UK Commission for Employment and Skills

LMI for All

Developing a Careers LMI Database:

Final Report (02/07/15)

Career Database Project Team

Warwick Institute for Employment Research

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

Pontydysgu

Graham Attwell, Philipp Rustemeier

Raycom

Raymond Elferink

Rewired State

Julia Higginbottom

Contents

List of tables and figures

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 Exchequer's Notice is required to access potentially disclosive data.
CSS	Cascading Style Sheets (CSS) is a style sheet language used for describing the look and formatting of a document written in a mark-up language. It is designed primarily to enable the separation of document content from document presentation, including elements such as the layout, colours, and fonts and can improve accessibility.
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.
NQF	NQF The National Qualifications Framework (NQF) is a former credit transfer system developed for qualifications in England, Wales and Northern Ireland. It was replaced in 2010 with the Qualifications and Credit Framework.
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).
NQF	National Qualification Framework sets out the level at which a qualification can be recognised in England, Northern Ireland and Wales. Only qualifications that have been accredited by the three regulators for England, Wales and Northern Ireland can be included in the NQF. This ensures that all qualifications within the framework are of high quality, and meet the needs of learners and employers.
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.
Staging	A staging site is a website used to assemble, test and review

	its newer versions before it is moved into production.
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.
SSIS	This is a platform for data integration and workflow applications. It features a fast and flexible data warehousing tool used for data extraction, transformation, and loading (ETL). The tool may also be used to automate maintenance of SQL Server databases and updates to multidimensional cube data.
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

LMI for All is a web portal, which provides access to a comprehensive and rich set of labour market information (LMI) that can be exploited by IT developers to produce a range of applications to help inform better career choices and decisions about learning and work. Through a pilot project that extended over a three year period, (2012 - 2015), the UK Commission for Employment and Skills (UKCES) has successfully demonstrated the feasibility of developing a comprehensive career LMI data tool that exploits open data sources that can be mainstreamed into service provision. Overall aims of the pilot project were:

- 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.

The purpose of this report is to document progress of LMI for All during the second phase of the pilot, detailing the data processing required to populate LMI for All, technical development supporting the LMI for All, current data available and the stakeholder engagement process to raise the profile of the offer.

Data have been made available through a purpose built web portal and data Application programming interface (API). This has been built on three successive iterations (a prototype, phase 1, followed by pilot phases 2A and 2), with the first iteration establishing the feasibility of using existing national data sources to develop a prototype LMI database for careers. By linking and opening up careers focussed LMI, the web portal provides a rich data source to improve the effectiveness and efficiency of organisations involved with, and/or directly providing services that support individuals in making better informed decisions about learning and work. This data source is freely available for any third party developers wishing to harness its potential in their own particular operational context. The purpose of these applications will be determined and developed by these third party developers.

The operational work for the project was undertaken by a consortia led by the Institute for Employment Research at the University of Warwick. Pontydysgu and Raycom led on the technical aspects and development of the LMI for All web portal, database and API. Rewired State delivered the hack and modding days that tested the LMI for All API and explored the feasibility of developing applications and web interfaces using the data.

The main web portal can be found at http://www.lmiforall.org.uk/. This contains information about the 'LMI for All' database and how it can be accessed using an Application programming interface API. The LMI for All data API can be integrated using the web explorer at http://api.lmiforall.org.uk/. Technical information about the data can be found at http://collab.lmiforall.org.uk/, where details will also be found about the current data and indicators included in the database. There is also a frequently asked questions section.

The ‘LMI for All’ database contains the following key labour market indicators, which for the first time are available from a single access point:

- Employment (historical time series 2000-2012);
- Projected employment levels (2012-2022);
- Future job openings (replacement needs);
- Weekly pay (2013);
- Changes in pay 2012-2013;
- Weekly hours (2013);
- Occupational descriptions;
- Skills, Abilities, Interests and Knowledge (based on US O*NET data);
- Unemployment rates;
- Current vacancies;
- Census data (details of geographical location of jobs and travel to work distances);
- First destinations of graduates.

The LMI data generally covers the following dimensions/characteristics:

- 369 detailed occupational categories (SOC2010 4-digit level);
- 75 detailed industries (roughly equivalent to SIC2007 2 dig level)
- Employment status (full-time, and part-time employees and self-employment);
- Highest qualification held (9 levels of the National Qualification Framework [NQF]);
- Countries and English regions within the UK; and
- Gender.

For the potential of these data to be maximised in the process of supporting individuals’ transitions into and through the labour market, they would ideally be transformed into applications designed for specific purposes for a particular beneficiary target group, combined with qualitative data (e.g. job profiles) and mediated by a career or employment practitioner. For example, supporting mid-career changers to upskill or re-skill for a different occupational area in which their skills set is relevant or assisting individuals in their choice of higher education courses relevant for particular jobs in the future.

Technical developments to ensure maximum levels of accessibility to, and integration of, open data have achieved a high level of success in responding to the project requirements. Technical solutions have been found to a number of challenges arising from the complexity of data sets and the overall demands on capacity. Data cube access to some of the LMI for All data was implemented. Data cubes offer a richer, multi-dimensional display of data that is especially well suited to creating cross-category charts in an application. One built on asheHours; and the other on ashePay.

These provide a set of data in a multidimensional structure containing the rules for calculation allowing data to be easily queried. These were constructed based on commonly run queries. Early in 2015, the Applications Programming Interface (API) for the LMI for All web portal was nominated for an Open Data Institute award, testifying to its quality, judged externally. A review of available data sources conducted as part of its Jobs Open Data Challenge, NESTA appointed external assessors who assigned LMI for All the highest score for data quality of all the sources considered.

An extensive stakeholder and communications engagement strategy has been pursued throughout, but with a particular emphasis during the final fifteen months of the pilot project, to raise awareness in the key target groups. These have comprised: the broad community of careers and employment guidance practice; developers, technologists; further education, higher education; and schools. A variety of methods were used, including: keynote presentations at conferences; workshop presentations at conferences; exhibition stands; article features in professional journals; discussions with stakeholder interest groups; presentations to target audiences; and the use of social media. The UK Commission took the lead on dissemination to the policy audience.

High levels of attendance at these events testify to the genuine interest in, and demand for the LMI for All product. However, there is a real danger that the impetus gained through this strand of work will be lost quickly, should the potential user community lose confidence in the longevity of the data portal, not least because investment decisions have to be made regarding the potential use of the dataset for particular operational contexts.

Overall, the first three years of pilot development of LMI for All has been successful in achieving three key goals:

- The development of a comprehensive data offer;
- the implementation of robust, secure, fit-for-purpose technical infrastructure; and
- An increased awareness and understanding throughout the stakeholder community of its existence as a high quality, free resource.

Whilst the database has been developed to a level where it can be, and is being, harnessed by a range of stakeholder groups for various purposes, further areas for development include: updating current databases and adding additional databases relevant to supporting decisions about learning and work; further enhancement and testing of the technical infrastructure; and additional work with stakeholder groups to ensure the potential, together with the likely processes of engagement, are understood and can be implemented within organisational contexts.

1. Introduction

1.1. Project overview

At the heart of the UK Commission for Employment and Skills' (UKCES) strategic objectives and business planning is robust business intelligence that will assist in 'creating the best opportunities for the talents and skills of people to drive competitiveness, enterprise and growth in a global economy' (UKCES, 2014, p.4)[1]. The need to strengthen and improve the quality of labour market information (LMI) for careers and employment practice is essential to inform the choices of individuals who wish to enter or re-enter the labour market, or wish to move between jobs. Easy access to improved data through a single portal has the potential to enhance careers delivery services.

Despite the increasing emphasis on the importance of labour market information and intelligence for supporting individual labour market transitions, access to a number of publicly funded and open large scale longitudinal databases (including the Annual Survey of Hours and Earnings (ASHE), the quarterly Labour Force Survey (LFS) and the Business Register and Employment Survey (BRES) has been limited. Recently, there have been significant efforts to provide online and open access to government datasets in the UK. The release of Public Sector Information (PSI) datasets is advocated on a number of grounds, including: the potential economic benefits of services being developed on top of PSI; the potential for greater democratic accountability through open PSI; the empowerment of citizens to drive local reform of government services based on local data; and the contribution that an 'open data' and 'linked data' industry can make to the competitiveness of the country. The availability of such PSI data sets in enabling the creation of applications based on Open and Linked Data is crucial.

