, industry sector and ethnic group. Flow tables only provided data on mode of travel. More detail was available for commuting flows between local authority districts than between small areas. The level of detail available was severely constrained because of fears over confidentiality, because of the small number of jobs located in many residential areas. The new 2011 workplace geography reflects the geography of employment, enabling much more detail on the characteristics of employment to be released. The list below lists the sort of information likely to be available. Unfortunately, there is as yet no timetable for the publication of this data.

The majority of the data is likely to be comparable across the UK, but Scotland and Northern Ireland tend to publish a slightly different range of tables to England & Wales.

- Occupational profile (at least SOC2010 major group in all areas and potentially 3-digit SOC in larger geographical areas) of employment located in an area
- Industrial profile (at least SIC 2007 section) of employment located in an area
- Qualification (NQF) profile of employment located in an area
- Distance travelled to work (e.g. 5, 10, 20, 50, 50+ kilometre bands or median distance) in a location by occupation
- Distance travelled to work (e.g. 5, 10, 20, 50, 50+ kilometre bands or median distance) in a location by qualification (NQF) level
- Occupational profile (at least SOC major group) of jobs available within a given commuting distance of a home postcode
- Industrial profile (at least SIC 2007 section) of jobs available within a given commuting distance of a home postcode
- Qualification (NQF) profile of jobs available within a given commuting distance of a home postcode

The key advantage of the Census data is the provision of data for small geographical areas and the information it provides on the distance workers have to travel to different types of job. As is the case for all Census data, the main disadvantage is the fact that it refers to a single point of time, and is published more than two years after the data was collected.

Figure C.1 Labour market questions in 2011 Census of Population



26 Last week, were you:

- Tick all that apply
- Include any paid work, including casual or temporary work, even if only for one hour

□ working as an employee? ➡ Go to **32**

□ on a government sponsored training scheme? ➡ Go to **32**

□ self-employed or freelance? ➡ Go to **32**

□ working paid or unpaid for your own or your family's business? ➡ Go to **32**

□ away from work ill, on maternity leave, on holiday or temporarily laid off? ➡ Go to **32**

□ doing any other kind of paid work? ➡ Go to **32**

□ none of the above

27 Were you actively looking for any kind of paid work during the last four weeks?

□ Yes □ No

28 If a job had been available last week, could you have started it within two weeks?

□ Yes □ No

29 Last week, were you waiting to start a job already obtained?

□ Yes □ No

30 Last week, were you:

- Tick all that apply

□ retired (whether receiving a pension or not)?

□ a student?

□ looking after home or family?

□ long-term sick or disabled?

□ other

31 Have you ever worked?

□ Yes, write in the year that you last worked

□□□□ ➡ Go to **32**

□ No, have never worked ➡ Go to **43**

33 In your main job, are (were) you:

□ an employee?

□ self-employed or freelance without employees?

□ self-employed with employees?

34 What is (was) your full and specific job title?

- For example, PRIMARY SCHOOL TEACHER, CAR MECHANIC, DISTRICT NURSE, STRUCTURAL ENGINEER
- Do not state your grade or pay band

35 Briefly describe what you do (did) in your main job.

36 Do (did) you supervise any employees?

- Supervision involves overseeing the work of other employees on a day-to-day basis

□ Yes □ No

37 At your workplace, what is (was) the main activity of your employer or business?

- For example, PRIMARY EDUCATION, REPAIRING CARS, CONTRACT CATERING, COMPUTER SERVICING
- If you are (were) a civil servant, write GOVERNMENT
- If you are (were) a local government officer, write LOCAL GOVERNMENT and give the name of your department within the local authority

38 In your main job, what is (was) the name of the organisation you work (worked) for?

- If you are (were) self-employed in your own organisation, write in the business name

□ No organisation, for example, self-employed, freelance, or work (worked) for a private individual