The ultimate aim for 'LMI for All' is to provide a single access point for multiple sources of LMI, which is openly accessible and shared in a way that would allow it to be used by a number of career related interfaces and is viable in the longer term. As part of the staged development in pursuance of this ultimate aim, the immediate aim was to develop a prototype database of careers LMI that can be used to test the feasibility of development of the longer term aim.

The 'LMI for All' pilot project has been funded and managed by the UK Commission for Employment and Skills. Operational work was undertaken by a consortia led by the Institute for Employment Research at the University of Warwick. Pontydysgu and Raycom led on the technical aspects and development of the LMI for All web portal, database and API. Rewired State delivered the hack and modding days that tested the efficacy of the LMI for All API and explored the feasibility of developing applications and web interfaces using the data.

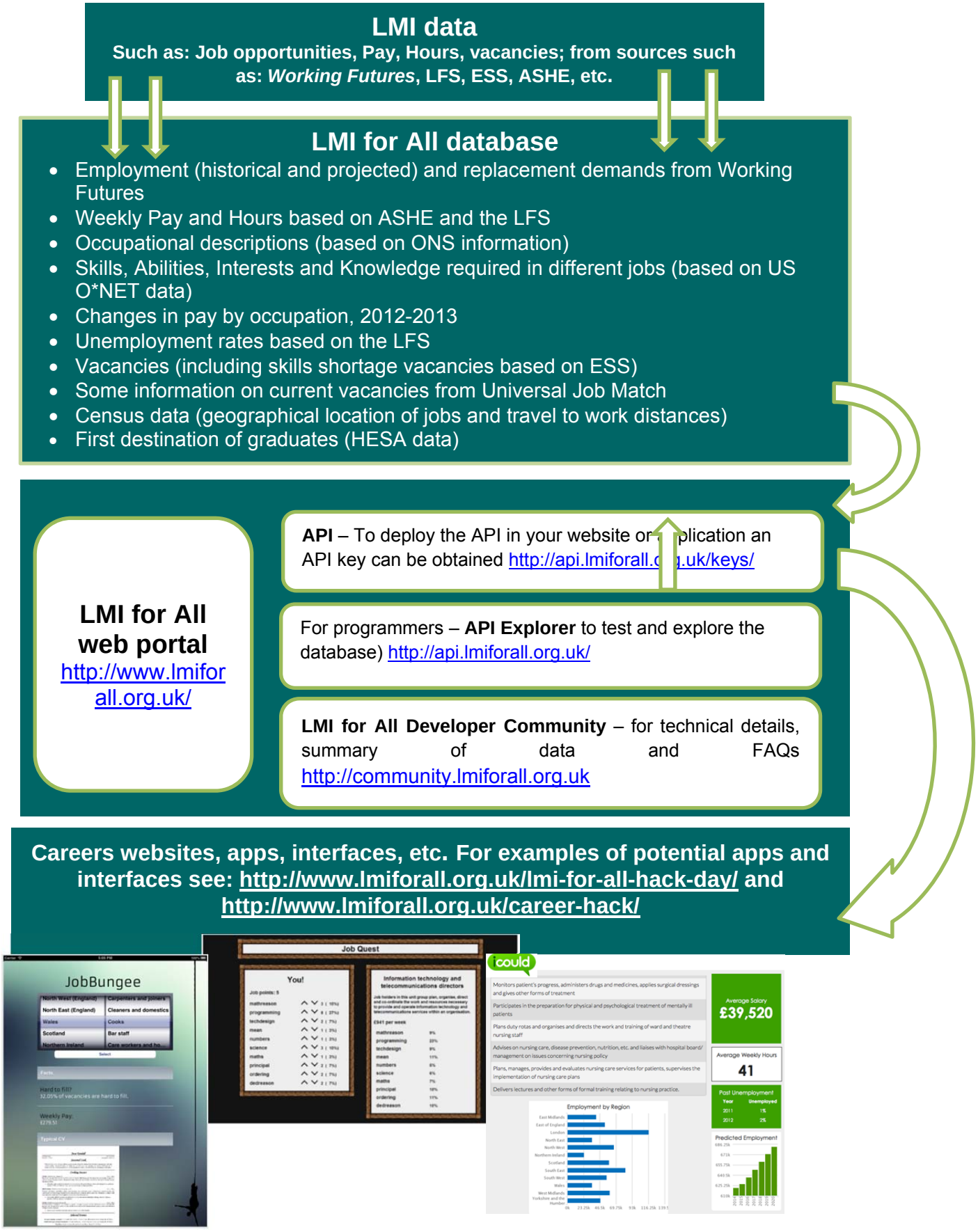
[1] UK Commission for Employment and Skills (2014). UKCES Strategy 2014-2017 and business plan 2014-2015. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/343231/14.08.13._Business_plan___strategy.pdf

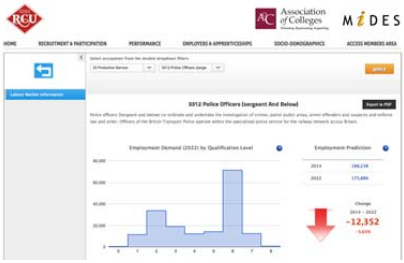
The project has opened up and linked career focused LMI, optimising access, so that individuals can be better supported in making decisions about learning and work. Three sequential, but distinct, research and development phases have ensured a successful outcome. A prototype stage (Phase 1, November 2012 to May 2013) tested the feasibility and viability of the foundation concepts and ideas. Phase 2A followed almost immediately and was completed satisfactorily in 2013 (see Bimrose et al., 2013, for a detailed account).

The report that follows describes the activities undertaken in the third and final stage, Phase 2B, which was largely based on recommendations from Phase 2A.

Figure 1.1, below, provides a visual representation of the database, web portal and API that have been developed from the project.

Figure 1.1 Representation of LMI for All database, web portal and API





1.2. Project aims and objectives

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.

These were represented in three separate, but inter-related work strands, specified by the UK Commission for Employment and Skills, identified below together with their related objectives, all of which have been fully met:

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 pilot feasibility project.
- 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.
- To disseminate findings to a wider audience, through various methods.

1.3. Report structure

This final project report focuses on the Phase 2B activity. It deals with the three different work strands separately: data development (section 2); accessibility and open data: technical developments (section 3); and stakeholder and communications (section 4). A summary and recommendations can be found in section 5, identifying the next steps necessary to secure LMI for All going forward. Its specific purpose is to document progress of LMI for All during Phase 2, detailing the data processing required to populate LMI for All,

technical development supporting the LMI for All, current data available and the stakeholder engagement process to raise the profile of the offer.

1.4. Following up recommendations from Phase 2A

Phase 2A of the project demonstrated the practical feasibility of developing a comprehensive careers LMI data tool designed to support individuals make better decisions about learning and work. LMI for All was, therefore, further developed to meet the LMI needs of these individuals (as well as other potential users in the longer term). Existing data were used, from robust and reliable (mainly official) sources. However, a number of gaps in the existing data were identified only some of which could be filled within the scope of the current project.

The main indicators used in the LMI for All database in Phase 1 (October 2012 – May 2013) continued to be at its core in Phases 2A and 2B, (June 2013 – March 2015). These include:

- 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[2]) 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).

Phase 2B also considered:

- Various refinements to the way these estimates are generated and presented (e.g. focusing on medians/deciles, rather than means).
- Some work outside the LMI for All project (e.g. refining the projections of employment at the 4-digit occupational level, which required an extension to the then current *Working Futures* database).
- The full, revised O*Net dataset, including Skills, Abilities, Interests and Knowledge, as well as a number of other skill related indicators;

Other possible indicators and enhancements considered for inclusion in the LMI for All database during Phase 2B, included:

- Further work to integrate Universal Job Match (UJM) vacancy data into the database more fully, once mapping to occupational categories has been resolved;

[2] 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.

- Making greater use of data from higher education, such as HESA information on the destination of graduates (this required detailed negotiation with data owners);
- Course information – although a great deal of information is available about courses of study and links to different career paths, this is not well coordinated or consistent - work was undertaken to assess the feasibility of bringing this into the database.
- The UK Census of Population, especially local labour market information (there is limited sub-regional information), including some commuting and workplace data);
- NOMIS, consideration of using the API to include workforce jobs data at regional level, the unemployment claimant count and data from the APS;
- Use of more information from the Cedefop pan-European employment database – this is equivalent to the UK Working Futures employment database (but only available at 2-digit occupational level).

It was concluded during Phase 2A that 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
- European Union labour Force Survey (EULFS, problems with availability and detail).

Early discussions took 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 was undertaken to promote uptake. Consideration was given to the resources this requires, 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 was 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.

This included a more strategic use of social media and dissemination at key events throughout Phase 2B, to ensure the web portal and API were promoted to create demand for the product and to maintain the momentum of interest.

The successful format of the hack and modding days carried out in Phase 2A was repeated in Phase 2B. These events were successful in not only helping to prove the viability of the database, but also ensuring that career stakeholders were able to contribute to the development process.

Active participation of key stakeholder representatives throughout the project was carefully designed to ensure engagement and raise awareness of the resource. Throughout Phase 2B, there was an on-going dialogue with organisations that expressed an interest in using the API and their feedback was gathered in order to inform further refinements and amendments to the database and API.

The intention was for communication of the web portal concept to go beyond traditional dissemination methods (e.g. newsletters, professional publications, presentations at various events, etc.). Visual representations of potential applications were made available to various audiences, in response to advice on priority target groups and their career needs collected from key stakeholders (see section 4).

1.5. Data overview

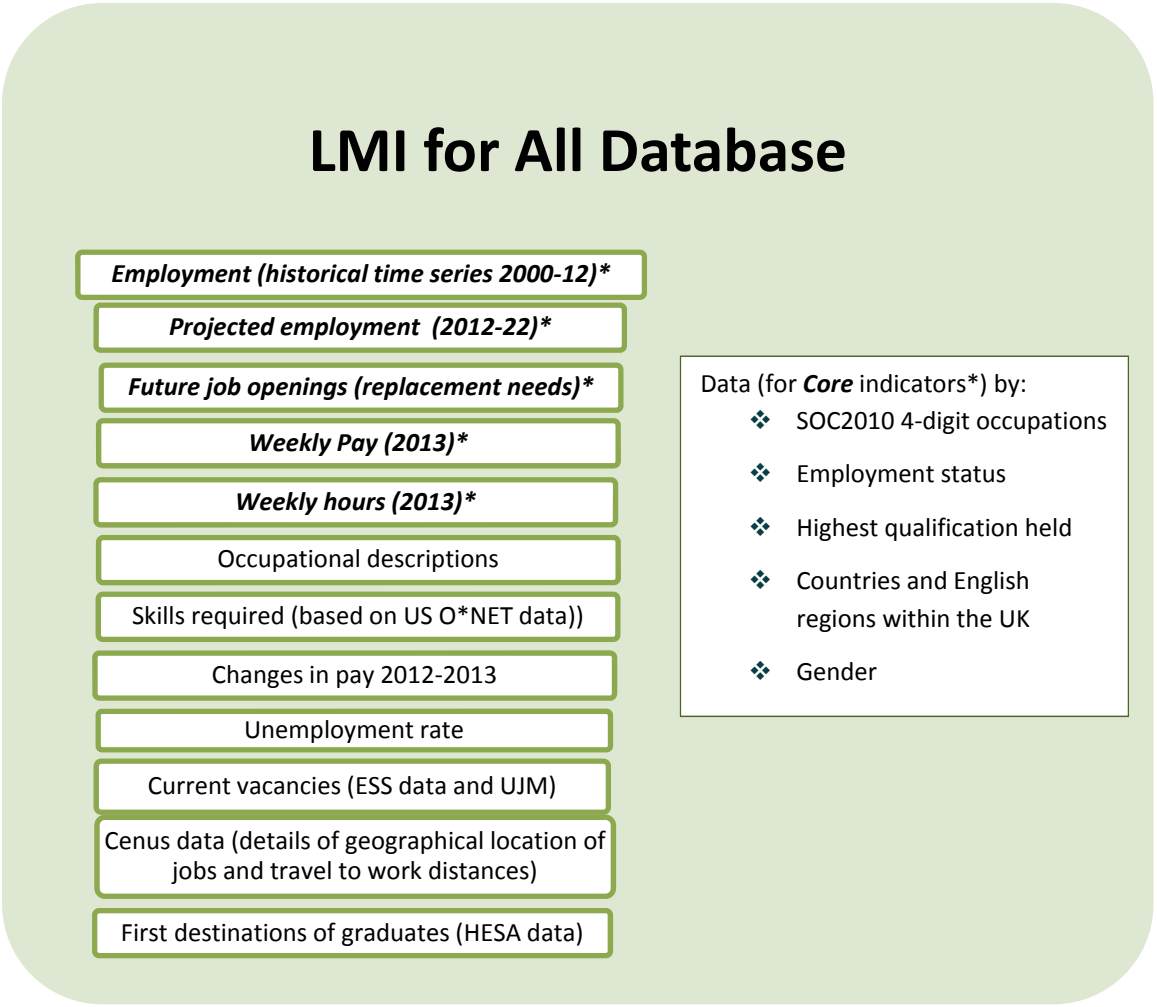
As of end of Phase 2B the LMI for All API contains key data from the following data sets, available from a single access point:

- Employment (historical and projected) and replacement demands from *Working Futures*;
- Weekly Pay based on the Annual Survey of Hours and Earnings and the Labour Force Survey;
- Weekly Hours based on the Annual Survey of Hours and Earnings;
- Occupational descriptions (based on ONS information;
- Skills, Abilities, Interests and Knowledge required in different jobs (based on US O*NET data);
- Changes in pay by occupation, 2012-2013;
- Unemployment rates based on the Labour Force Survey;
- Vacancies (including skills shortage vacancies based on the Employer Skills Survey) and some information on current vacancies from Universal Job Match (UJM) made available by Monster/DWP;
- Census data (details of geographical location of jobs and travel to work distances); and
- First destination of graduates (HESA data).

Sources of these data include: the *Working Futures* employment database; the Labour Force Survey; Annual Survey of Hours and Earnings; UKCES Employer Skills Survey; and the O*NET skills database. Also included in the database are the ONS occupational descriptions. A detailed overview of the data included is presented in Section 2. Figure 1.5 provides a summary.

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.

Figure 1.5 Overview of data and variables in the LMI for All database



2. Data development

2.1. Approach to providing data

The LMI for All database requires detailed data if it is to be useful for services that support individuals in making better informed decisions about learning and work. Individuals and those supporting career transitions have an interest in knowing which jobs are available, distinguishing sector, occupation and typical qualifications required, as well as 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);[3]
- Sector (up to the 2-digit level of SIC2007, about 80 categories);
- Geographical area (12 English regions and constituent countries of the UK);[4]
- Gender and employment status (full-time, part-time employees and self-employed).

The 369 SOC 4-digit occupational categories lie at the heart of the database prepared for LMI for All. Information at this level of detail is provided everywhere possible, although not all data are available at that level of detail.

The ***core*** data provided comprises detailed information, as described above, for:

- Employment (time series of historical and projected levels, plus (for the future only) projected replacement needs (RDs));
- Pay (for a recent year (currently 2013, for employees only);
- Hours (for a recent year (currently 2013, for employees only).

In addition, less detailed information is provided on the following:

- Occupational descriptions for 4-digit occupations (based on ONS information);
- Skills, Abilities, Interests and Knowledge data mapped to 4-digit occupations based on US O*NET information;
- Changes in pay by detailed 4-digit occupation (between 2012 and 2013);
- Unemployment rates (based on LFS data);
- Hard to fill vacancies (currently limited primarily to data from the UKCES Employer Skills Survey);
- Current job vacancies based on UJM API;
- Census (limited information for 2011 on occupational employment at a detailed geographical level and on travel to work distances); and

[3] 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.
[4] Plus in some cases additional information on: age; gender; status; and qualification (highest held).

- Occupational destinations of graduates by detailed 4-digit occupation (based on HESA 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 for the purpose of providing very detailed labour market information to support career transitions.

The 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. These 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 very detailed 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 was 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. Following detailed discussions with ONS, it was concluded that:

- First, the aggregation of information on employment by industry to some 75 industries (by English region and UK nation) could avoid problems of disclosure;[5] 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. Summary of data and indicators included in the portal

2.2.1. Core Indicators

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

For reasons discussed above, these are taken from the *Working Futures* model, which is in turn based on BRES and LFS data. 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 for careers purposes 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 historical 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

[5] Without the necessity for a Chancellor of the Exchequer's Notice (CEN), for details, see Annex A. Information at a more detailed sub-regional level cannot be provided without running into such problems, as well as concerns about statistical reliability of the estimates.

may be less reliable. These criteria are based on rules developed for the main *Working Futures* database. The rules used are based on the practice recommended by ONS for use of LFS data:

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 details are given Annex A.