42 In your main job, how many hours a week (including paid and unpaid overtime) do you usually work?

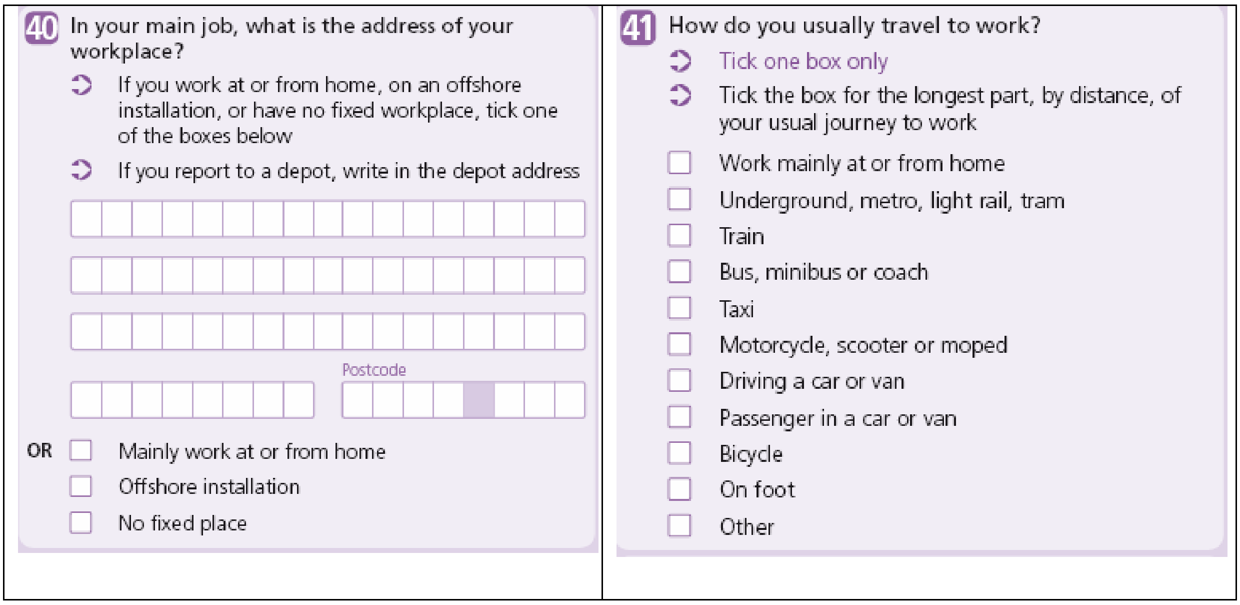
□ 15 or less

□ 16 - 30

□ 31 - 48

□ 49 or more

Figure C.2 Journey-to-work questions in 2011 Census of Population



40 In your main job, what is the address of your workplace?

- If you work at or from home, on an offshore installation, or have no fixed workplace, tick one of the boxes below
- If you report to a depot, write in the depot address

Postcode

OR □ Mainly work at or from home

□ Offshore installation

□ No fixed place

41 How do you usually travel to work?

- Tick one box only
- Tick the box for the longest part, by distance, of your usual journey to work

□ Work mainly at or from home

□ Underground, metro, light rail, tram

□ Train

□ Bus, minibus or coach

□ Taxi

□ Motorcycle, scooter or moped

□ Driving a car or van

□ Passenger in a car or van

□ Bicycle

□ On foot

□ Other

C.6 *Working Futures* projections at 4-digit level

The current published *Working Futures* database provides information by occupation at a 2-digit level of SOC2010 (although 3-digit level results have also been produced occasionally, linked to additional work on the qualifications dimension).

In principle, more detailed projections are technically feasible but this is limited by the quality of the available data upon which the analysis is based (primarily the LFS).

In the prototype database the possibility of using common growth factors applied to all occupations within a 2-digit category (i.e. assuming fixed shares) was explored but not fully implemented. This has been taken to an operational level in Phase 2A.

The LFS enables reasonably robust estimates of the shares of employment in SOC 2-digit categories that are employed in the 4-digit categories they contain at the all industry level.

These have been used in Phase 2A to expand the *Working Futures* database to the 4-digit level based on an assumption of these shares being fixed. Similar assumptions have been imposed to generate estimates of replacement demands at the 2 and 4-digit levels.

The use of fixed shares provides an indication of the patterns of jobs available that is useful from a careers guidance perspective. However it does not make use of information on how these shares are changing over time. Although it is not possible to carry out such an analysis across all dimensions of employment simultaneously at a more aggregate level, these changing patterns over time can be explored and used to develop a more refined view of changing occupational structure over the next decade at the 4-digit level. This should be explored in Phase 2B or as part of the next round of *Working Futures* projections.

C.7 Cedefop database

Cedefop publish a range of skills demand and supply projections that are available in the public domain.[34] IER is the lead organisation responsible for producing these results. For the past 5 years IER, in collaboration with others, have developed an historical employment database and projections at a pan-European level on behalf of Cedefop. This replicates many of the same features of the *Working Futures* employment database. Details can be found at: http://www.cedefop.europa.eu/EN/about-cedefop/projects/forecasting-skill-demand-and-supply/skills-forecasts.aspx These estimates are based on the ELFS (see below) plus some other data. They provide a consistent historical as well as a forward looking dataset that could be exploited in the LMI for All project.

The Cedefop data could be used to add a European dimension to the assessment of future job prospects to complement the information available for the UK from *Working Futures*.

The Cedefop data are presented using ISCO88 2-digit categories. In Phase 2A the team have explored the feasibility of developing a suitable mapping to the SOC2010 categories and the overall practicality of adding this information to the database.

Some of the data are available on line. More detailed information is available to Cedefop *Skillsnet* members in the form of Excel Workbooks. IER has access to the full database and can supply it in a more user friendly form for the *LMI for All* project.

In principle, the data can be used to generate employment information, including replacement demands, for each of the 27 EU Members States plus a few additional countries such as Norway and Switzerland.

In practice, there are a few issues:

- The data are currently classified using ISCO 88 which is not directly comparable with SOC 2010 – However, a broad brush mapping can be derived (see note below).
- The new data to be published in 2013/2014 will use ISCO 08. This is broadly compatible with SOC 2010. IER and ONS have been working on developing mappings (see note below).
- The current Cedefop projections are primarily focused on the 2-digit level. Development of information at a more detailed level is being explored, but data limitations are problematic. Information at a 4-digit level is unlikely to be available in the foreseeable future

Recommendations

- Use currently available 2-digit information, based on ISCO88, adopting a broad brush mapping to SOC 2010 2-digit categories in the short term.
- Move to the revised ISCO08 data as soon as they are available (early 2014) and

[34] See http://www.cedefop.europa.eu/EN/about-cedefop/projects/forecasting-skill-demand-and-supply/index.aspx

exploit more detailed information if and when it is published.

Note: SOC2010 – ISCO08 Mapping

Full details of current ONS thoughts on mapping between SOC 2010 and ISCO08 are on the ONS website at: http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/index.html

The international 'Resolution Concerning Updating the International Standard Classification of Occupations' coordinated by the International Labour Office (ILO) resolved on the 6th December 2007 to update ISCO88. The resolution stated:

Each country collecting and processing statistics classified by occupation should endeavour to compile data that can be converted to ISCO08, to facilitate the international use and comparison of occupational information.

Each country should provide information to the ILO about how the groups defined in national classification(s) of occupations can best be related to ISCO08.

ONS have developed a crude mapping at: http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-to-isco08-mapping.xls

Wherever possible the 369 SOC2010 Unit Groups have been mapped to one ISCO08 Unit group. However, in certain cases this has not been possible.

- A few SOC2010 Unit Groups shows have been mapped to two ISCO08 Unit Groups on a 50:50 split: (15/30);
- and Armed Forces are divided into 2 (40/60);
- There are 145 ISCO 4-digit codes with no direct match to SOC 2010 (IER/ONS are working on this, see below).

IER are currently working with ONS on preparing a more detailed mapping. ONS has assigned codes on index entry level and IER are making suggestions to improve it. Table C.1 below shows an example of the work being done. Once completed ONS will consider the suggestions, and adapt some (or all) of them. It is clear that the mapping process is challenging and that a simple solution is not likely in the foreseeable future.

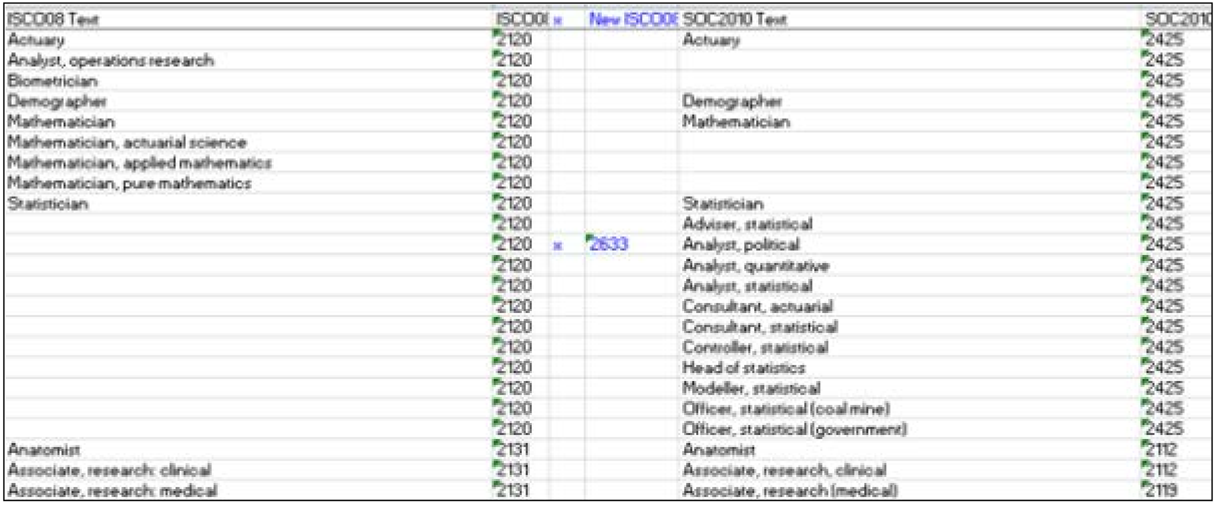
A crude probability mapping from SOC2010 to ISCO 88 has also been developed by ONS (but this simply assumes the same 1:1, 50:50 split or 40:60 split as set out above).

Previously SOC2000 was mapped to ISCO88COM (a European variant of ISCO 88). There is no mapping from the old SOC2000 and ISCO88 classifications to the new ones.