The published *Working Futures* database also provides projections by occupation at a 2-digit level of SOC2010. In principle, more detailed projections are feasible but this is limited by the quality of the available data upon which the analysis is based (primarily the LFS). In the LMI for All project 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 and then taken to an operational level. 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 as projections (rather than forecasts). 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).

The LFS enables reasonably robust estimates of the current shares of employment in SOC 2-digit categories that are employed in the 4-digit unit groups they contain at the all industry level. In principle, trends within the 4-digit occupation can also be considered and used to develop more realistic projections although only a small number of historical observations are available. In practice, fixed shares for 4-digit categories within the broader 2-digit categories were applied. This was implemented in Phase 2b of the LMI for All project.[6] The initial attempts to produce 4-digit projections in *Working Futures* based on this assumption ran into problems for some detailed categories such as chefs. This occupation has fairly positive growth prospects but is part of a larger 2-digit occupational grouping for which employment was projected to decline across all industries quite sharply. It was not possible to generate plausible projections for chefs, which are heavily concentrated in the hotel and restaurant sectors within that total. An amended set of 2-digit occupational projections across all industries was therefore produced for LMI for All, which differed slightly from the original published *Working Futures* estimates.

[6] There is also scope for considering variations by industry (although sample sizes in the LFS would preclude doing this at a more detailed a level than the six broad sectors used in *Working Futures*).

Figure 2.2 Data overview – LMI for All

All the indicators are currently available in the database although the present version of the API does not provide access to all of these. Providing extended access to these dimensions will require a rewrite of the whole endpoint and we recommend making all this available in an API v2, whilst noting the need to allow access to the original API for those with currently running applications based on this version.

Dataset	Employment (Historical)	Employment (Projected)	Employment (Replacement Demand)	Pay and earnings	Hours	Unemployment Rates	Number of vacancies	Skills, Abilities, Interests	Occupational descriptions	Current vacancies	UK Census Population	Higher education destinations
Source	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)/ Labour Force Survey (LFS)	Annual Survey of Hours and Earnings (ASHE)	Labour Force Survey (LFS)	Employer Skills Survey (ESS)	O*NET database	ONS Standard Occupational Classifications	Universal JobMatch	UK Census Population (England and Wales)	HESA
Indicator	Number of jobs (employee, self-employed)	Number of jobs (employee, self-employed)	Number of jobs openings between selected years (employee, self-employed)	Average full-time earnings; plus indicative estimates of medians and deciles, and limited data on changes in pay 2012-2013	Average weekly hours	ILO Unemployment rate	Number of vacancies, Hard-to-fill vacancies, Skills shortage vacancies, Occupation	Skills, Abilities, Interests	Structure and descriptions of occupations	(Available through fuzzy search)	Information on geographical patterns of employment and travel to work distances	Destination of graduates immediately after graduation
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, Age	Occupation, Industry, Geography, Gender, Status	Occupation, Industry, Qualification, Geography, Gender, Status	Occupation, Industry, Geography	Occupation	Occupation	Occupation	Occupation (1,2,3 digit), Geography**	Occupation, Qualification, Qualification required for job, Subject of study
Period	2000-2012 #	2012-2022 #	2012-2022 #	2013	2013	2011/2012	2011	2013	2010	2014	2011	2011/12-2012/13
Available updates (if known)	every 2-3 years	every 2-3 years	every 2-3 years	annually (or quarterly if required)	annually	annually (or quarterly if required)	every 2-3 years	every 2-3 years	only required when SOC is updated	constant	every 10 years	annually

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).

** Geography available for Output Areas, Lower and Middle Super Output Areas and the hierarchy of local government areas from wards to regions and nations

For 2000-2011 SOC2010 2-digit data is only available

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 because it is based on a larger sample. However, despite this, ASHE 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 and is based on a combination of information from both ASHE and the LFS. Analysis of pay using earnings equations is a well-established way of understanding the key factors that influence pay. In order to ensure that the predicted pay figures match up with the published official data at a "headline" level, 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. This is now done for both part-time and full-timers.

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.

Finally additional analysis has also been included to enable estimates of deciles, including median pay levels, to be derived from the detailed estimates of mean pay. These estimates are based on assumptions that pay is log normally distributed rather than the statistical properties of the original sample data in ASHE or the LFS.

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'. This was explored in Phase 2b of the project. In practice, a non-parametric method has been used based on the published data. The occupational patterns of weekly hours in the ASHE data set are assumed to apply for all industries and constrained to the published ASHE hours for

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

2.2.2. Other Indicators

Occupational descriptions (ONS)

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

ONS have prepared a 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 all of the 369 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

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 add

considerable value from a careers guidance perspective and are therefore included in the full LMI for All database.

The full set of US O*NET indicators now comprises:

Indicator	**Description**[7]
Abilities	O*NET-SOC codes (occupations) Ability scores – enduring attributes of the individual that influence performance (e.g. cognitive, physical, psychomotor and sensory)
Skills	O*NET-SOC codes (occupations) Skill scores – developed capacities that facilitate learning or the more rapid acquisition of knowledge (e.g. basis, complex problem solving, resource management, social, systems and technical skills)
Interests	O*NET-SOC codes (occupations) Interests scores – preferences for work environments and outcomes (e.g. realistic, investigative, artistic, social, enterprising and conventional)
Content Model Reference	Content Model elements and descriptions
Education, Training, and Experience Categories	Categories associated with the Education, Training, and Experience content area
Education, Training, and Experience	O*NET-SOC codes (occupations) per cent frequency data associated with Education, Training and Experience
Job Zone Reference	Job Zone data in seven tab delimited fields
Job Zones	O*NET-SOC code (occupations) and its corresponding job zone number
Knowledge	O*NET-SOC codes (occupations) Knowledge scores – organised set of principles and facts applying in general domains
Level Scale Anchors	Scale anchors associated with the four content areas
Occupation Data	O*NET-SOC codes (occupations), occupational titles and definition/description
Occupation Level Metadata	O*NET-SOC codes (occupations) and the associated Occupation Level Metadata
Scales Reference	Scale information by which the raw values are measured
Task Categories	Categories associated with the Task content area
Task Ratings	O*NET-SOC codes (occupations) Task Ratings scores

[7] For more information on these indicators see: https://www.onetonline.org/find/descriptor/

Task Statements	O*NET-SOC codes (occupations) Task Statements scores
Work Activities	O*NET-SOC codes (occupations) and the associated Content Model Work Activity data – general types of job behaviours occurring on multiple jobs (e.g. information input, interacting with others, mental processes and work output)
Work Context Categories	Categories associated with the Work Context content area – physical and social factors that influence the nature of work (e.g. interpersonal relationships, physical work conditions and structural job characteristics)
Work Context	O*NET-SOC codes (occupations) Work Context scores
Work Styles	O*NET-SOC codes (occupations) and the associated Content Model Work Styles data – personal characteristics that can affect how well someone performs a job (e.g. achievement/effort, adaptability/flexibility, analytical thinking, attention to detail, concern for others, cooperation, dependability, independence, initiative, innovation, integrity, leadership, persistence, self-control, social orientation and social tolerance)
Work Values	O*NET-SOC codes (occupations) and associated Content Model Work Values data associated – global aspects of work that are important to a person's satisfaction (e.g. achievement, independence, recognition, relationships, support and working conditions)
Green Occupations	O*NET-SOC codes (occupations) and associated Green occupations associated
Green Task Statements	O*NET-SOC codes (occupations) and associated Green Task Statements data associated

More detailed information on O*NET indicators and descriptors see Annex B.

Unemployment (LFS)

The unemployment rate is an important indicator for supporting careers transitions. The unemployment rate represents the probability of a worker of a given type, or living in a particular location, being unemployed. The unemployment rate in an occupation is a key indicator, providing information on the likelihood of securing employment. Various sources provide information on unemployment by occupation including the Census of Population and the official series on claimant unemployment made available on NOMIS. However, only one source offers the possibility of developing a consistent time series on the unemployment rate by detailed occupation classified using SOC2010.[8] 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. In principle, the unemployment rate can also be calculated by age, gender and occupation for statistical regions from the LFS.

[8] The official claimant series uses SOC2000 and hence cannot be used.

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) and so is increasingly out of data and irrelevant as an indicator of the current state of the labour market. It is not therefore used in the LMI for All portal.

Vacancies (UKCES ESS and Monster/DWP)

General considerations

The number of vacancies is another key indicator for supporting individuals in making better decisions about learning and work. They provide a measure of the number of jobs potentially available to job-seekers. Historically, the Department for Work and Pensions (DWP) and its predecessors 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 was discontinued and has been replaced by information on raw vacancies generated by DWP/Monster. Unfortunately these data are not coded using SOC, so no occupational data coded to the SOC10 are currently available from this source.