A broad brush mapping from the old ISCO88 categories to SOC2010 categories at a 2-digit level is possible but users would need to be advised that this is approximate. This is probably adequate for the purpose of careers guidance and advice where the aim is to provide general information on the type of jobs likely to be available rather than a precise picture of employment numbers.

Table C.1 Mapping from ISCO08 to SOC2010

1	2	3	4	5



ISCO08 Text	ISCO0	×	New ISCO0	SOC2010 Text	SOC201
Actuary	2120			Actuary	2425
Analyst, operations research	2120				2425
Biometrician	2120				2425
Demographer	2120			Demographer	2425
Mathematician	2120			Mathematician	2425
Mathematician, actuarial science	2120				2425
Mathematician, applied mathematics	2120				2425
Mathematician, pure mathematics	2120				2425
Statistician	2120			Statistician	2425
	2120			Adviser, statistical	2425
	2120	×	2633	Analyst, political	2425
	2120			Analyst, quantitative	2425
	2120			Analyst, statistical	2425
	2120			Consultant, actuarial	2425
	2120			Consultant, statistical	2425
	2120			Controller, statistical	2425
	2120			Head of statistics	2425
	2120			Modeller, statistical	2425
	2120			Officer, statistical (coal mine)	2425
	2120			Officer, statistical (government)	2425
Anatomist	2131			Anatomist	2112
Associate, research: clinical	2131			Associate, research, clinical	2112
Associate, research: medical	2131			Associate, research (medical)	2119

Notes:

1. ISCO08 index entries
2. ISCO08 code (assigned by ONS for SOC-only index entries)
3. IER'ss suggestion for ISCO08 code change
4. SOC2010 index entries, matched to ISCO08 entries where possible by ONS
5. SOC2010 code (assigned by ONS for ISCO-only index entries)

Table C.2 Map from ISCO 88 to SOC2010 at 2-digit level

ISCO88 Categories as used in Cedefop Projections	2010			SOC2010 categories as used in *Working Futures*	***2010***
11 Legislators and senior officials	55	1.1	(	11 Corporate managers and directors	2,015
12 Corporate managers	3,764	1.1	(		
13 Managers of small enterprises	1,177	1.2		12 Other managers and proprietors	1,000
21 Physical, mathematical and engineering science	1,284	2.1		21 Science, research, engineering and technology professionals	1,593
22 Life science and health professionals	403	2.2		22 Health professionals	1,296
23 Teaching professionals	1,270	2.3		23 Teaching and educational professionals	1,364
24 Other professionals	1,496	2.4		24 Business, media and public service professionals	1,591
31 Physical and engineering science associate professionals	748	3.1		31 Science, engineering and technology associate professionals	501
32 Life science and health associate professionals	965	3.2		32 Health and social care associate professionals	323
33 Teaching associate professionals	178	3.3-		34 Culture, media and sports occupations	569
34 Other associate professionals	2,350	3.3-		35 Business and public service associate professionals	2,074
41 Office clerks	2,869	4.1		41 Administrative occupations	2,738
42 Customer services clerks	942	4.1		42 Secretarial and related occupations	961
51 Personal and protective services workers	3,455	6.1	}	33 Protective service occupations	458
			}	61 Caring personal service occupations	2,094
			}	62 Leisure, travel and related personal service occupations	625
			}	72 Customer service occupations	617
52 Models, salespersons and demonstrators	1,683	7.1		71 Sales occupations	1,991
61 Skilled agricultural and fishery workers	436	5.1		51 Skilled agricultural and related trades	399
71 Extraction and building trades workers	1,450	5.3		53 Skilled construction and building trades	1,152
72 Metal, machinery and related trades workers	875	5.2		52 Skilled metal, electrical and electronic trades	1,330
73 Precision, handicraft, craft printing and related trades	114	5.4	}	54 Textiles, printing and other skilled trades	645
74 Other craft and related trades workers	149	5.4	}		
81 Stationary plant and related operators	145	8.1	}	81 Process, plant and machine operatives	822
82 Machine operators and assemblers	575	8.1	}		
83 Drivers and mobile plant operators	1,073	8.2		82 Transport and mobile machine drivers and operatives	1,128
91 Sales and services elementary occupations	2,258	9.2		92 Elementary administration and service occupations	2,628
92 Agricultural, fishery and related labourers	136	9.1	}	91 Elementary trades and related occupations	544
93 Labourers in mining, construction, manufacturing and	1,140	9.1	}		
All occupations	31,049			All occupations	30,458

C.8 Other European datasets

In principle, there are a number of pan-European datasets that might be useful to add to the LMI for All database. These include:

1. European Labour Force Survey (ELFS);
2. Other surveys including:
    a. Eurofound survey of living and working conditions;
    b. Eurobarometer;
    c. European Values Survey; and
    d. European Social Survey

These are briefly summarised here.

In practice, although they contain some interesting and useful data they are generally not suitable for including in the database because the sample sizes are inadequate to provide reliable data at a detailed and consistent level by occupation.