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. In principle, 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 supporting individuals make better decisions about learning and work it was recommended NOT to include this source but to wait for the Monster/DWP data to be made available on a SOC2010 basis.

ESS data on vacancies

At present there is only one statistical source for vacancies that can be used in the LMI for All database to provide information classified to SOC2010. This is the Employer Skills Survey (ESS), carried out once every two years since 2001, and now managed by UKCES.

The detailed UKCES Employer Skills Survey (ESS) collects information on skill deficiencies, including vacancies. It is a sample survey covering some 90,000 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 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 by the two most recent surveys is 2011 and 2013. The ESS has been conducted on a similar basis roughly every two years. Results from the 2013 survey are the first ESS to cover the entire UK and the first to use the SOC2010 classification.

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 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. 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 decreases. Estimates based on an unweighted cell count of less than 50 should not be reported. 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 for supporting individuals make better decisions about learning and work is that it does not provide a picture of all jobs currently available – but a measure of the number of vacancies employers had when the survey was conducted. The latest data relate to 2013. Nor is it comprehensive, focusing on up to six occupations in the sampled firms. However, until an alternative source, such as 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 ONS 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).

General Vacancies (Monster/DWP)

In principle, the data on vacancies collected by Monster on behalf of DWP provides a key dataset for LMI for All. Detailed information on the number of jobs available classified by

occupation is a crucial element for supporting individuals make better decisions about learning and work. Such information used to be available via DWP as Jobcentre Plus vacancies (see discussion in Annex C.4).

The Monster contract with DWP includes 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 (UJM) does not follow standard statistical definitions at present. This lack of standardisation has been the subject of debate in the Labour Market Statistics User Group. The 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.

As noted above the data on vacancies collected by Monster on behalf of DWP replaced 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 uses a system of classification based on job titles that does not match any UK occupational standard. Without a mapping between the categories used by DWP/Monster and SOC2010 4-digit categories used in the LMI for All database, this information is therefore of limited value.

Consequently, the LMI for All Technical Team have implemented a "fuzzy matching" based on reported job titles, which provides a feed of vacancy information from the DWP/Monster website. This includes details of actual vacancies rather than any attempt to quantify the overall number of vacancies or estimate a vacancy rate. The information has limited value, as it is not fully integrated into the main database coded to SOC2010 4-digit occupational categories (although it does allow the user to explore specific opportunities available in their local area). It is worth noting that this is one of the most heavily used indicators within LMI for All, reflecting its perceived importance to both developers and end-users.

A meeting took place with representatives from Monster in October 2013 to discuss the requirements for vacancy data for LMI for All database. Problems with using Monster data were identified and explored, including mapping job titles to UK SOC. Further exploratory meetings and correspondence took place between Monster representatives and IER (namely Professor Peter Elias and Professor Rob Wilson). This focused on the adoption of the IER's CASCOT[9] software package as a possible solution to the mapping problem.

[9]CASCOT, Computer Aided System for Coding Occupational Titles, is a computer program designed to make the coding of text information to standard classifications simpler, quicker and more reliable. The software is capable of occupational coding and industrial coding to the UK standards developed by the UK Office for National Statistics. For more information see: http://www.warwick.ac.uk/go/ier/software/cascot

In 2014 IER undertook a separate feasibility study for Monster to assess if it was possible to recode the Monster/DWP data using a version of CASCOT. This led to a follow-on project and annual licencing arrangements. In principle, this should allow Monster to make data available recoded to SOC2010. At present such data are not in the public domain nor included in the LMI for All portal. A recent media report indicated that Government support for UJM was being withdrawn in the near future.

Recent (March 2015) correspondence between IER and Monster attempted to clarify the position. There seem to be no immediate plans by DWP/Monster plan to publish time series and detailed data on vacancies coded to SOC 4-digit occupation to replace the old DWP Jobcentre Plus (JCP) data series.

The cessation of the DWP Jobcentre Plus (JCP) data series has left a major gap (as highlighted in blogs such that by Educe (2015, http://www.educe.co.uk/?p=1183).

DWP has not asked Monster to provide any data analysis either to them or to a wider audience. So no replacement for the old DWP Jobcentre Plus (JCP) data series or anything else is imminent. This is disappointing and in the context of general ambitions for more open data and the desire to develop systems such as LMI for All that cover all relevant data.

Monster are now using CASCOT to code job postings with a 4-digit SOC code in order to produce a labour market information tool that Monster launched in April of this year (Monster invented online recruiting more than twenty years ago). Monster has been working with the Centre for Economic and Social Inclusion (CESI) for the last year in developing this product. It has been specifically designed to help Further Education colleges understand better the employment opportunities in their local area. The tool works by collecting job vacancy data from job boards and then it matches the advertised occupations with the college courses. In this way curriculum planners can see the relevance of the courses provided to the local employment market. They can also see where there is demand for skills that the college is not provisioning for. The product (see http://www.labourplanning.com) is going through an iterative development (Agile) and will continue to be enhanced in line with customer requirements.

The work carried out by Monster has been at its own initiative and expense (including the purchase of CASCOT). The assumptions made in the processing of job vacancy data (cleaning, deduplication and coding) may not suit everyone and all purposes.

Setting up, managing and populating an ‘open’ data store for free access through an API would require a significant further investment by DWP or some other organisation. In order to fill a significant gap on the LMI for All database such an investment should be a high priority. This would involve placing detailed time series information on vacancies back into the public domain. This should be along the same lines as the old DWP Jobcentre Plus (JCP) data series. This provided estimates of the numbers of notified and unfilled vacancies, and the duration of vacancies, classified to occupations using the SOC2000 classification. From an LMI for All perspective a similar series classified to SOC2010 is needed, preferably made available via an API.

Census of Population variables

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 supporting individuals make better decisions about learning and work. 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 the perspective of supporting individuals make better decisions about learning and work 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 (27th March 2011). For further Details see Annex C.

Three sets of variables derived from the 2011 Census of Population have been included in the LMI for All database. These add some detail to the picture of local employment patterns although the data are of course increasingly out of date. The focus here is on geographical patterns rather than detailed occupational structure.

The three data sets are as follows:

Occupational breakdown of residents in employment: This data set presents the number of people aged 16-74 living in the area and in work during the week before the Census date who were working in each SOC2010 sub-major group. The data is provided for all 232,297 Output Areas in the UK. The Output Areas are referred to by their Office for National Statistics codes and by two types of geographical code: the 1 metre Ordnance Survey grid reference of the geographical centroid of the Output Area and the latitude and longitude of this point. These geographical references can be used to calculate the number of workers in a given occupation within a given distance of a location.

Occupational breakdown of jobs in a location: This data set presents the number of people aged 16-74 working in one of the 53,579 workplace zones in England and Wales for each 3 digit SOC2010 occupation in the week before the Census was taken. Workplace Zones are groupings of Output Areas designed to preserve the confidentiality of employers. They are referred to by their Office for National Statistics codes and by two types of geographical code: the 1 metre Ordnance Survey grid reference of the geographical centroid of the Output Area

and the latitude and longitude of this point. These geographical references can be used to calculate the number of jobs in a given occupation within a given distance of a location.

Mean distance travelled to work in a location: This data set presents the mean distance (in kilometres) between home and work location for people in work within the week preceding the Census date. Mean distances are calculated for persons aged 16 to 74, 16 to 24, 25 to 49 and 50 to 74 for all output areas. In England and Wales, mean distances are also calculated for men and women aged 16-74 and for people aged 25 to 34 and 35 to 49. Data is provided for the 227,760 Output Areas in Great Britain. They are referred to by their Office for National Statistics codes and by two types of geographical code: the 1 metre Ordnance Survey grid reference of the geographical centroid of the Output Area and the latitude and longitude of this point.

First Destination of Graduates (HESA data)

HESA data provide a rich source of information on the pathway of individuals through Higher Education and the first destinations of many graduates. In principle, this data set provides useful information on the kinds of qualifications held by those entering different occupations by both the subject/field of study the level of qualification held.

Data are collected in the HESA graduate destination survey, which contains SOC classification. This allows mapping from courses studied to job destination. Currently much of this information is only made available subject to a fee. Following detailed consultation and negotiation with the data owners led by the UKCES detailed information has been made available for use in LMI for All. The authors and UKCES acknowledge that these data are made available with the kind permission of HESA.