They would have more value if the database were to be extended to cover the needs of other users such as more general labour market analysts.

European Labour Force Survey (EFLS)

General description of the dataset

The European Union Labour Force Survey (EU LFS) is conducted in the 27 Member States of the European Union, 3 candidate countries and 3 countries of the European Free Trade Association (EFTA) in accordance with Council Regulation (EEC) No. 577/98 of 9 March 1998. At the moment, the LFS microdata for scientific purposes contain data for all 27 Member States and in addition Iceland, Norway and Switzerland.

The EU LFS is a large household sample survey providing quarterly results on labour participation of people aged 15 and over as well as on persons outside the labour force. All definitions apply to persons aged 15 years and over living in private households. Persons carrying out obligatory military or community service are not included in the target group of the survey, as is also the case for persons in institutions/collective households.

The national statistical institutes are responsible for selecting the sample, preparing the questionnaires, conducting the direct interviews among households, and forwarding the results to Eurostat in accordance with the common coding scheme.

The data collection covers the years from 1983 onwards. In general, data for individual countries are available depending on their accession date.

The Labour Force Surveys are conducted by the national statistical institutes across Europe and are centrally processed by Eurostat:

- Using the same concepts and definitions
- Following International Labour Organisation guidelines
- Using common classifications (NACE, ISCO, ISCED, NUTS)
- Recording the same set of characteristics in each country

In 2011, the quarterly LFS sample size across the EU was about 1.5 millions of individuals. The EU-LFS covers all industries and occupations.

A significant amount of data from the European Labour Force Survey (EU LFS) is also available in Eurostat's online dissemination database, which is regularly updated and available free of charge. The EU LFS is the main data source for the domain ‘employment and unemployment’ in the database. The contents of this domain include tables on population, employment, working time, permanency of the job, professional status etc. The data is commonly broken down by age, sex, education level, economic activity and occupation where applicable.

Several elements of indicator sets for policy monitoring are also derived from the EU LFS and freely available in the online database. The structural indicators on employment include the employment rate, the employment rate of older workers, the average exit age from the labour force, the participation in life-long learning and the unemployment rate. The sustainable development indicators also include employment rates by age and educational attainment as well as the population living in jobless households and the long-term unemployment rate.

Data made available via Eurostat are annoymised by suppression if necessary.

Microdata from the ELFS is available from Eurostat but confidentiality concerns mean that access to the data is tightly controlled, many variables are not available in all countries and limited detail is made available on sensitive variables. Publically available data are available in xls format to download from the Eurostat website. The standardisation of the data means that it could be integrated in to the Careers LMI database providing a European perspective on employment, unemployment rates, workforce characteristics, educational attainment and earnings. Because of concerns about confidentiality and statistical robustness Eurostat only make the data available in restricted format. These data would, therefore, need to be presented at an aggregated industry, occupational and regional level. See: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/lfs

The *recommendation* for the EULFS is that the European LFS should not be included in LMI for All since national employment data from the LFS on the Eurostat website are limited to the ten-fold ISCO classification of occupations and the microdata are not suitable for accessing for this purpose.

Other regular European surveys (such as the Eurobarometer, the European Values Survey and European Social Survey and the European Working Conditions survey) can also provide contextual information on issues such as attitudes towards labour migrants in different countries. working conditions, etc.

Eurofound Working Conditions Survey

The European Working Conditions Survey provides an overview of working conditions in Europe. It assesses and quantifies working conditions of both employees and the self-employed across Europe on a harmonised basis, including:

- Analysis of relationships between different aspects of working conditions;
- Identification of groups at risk and issues of concern, as well as of progress;
- Monitoring of trends by providing homogeneous indicators on these issues;
- Contributing to European policy development.

The scope of the survey questionnaire has widened substantially since the first edition in 1990, aiming to provide a comprehensive picture of the everyday reality of men and women at work.

Themes covered today include gender equality, employment status, working time duration and organisation, work organisation, learning and training, physical and psychosocial risk factors, health and safety, work-life balance, worker participation, earnings and financial security, as well as work and health.

In each wave a random sample of workers (employees and self-employed) has been interviewed face to face. Following the European enlargements the geographical coverage of the survey has expanded to now cover the whole if the EU plus a number of neighbouring and accession countries.

While very interesting from a general labour market analysis perspective it is of less relevance in a careers guidance and advice context. It is also based on a relatively small sample (around 44,000 across all countries covered) which means that it is unable to produce any detailed data by occupation. Consistent classification is also an issue (SOC/ISCO, see discussion above under Cedefop).

As a result the recommendation is that it should NOT be included on grounds of:

- Lack of relevance;
- Small sample size.

The same applies to the remaining surveys discussed below.

European Social Survey

The European Social Survey (the EurSS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. The EurSS was established in 2001.