The full set of HESA indicators now comprises:

Variable	Description	Details
ACYEAR	Academic year	2011/2012 and 2012/2013
F_SOCDLHE2010	Standard occupational classification	SOC2010, 4-digit Level
F_LEVEL	Level of qualification obtained	(DOC - Doctorate, MAS - Masters, OPG - Other Postgraduate, FID - First degree, OUG - Other undergraduate)
F_QUALREQ	Qualification required for job	(11 - Yes: the qualification was a formal requirement, 12 - Yes: while the qualification was not a formal requirement it did give me an advantage, 13 - No: the qualification was not required, 14 - Don't know, Unk - Unknown)
F_XJACS201NEW	Subject of study (2012/13)	Principal subject of study.
F_XJACS201OLD	Subject of study (2011/12)	Principal subject of study.
TOTAL	Number of cases	(NB, this includes decimals since there is an apportionment of courses split between different areas).

In principle this data set helps to fill part of the gap between the course of study individuals undertake and the jobs they end up in. Obviously, it only covers part of the picture. In particular it is focussed just on those going through the higher education system. It is also restricted to the jobs that higher education graduates go to soon after graduation (rather than their longer term destinations). Nevertheless, it provides some useful information of interest to those wishing to pursue particular careers or wanting to find out what particular course of study might best qualify them for. The data can be used to consider what occupation graduates with particular qualifications typically end up in. It can also be used to work backwards from an occupation to the types and levels of qualification typically held by those starting out in such jobs.

2.3. Data development summary

The data development strand of Phase 2B has identified, through expert knowledge and stakeholder consultation, the key information used by individuals in making decisions about learning and work, as well as that used by others supporting those decisions and transitions. This has included data and information on: employment rates and forecasts; qualifications; replacement demands; unemployment rates; pay; hours worked; vacancies and vacancy estimates; occupational descriptions; graduate destinations; geographic location of work; travel to work areas; plus occupational skills. These data have been processed and offered as part of the LMI for All service ensuring that issues of quality and disclosiveness have been addressed. Whilst course data were identified as important in learning and work decisions, no viable set of data are currently available due to issues of mapping to SOC, comprehensiveness and/or quality of data. Similar issues with vacancy information have been identified, but could not be resolved with the timeframe of the project. A range of data from other sources were also examined and discounted for a number of reasons. UK wide data have been included disaggregated by region and devolved nation.

Data from a number of sources (namely LFS, ASHE, BRES, Census, *Working Futures* and ESS) have been prepared and made available through the purpose built web portal and data Application programming interface (API) as part of the LMI for All service. The LMI data generally covers the following dimensions/characteristics:

- 369 detailed occupational categories (SOC2010 4-digit level);
- 75 detailed industries (roughly equivalent to SIC2007 2 dig level)
- Employment status (full-time, and part-time employees and self-employment);
- Highest qualification held (9 levels of the National Qualification Framework [NQF]);
- Countries and English regions within the UK; and
- Gender.

For the potential of these data to be maximised in the process of supporting individuals' transitions into and through the labour market, they would ideally be transformed into applications designed for specific purposes for a particular beneficiary target group, combined with qualitative data (e.g. job profiles) and mediated by a career or employment practitioner.

3. Accessibility and open data: technical developments

This section outlines the technical developments that have been undertaken during the initial phases of the project, focussing on accessibility and open data issues. It highlights technical issues encountered and solutions found.

A major issue for Phase 2B of the project was the server and technical infrastructure required. LMI for All is a public service, supported by limited public funding, so caution was required. A series of scenarios were therefore presented to UKCES to estimate the infrastructure required and provide costings for different Phases of 2B. During phase 2B, the technical side of the project faced several challenges. As public usage of the LMI for All service grows, it is important not to disrupt service delivery. Hence a process had to be implemented where data is first staged on a separate staging system, and then moved over to the public system with minimal interruptions. Where needed, the data team prepared the data, making sure the datasets were consistent, complete and valid. These constructed raw data files are then picked up by the technical team and imported into a holding database from where they were further processed, validated and integrated into the LMI for All database.

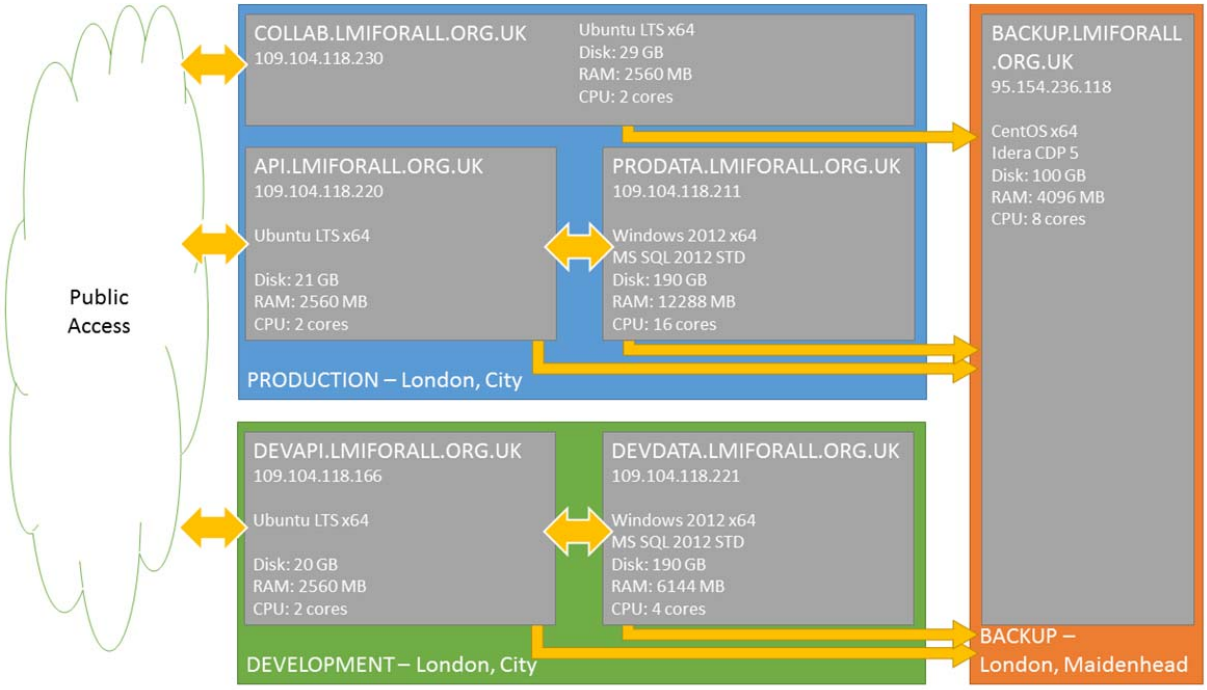
Furthermore, in addition to keeping abreast of updates from the data team, there was a need to develop a more comprehensive view on the already existing data for some of the topics covered in LMI for All. Data cubes are a type of multidimensional database that allows “overview”-oriented queries that examine cross-sections of the data. These were implemented for the ASHE and Working Futures datasets.

3.1. Platform and database

The platform consists of five servers in three separate environments:

All servers are cloud servers, hosted from London based data centres and run by Dediserve Ltd. (http://dediserve.com). To provide a better disaster recovery the backup facilities are run from a different data centre than the development and production environment.

Figure 3.1.1 Overview of LMI for All platform and database



The development environment contains exact copies of the API and PRODATA servers in the production environment and is being used to develop and test new functionality and data-updates before they go live in production. Below is an overview of the two database servers.

Table 3.1 Overview of database servers

IP: 109.104.118.211 HOST: PRODATA OS: Windows Server 2012 version 6.2 (build 9200) USERS: Administrator [pwd: ohQkxkZWcIHT] Guest (default config) SQL SERVER ver: Microsoft SQL Server 2012 - 11.0.2100.60 (X64) Feb 10 2012 19:39:15 Copyright (c) Microsoft Corporation Standard Edition (64-bit) on Windows NT 6.2 <X64> (Build 9200:) (Hypervisor) SQL server authenticated: lmi4all_api [pwd: lmi4all] Windows authenticated: PRODATA\Administrator NT SERVICE\SQLWriter NT SERVICE\Winmgmt NT Service\MSSQLSERVER NT AUTHORITY\SYSTEM NT SERVICE\SQLSERVERAGENT NT SERVICE\ReportServer	IP: 109.104.118.221 HOST: DEVDATA OS: Windows Server 2012 version 6.2 (build 9200) USERS: Administrator [pwd: 2puX10B7tzdW] Guest (default config) SQL SERVER ver: Microsoft SQL Server 2012 - 11.0.2100.60 (X64) Feb 10 2012 19:39:15 Copyright (c) Microsoft Corporation Standard Edition (64-bit) on Windows NT 6.2 <X64> (Build 9200:) (Hypervisor) SQL server authenticated: lmi4all_api [pwd: lmi4all] data [pwd: eong)ah6U] Windows authenticated: PRODATA\Administrator NT SERVICE\SQLWriter NT SERVICE\Winmgmt NT Service\MSSQLSERVER NT AUTHORITY\SYSTEM NT SERVICE\SQLSERVERAGENT NT SERVICE\ReportServer

The databases can be approached using Microsoft SQL Server Management Studio (MS SSMS), but also by a variety of alternative tools, including TOAD, Navicat or JetBrains 0xDBE.