Currently in the midst of its sixth round, this biennial cross-sectional survey covers more than thirty nations and employs the most rigorous methodologies. The EURSS information brochure outlines the origins and development of the project. In addition two collections of findings are available: one summarises key findings from the first three rounds of the survey; the other focuses on 'topline' results relating to Trust in Justice data collected in round five.

Eurobarometer

This is a series of public opinion surveys and reports undertaken for the European Commission. It focuses on issues relating to the European Union member states, with a sample size of around 1000 in each country. A longitudinal element enables the tracking and comparison of public opinions on, for example, gender roles, family, youth, elderly, immigration, regional identity, science and technology and working conditions over time.

The topic/focus of the survey changes. One of the recent concerns of the survey has been labour migration and mobility in Europe and it is possible to identify recent trends in the types of individual willing to work in another country and the types of work they undertake. This survey could not be linked in a formal manner to other data sources. Instead, it would provide useful contextual background information.

European Values Study

The European Values Study is a large-scale, cross-national, and longitudinal survey research program on basic human values. It provides insights into the ideas, beliefs, preferences, attitudes, values and opinions of citizens all over Europe. It is a unique research project on how Europeans think about life, family, work, religion, politics and society.

The European Values Study started in 1981, when a thousand citizens in the European Member States of that time were interviewed using standardized questionnaires. Every nine years, the survey is repeated in an increasing number of countries. The fourth wave in 2008 covers no less than 47 European countries/regions, from Iceland to Azerbaijan and from Portugal to Norway. In total, about 70,000 people in Europe are interviewed.

A rich academic literature has been created around the original and consecutive surveys and numerous other works have made use of the findings. In-depth analyses of the 1981, 1990 and 1999 findings with regard to Western and Central Europe, and North America reinforced the impression that a profound transformation of modern culture is taking place, although not at the same speed in all countries. Cultural and social changes appear dependent upon the stage of socio-economic development and historical factors specific to a given nation. The latest wave provides further insights in this matter.

As with Eurofound Working Conditions Survey the limited sample size and lack of immediate relevance suggest that none of these surveys should be a priority for inclusion in the LMI for All database.

C.9 Course Information

Information and data on courses and training available across the UK are an important element in a database focused on careers guidance and advice. Unfortunately this is not held in any one central database.

Compiling a comprehensive list of further and higher education training and courses is complex, not least due to the number and range of courses available. Accessing such data and incorporating it in to the LMI for All database will require a comprehensive mapping of courses to occupational codes. These issues are discussed in more detail in Section 2.3.10 of the main report.

C.10 HESA and related data

Information and data on the passage of individuals through the higher and further education system are another important element in a database focused on careers guidance and advice.

There are a number of sources of relevant information, including the first destinations of graduates from higher educational establishments. For example, the data HESA collect in their graduate destination survey are classified using the SOC classification and in principle allow mapping from courses studied to job destination. Currently much of this kind of information is not freely available (and often is subject to a charge).

Annex D: Use cases for hack day developers

The following use cases were given to the hack day developers to illustrate the types of career-related issues for which individuals require high quality, impartial labour market information to progress their career decision-making. These illustrative cases provided the developers with some context to the database and potential uses.

- Robin, 15, is at secondary school and is considering a career in nursing. He is wondering what training is involved, how he will pay for it and whether there will be job opportunities for nurses in his home area when he finishes his training.
- Jo, 27 is an electronic engineering technician who is unemployed and having difficulty finding a job in her local area. She has already checked out the local job centre, job websites and newspapers for any current job vacancies posted for electronic engineering technicians. Although Jo would like stay in the local area, she is willing to travel to find work.
- Liam, 42 is an unemployed bricklayer. He has been checking the local newspapers, reviewing various job posting Web sites and talking to family and friends, but has not yet found anything suitable. He has just heard about a major construction project starting in the area and wants more information.
- Angela, 37, is a qualified and experienced pharmacist who wants to immigrate to Canada, but is not sure what information she needs or where to start her research.
- Kola is a Year 12 student at school, thinking about applying to university. When he graduates, he wants to find a job near home. No-one else in his family has gone into higher education and his father says that university is too expensive. He wants Kola to leave school at 18 and get a job in a local hotel as junior manager. Kola wants to know how to get into ICT.
- James is 18, unemployed and lives in a rural area. He left school at 16 with low exam grades. He has been on several different training schemes, but has not added to his qualifications. James wants a job on a building site, driving a bus or in a shop – but is not willing to consider ‘low paid jobs’. Public transport is expensive and limited. He has heard that Tesco are building a new distribution depot and would like more information, but does not have a current CV.
- After taking time out of education for two years to save up before going to University, Caroline studied Sociology and achieved a 3rd class Honours degree. She is unsure what options are open to her. Before starting her course, Caroline worked in a variety of different temporary jobs, such as waiting / bar staff, support worker at a Youth Centre and in a shoe shop. She continued with bar work whilst at University. Caroline is able to identify the skills she acquired whilst at University: confidence; independence; initiative and the ability to work either by herself or with others.