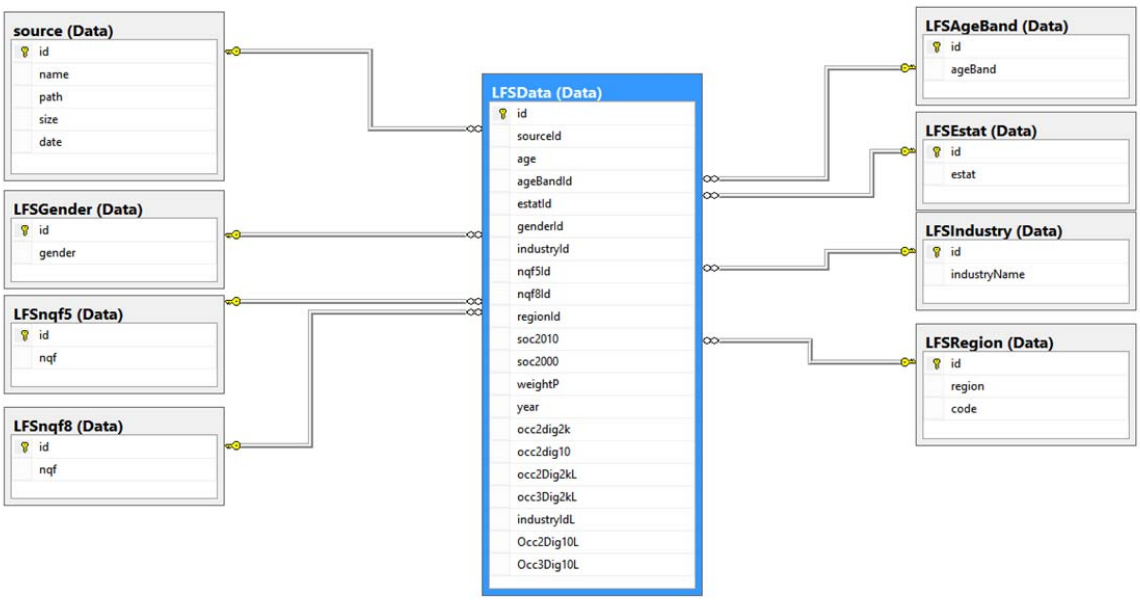
In each of the databases, the same data model that consists of two databases is hosted:

- RAW – for staging (constraints, updates etc.);
- PRODUCTION – for reporting (API, cubes).

All tables are prefixed with the name of the source data, i.e. WFdata, WFregion (source Working Futures = WF).

For our structures a so-called STAR model is used. The star schema consists of one or more fact tables referencing any number of dimension tables. This fits well with the data that needs to be stored for LMI for All. An example for LFS data is included next.

Figure 3.1.2 STAR model illustrated by LFS data



3.2. Extract, Transform and Loads (ETLs)

SSIS packages to automise the data loads have been created. These packages are able to extract the data from the raw data files, transform them into the right structure, validate the values and load them into the database.

Currently, the ETLs are running in debug mode because the formats of the raw data are still under development as the structure of the data is still evolving. Currently, the data are not stable enough to allow for automated processing. Once stability is reached the SSIS packages can be set to run automatically upon upload of new raw data files.

3.3. Data security and data disclosure

All servers are protected by firewalls that allow only traffic over a minimum number of ports from preset IP addresses. This means that apart from limited developer access the Database servers can only be approached from the API servers in the same environment. This means prodata.lmiforall.org.uk only allows connections from api.lmiforall.org.uk and devdata.lmiforall.org.uk only allows connections from devapi.lmiforall.org.uk.

Following detailed discussions with ONS, the research team are confident that the data we have provided are neither disclosive nor confidential. ONS have agreed to place this level of detail into the public domain. They have also agreed that, as the data being presented are econometric rather than 'raw' survey based estimates, these do not fall foul of the Statistics of Trade Act.

3.4. Wiki for tracking project development

The wiki is used for both internal communication and internal and external data documentation and has been substantially overhauled and redesigned. The public area provides extended documentation for developers, whilst the restricted area provides spaces for communication between the data team and the developer team.

3.5. LMI for All web portal

The LMI for All website has been substantially redesigned. Whereas, previously it was based on a static web page, it is now running on the Wordpress Content Management system. This allows the dynamic updating of content, easy management of navigation and navigation motifs, the use of widgets, advanced CSS settings as well as the creation and management of text based and multi media content.

The system allows for user permissions management with different privileges accruing to different kinds of accounts. Google Analytics has also been installed on the site.

The site has been redesigned to attempt to provide different content for different user groups, such as for example careers professionals, managers of careers organisations and application developers. The present structure of the site is as follows:

- Home;
- About;
- Gallery: career hack;
- Widget;
- Developers: LMI key; resources; collaboration space;
- Documentation: API explorer; wiki; data documentation; service level agreement;
- FAQs;

- Terms and conditions.

The website also supports widgets giving access to the LMI for All Twitter feed (@lmiforall). The LMI for All widget is embedded in the site.

3.6. Data cubes

A cube is a set of data that is usually constructed from a subset of a data warehouse and is organized and summarised into a multidimensional structure defined by a set of dimensions and measures. Currently, there are three cubes; one built on asheHours; one on ashePay, and a third one built on Working Futures. Where regular API queries provide a time series constrained to one expression of a variable across the years, data cubes allow users to cross examine two variables directly (for example, employment by gender versus region, wages by qualification across industries, and so on). This instantly provides data of a type that is especially suitable for charting and publication and is also more comprehensive. At the moment, the cubes only support two-dimensional results (i.e., one variable on columns and one on rows), but higher dimensional queries may be possible in the future, with some development work. Currently, these cubes are live on DEVDATA only. Where regular API queries provide a time series constrained to one expression of a variable across the years, data cubes allow users to cross-examine two variables directly (for example, employment by gender versus region, wages by qualification across industries, and so on). This instantly provides data of a type that is especially suitable for charting and publication, and is also more comprehensive. These cubes provide a set of data in a multidimensional structure containing the rules for calculation allowing data to be easily queried. These were constructed based on commonly run queries.

ASHE Pay

This cube contains weekly pay estimate data for one year (the most recent we could obtain). Note that these are estimates. For privacy reasons, the actual numbers are not made public, but the estimates are designed to be close enough to the actual data for meaningful statistics. For more detail, see section 2.2.1, above. The Pay estimates are based on a combination of data from ASHE (Annual Survey of Hours and Earnings) and the Labour Force Survey (LFS). The estimates are derived from data for 2012. Thanks are due to the Secure Data Service at the UK Data Archive for providing access to the Annual Survey of Hours and Earnings (ASHE) data to enable the econometric analysis on which these numbers are based.

- Dimensions: Gender, qualification, region, industry, SOC, year
- Measures: Pay

ASHE Hours

This cube contains information on the weekly hours worked, by occupation and a number of other factors. Similar to the pay data, these are privacy-conscious estimates that are designed to be close enough to the real numbers to be useful. The data is drawn from ASHE (the Annual Survey of Hours and Earnings).

- Dimensions: Gender, region, industry, SOC, year

- Measures: Hours

New data can be included as and when new surveys or other data sources are updated or refreshed. In the case of Pay and Hours, this is dependent on the timetable ONS adopt for publishing new survey results. There is not a single date necessarily as different versions of the data set are made available to different users at different times. For example, the most detailed data are made available at the end of the process when they are deposited with the Secure Data Service. Both Pay and Hours, as well as Employment estimates, depend on the updating and refreshment of the *Working Futures* database. Employment estimates in the latter are used as weights in preparing the detailed estimates of Pay and Hours to ensure consistent with published headline data. The next update of *Working Futures* is currently planned to be completed in the Spring of 2016.