- Russell is an employer looking to establish a business in the area. He is looking for current wage rates, a list of similar employers located in the area and some general information on the local economy.

- Gillian, 43, left school with three ‘A’ levels and trained in Beauty Therapy. After practicing for several years, she became dissatisfied and re-trained as an acupuncturist. She then worked freelance for three years and then stopped work to have a family. After an absence of nine years from the labour market, she needs to return to work to contribute to the family budget.

- Priti is in Year 11 at school, with predicted ‘C’ grades for most subjects at GCSE, with ‘A’ predicted for art. She loves animals and her parents think she should be a vet, because she would get good money and could live in the country. Recently, she did her work experience in a solicitor’s office, which she enjoyed. Because she enjoys television programmes involving legal cases, she is now thinking she might like to be a solicitor. She wants to know what subjects she would need at ‘A’ level.

- Tom was made redundant last year (for the fourth time). He’s 57 years old, but not ready to finish work. He left school at 16 and started as a craft engineering apprentice. He did Technician1 & 2 alongside the craft apprenticeship, but did not qualify as a technician. Over his career, he has worked in purchasing and management – each time training and qualifying for the work he undertook. Now he’s past 55 years old, people assume he’s looking to retire, so don’t take him seriously.

Annex E: Summary of careers stakeholder pre-event questionnaire

All careers stakeholders who attended the hack day were asked to complete a pre-event questionnaire. This provided the developers with important contextual information and an understanding of their use of LMI and access to technology.

How do they use LMI in their job?

- Used in delivery of careers guidance services: supporting individuals to make informed decisions about career paths and understand the changing world of work
- Background research for presentations, bulletins, workshops, resources, create learning resources e.g. to introduce sector/professional events for students, reconcile aspirations with employer needs
- Research (local and national data) and disseminate LMI to colleagues and wider audiences
- Explore how LMI is currently being used and possible future application in services
- Analyse destinations (DLHE) data for performance monitoring and planning at College, School/Department and programme level
- Analysis of national DLHE data for competitor analysis
- Use LMI in policy work to try and understand the numbers of technicians in the workforce, which sectors they work in and their education levels
- Inclusion in online materials (i.e. jobs and regional pages)

What problems do they currently experience with accessing LMI?

- Accurate, reliable and impartial LMI data is hard to source
- Finding up-to-date data as labour markets often change faster than source data are updated
- Understanding geographical coverage (local, national, international)
- Understanding how to deliver LMI and make it appealing
- Understanding and reconciling contradictory information and being able to compare data from different sources
- Data not readily available, not complete or consistent, i.e. not all industry data are available, data not disaggregated to required level
- Lack of data on future trends
- Lack of understanding of the importance of LMI
- No time for LMI CPD

- Unable to access data for an individual job profile and macro level (i.e. by sector and county)
- Language can be difficult for clients

What are the most common sources of LMI currently used?

- Colleagues (including their reports from careers visits and conferences)
- Local knowledge and newspapers
- Online sources, articles and databases, i.e. Going Global, Prospects, DLHE datasets, NOMIS
- Employer engagement
- Social media, i.e. Twitter
- Government, i.e. ONS, JCP for vacancies
- Employer organisations, i.e. CBI, Birmingham Chamber, SSCs
- Professional bodies, i.e. Law Society, ICG, Montrose Public Affairs Consultants Ltd, local Job Centre LMI newsletter
- Research centres, i.e. IER
- National Careers Service website (in the past Jobs4U directory)
- Regional skills observatories and other regional sources

In an ideal world, they would love to be able to access LMI on…

- One portal accessible on a laptop or phone
- A local, county, regional, national and international level
- Local data and trends by sector and role, by gender, age and ethnicity and socio-economic
- Application and growth rates for apprenticeships
- Industry/employment forecasts
- Local vacancies
- Job opportunities, i.e. opportunities for mature job changers, part-time workers, local young people, those without a degree
- Occupations by employer size/type, opportunity type (e.g. graduate schemes, other entry level job)
- Occupations by qualification level (i.e. proportion with degrees, professional qualifications etc.)

- Profession-based LMI e.g. lawyers by employer type, practice specialisation, etc.
- Forecasts and trends, especially data that could be interrogated by rate of change over time (years/decades) showing emergent patterns in employment by sector/role to detect growth and decline within areas
- Linked data – careers videos, articles and data

What technology is used in their job?

- Devices are mainly owned by the employer, but two individuals also use their own devices
- Majority use a combination of devices for their job
- Majority use a pc or laptop, one individual uses a dated notebook with a dongle
- A few have access to smart phones (android, blackberry and iphone)
- Three use a tablet/ipad

Most are permitted to install apps, but this would depend on context and would need to be arranged with their IT department. One reported that they would not be able to install apps, whilst others were unsure or did not have a device for apps.

References

Bimrose, J. (2012). *Proposal for 'Developing a Careers LMI Data Tool', Research and Evaluation Framework Agreement, Category 3 – Programme and Pilot Evaluation, Department for Business Innovation and Skills, On behalf of the UK Commission for Employment and Skills*, 12th October 2012. Coventry: Institute for Employment Research, University of Warwick.

Bimrose, J. and Wilson, R. (2013a). *Data Development Plan: Pay and Employment.* Institute for Employment Research, University of Warwick.

Bimrose, J. and Wilson, R. (2013b). *LMI for All: Business Case for Access to More Detailed Data on Pay and Employment.* Coventry: Institute for Employment Research University of Warwick.

Bimrose, J., Wilson, R., Elias, P., Barnes, S-A., Millar, P., Attwell, G., Elferink, R., Rustemeier, P., Beaven, R., Hay, G. and Dickerson, A. (2012). *LMI for All Career Database Project - Processes Adapted and Lesson Learned*. London: UK Commission for Employment and Skills.

Dickerson, A and Wilson, R. (2012). *Developing Occupational Skills Profiles for the UK: A Feasibility Study, UKCES Evidence Report 4*. Wath upon Dearne: UK Commission for Employment and Skills. Retrieved from: http://www.ukces.org.uk/assets/ukces/docs/publications/evidence-report-44-developing-occupational-skills-profiles-for-the-uk-a-feasibility-study.pdf

HM Government (2012). *Open Data White Paper: Unleashing the Potential*. Norwich: TSO. Retrieved from: http://data.gov.uk/library/open-data-white-paper

HM Treasury & Department for Business, Innovation & Skills (2012). *Plan for Growth: Implementation Update (March 2012)*. London: HM Treasury. Retrieved from: http://www.hm-treasury.gov.uk/ukecon_growth_index.htm

Holman, J. & Finefold, P. (2010) STEM Carees Review Report. Report to the Gatsby Charitable Foundation. Retrieved from: http://www.nationalstemcentre.org.uk/res/documents/page/STEM%20CAREERS%20REVIEW%20NOV%202010.pdf

McMenamin, D.G. and Haring, J.E. (2006). An appraisal of nonsurvey techniques for estimating regional input-out-put models. *Journal of Regional Science* 14(2): 191-205.

Miller, R.E. and Blair, P.D.(2009). *Input-Output Analysis: Foundations and Extensions*, Second Edition. Cambridge: Cambridge University Press.

National STEM Centre (2011) STEM Careers Seminar: 28 September 2011. National STEM Centre, University of York. Retrieved from: http://www.cegnet.co.uk/newsletters/nov11/files/PFSTEMCareersSeminarreport26Oct11.pdf

Science for Careers: Report of the Science and Society Expert Group, Diana Garnham, BIS, Feb 2010. Retrieved from: http://www.sciencecouncil.org/sites/default/files/ScienceforCareers.pdf

Tippins, N.T. and Hilton, M.L. (eds.)(2010) *A Database for a Changing Economy: Review of the Occupational Information Network (O*NET)*. Panel to Review the Occupational Information Network (O*NET). Washington DC, USA: National Research Council. Retrieved from: http://www.nap.edu/catalog/12814.html

Toh, M-H (1998). The RAS Approach in Updating Input–Output Matrices: An Instrumental Variable Interpretation and Analysis of Structural Change. *Economic Systems Research* 10(1): 63-78.

UKCES (2010b). *Labour market information, information communications and technologies and information, advice and guidance.* Wath upon Dearne: UK Commission for Employment and Skills.Retrieved from: http://www.ukces.org.uk/assets/ukces/docs/publications/lmi-ict-and-iag.pdf

UKCES (2010c). *The use of LMI in online career direction and learning.* Wath upon Dearne: UK Commission for Employment and Skills. Retrieved from: http://www.ukces.org.uk/assets/ukces/docs/publications/the-use-of-lmi-in-online-career-direction-and-learning.pdf

UKCES (2011b). *Helping individuals succeed: Transforming career guidance*. Wath upon Dearne: UK Commission for Employment and Skills. Retrieved from: http://www.ukces.org.uk/publications/helping-individuals-succeed

Wilson, R. A., and Homenidou, K. (2012). *Working Futures 2010-2020: Main Report.* Wath upon Dearne: UK Commission for Employment and Skills. Retrieved from: http://www.ukces.org.uk/publications/er41-working-futures-2010-2020

Wilson, R. A., and Homenidou, K. (2012b). *Working Futures 2010-2020: Technical Report.* Wath upon Dearne: UK Commission for Employment and Skills.

Wilson, R.A (2010) *Lessons from America: a Research and Policy Briefing*. UKCES Briefing Paper Series. Wath upon Dearne: UK Commission for Employment and Skills. Retrieved from: http://www.ukces.org.uk/briefing-papers/lessons-from-america-a-research-and-policy-briefing-paper