3.7. Maintenance of the API and further development

3.7.1. Technological foundations

The API runs on commodity servers and can be deployed on any operating system that supports the Java Virtual Machine. Linux is preferred (the technical team uses Ubuntu Server LTS), but Windows should work fine. Deployment of the API requires the following components:

- A server and operating system, as specified above.
- The Java Virtual Machine (http://www.oracle.com/technetwork/java/javase/downloads/index.html) The API is compiled using Java 7, but should work on the new Java 8. Archived versions of Java 7 are available here.
- A Java Web Application-compatible web server. The API is delivered as a WAR application container file, which is a standardised format that can be run on a variety of compatible web servers. The tech team uses Jetty (http://eclipse.org/jetty/).
- An installation of Apache Solr (http://lucene.apache.org/solr/) to facilitate SOC code search. A bundle of Solr pre-loaded with SOC codes is provided with the API code.
- The API application itself, packaged into a WAR container file, and
- A swagger frontend (https://github.com/swagger-api/swagger-ui, specification level 1.1) to construct the API Explorer.
- (Optional/Recommended) A reverse proxy and caching server, such as nginx (http://nginx.org/), to tie the various components of the API together in a more pleasing manner. Nginx is also required for SSL support.
- (Deprecated) A PostgreSQL (http://www.postgresql.org/) database to store API keys and developer accounts. This is currently required, but the built-in API key system has never seen wide use. If usage numbers of the API keep increasing, we recommend migrating to a dedicated API key management system that works independently of the main API.

Generally, the API is deployed as follows:

- Install and configure PostgreSQL (or use a version of the API without built-in key management).
- Install, configure and start the web application server.
- Deploy Solr to the web application server.
- Deploy the API to the web application server.
- Install nginx and make the API available under '/api'. Optionally, install certificates and enable SSL support in nginx.
- Deploy the swagger frontend to a directory managed by nginx, and make it available under '/'. The frontend is a small, static application and doesn't need the capabilities of a full web application server. Nginx's built-in capabilities are entirely sufficient. In case nginx isn't available, deploy frontend to web application server.
- Secure other URL parts and ports against attack.

3.7.2. Building and Deploying the API

The API is progammed in Scala (http://www.scala-lang.org/), a programming language that runs on the Java Virtual Machine. However, to build and redeploy the API to the current production system, no actual programming knowledge is needed since we have automated the process. We recommend a Linux or OSX system to build the API, but Windows should work with minor modifications. The following software needs to be installed to do this:

- **Java 7** (Note: for compatibility reasons, this has to be Java 7 specifically. 8 will NOT work.)
- **SBT** (http://www.scala-sbt.org/) This is a build tool that will download all the other necessary components automatically.
- **Python** (https://www.python.org/, version 2.7.x) with the Fabric (http://www.fabfile.org/) package installed.

Once these requirements have been met, simply typing

```
fab build
```

in the API code directory will compile the application container file in the target directory. This is ready for immediate deployment. Typing

```
fab build deploy
```

will build the API, upload and deploy it to the production API server, and reload the server automatically. To perform the automated upload and deployment, the fab script will ask for the 'jetty' account password on the server. The current passwords are:

Passwords

Account	Password	Use for
`root`	`ranunkel`	Administration/entire Server.
`jetty`	`rumburak`	API management/deployment.

3.7.3. API monitoring

There are two levels of monitoring: internal API errors in response to individual user queries, and monitoring of outages of the whole system.

3.7.4. Query Error Monitoring

To monitor query failures, the API server keeps track of query logs. Web queries are always handled with a completion code. A completion code of 200 means that the query succeeded. A code in the 400 range means there was an error, probably on the side of the user (such as 400 Bad Request or the famous 404 Not Found). A code in the 500 range means an error occurred on the side of the server. All such error codes are detected by a watcher and automatically emailed to staff (currently a member of the technical team). The error rate for the 500 class of internal error averages about one to two errors a month. 400-class errors are more numerous, since they are errors on the part of the user, or nonsense requests (which are reported to the user and then discarded without interfering with operations). 400-class errors occur about 200-300 times per day.

To change the recipient of error alerts, log in to the API server using the 'root' account, and edit the file '/root/.swatchrc', replacing the email address on the third line. Typing

```
pkill swatch
```

and then

```
/usr/bin/swatch -c/root/.swatchrc -t/var/log/nginx/api.log –
daemon
```

will kill and then restart the watcher. Rebooting the server has the same effect, but takes longer.

3.7.5. Outage monitoring

To monitor outages of the whole server (in which case the watcher, which also runs on the server, won't be able to alert anyone), a second layer of monitoring is performed by an outside service. This also monitors if the API server website is reachable from the Internet at all. Since this is a third-party service, there is a choice among different providers. The tech team uses UptimeRobot, which is free. Generally, the combined outage of the public system (not including the vacancies service) is below three hours in the last year from the technical side, plus about two days of outages and irregular operations due to a DDOS attack on our server hosting provider, dediserve Ltd.

3.7.6. Future extensions of the API

From the technical side, the API is running fine and supports the provided datasets fine. Depending on how successful the project is in the end, it might be prudent to remove the current built-in API key system (which is underused anyway) and move to a dedicated API key management solution. This keeps API code and key management code separate (thus also separating concerns), and greatly simplifies both development and key management, since each part only has to concern itself with its immediate objective. The company 3scale offers one such (paid) service, but there are also free and self-deployed solutions such as ApiAxle.

3.8. Accessibility and open data summary

Technical developments to ensure maximum levels of accessibility to, and integration of, open data have achieved a high level of success in responding to the project requirements. A purpose built web portal and data Application programming interface (API) have been successfully based on lessons learned from the pilot feasibility project. Feedback from developers and key stakeholders have informed the subsequent iterations of the data tool, gathered as part of the project process.

Technical solutions have been found to a number of challenges arising from the complexity of data sets and the overall demands on capacity. Data cube access to some of the LMI for All data was implemented. Data cubes offer a richer, multi-dimensional display of data that is especially well suited to creating cross-category charts in an application. One built on asheHours; and the other on ashePay. These provide a set of data in a multidimensional structure containing the rules for calculation allowing data to be easily queried. These were constructed based on commonly run queries. Early in 2015, the Applications Programming Interface (API) for the LMI for All web portal was nominated for an Open Data Institute award, testifying to its quality, judged externally. A review of available data sources conducted as part of its Jobs Open Data Challenge, NESTA appointed external assessors who assigned LMI for All the highest score for data quality of all the sources considered. This nomination together with positive feedback from developers and technical stakeholders has not only supported the longer term need for this service, but proven that the LMI for All service can be operationalised.

4. Stakeholder engagement and communication

The development of strong partnership arrangements across a range of different categories of stakeholders and partners was regarded as essential to the success of the LMI for All project. The following principles, specified by the UK Commission for Employment and Skills, set the framework for this activity:

- **User focussed:** Engagement with developers at every stage of development was crucial to make changes to ensure the data tool is user-friendly.
- **A 'work with' approach:** Working with partners and stakeholders to identify where links could be made to add value to existing products/projects, as well as maximising benefits of the relationship to the project was crucial.
- **Focus on the overall objective:** The overall objective for LMI for All was to create a data tool that developers would use to create products to support individuals make better decisions about learning and work. Initial work with partners should, therefore, not focus solely on the development of the data tool, but also form the basis for raising demand and publicising the completed data tool.

4.1. Testing the database API

Originally, the stakeholder engagement and communication for Phase 2B of the project was designed to take place through two major sets of activities:

- The first related to the testing the detail and technical aspects of the data tool with developers to ensure that the database is accessible and useful. This was undertaken by organising a third iteration of Hack and Modding days, which mirrored the processes undertaken under for the same purpose in the Prototyping Phase and Phase 2A.
- The second was to be through a series of events with stakeholders designed to increase awareness of the data tool; gain feedback that can inform the final development of the data tool; and explore the potential for other websites to draw on the data tool using the API. A conference with not more than 100 participants was to be organised, presenting work to date and focusing on how the usefulness of data and how it might be used by stakeholder organisations. In addition, a series of three stakeholder engagement workshops was planned, involving not more than 30 participants each. These would present a more focused opportunities to gain feedback to inform the final development of the data tool and explore the potential for linking to the data tool from other websites with targeted stakeholder groups (for example, career practitioners and their managers).

Whilst the Hack and Modding Days were retained, the second part of the stakeholder engagement plan was amended by the UK Commission. Instead of a large conference, a number of small-scale events for target audiences of potential were delivered. Different categories of partners were identified, as follows:

- **Potential users:** The main intended audience for the data tool is developers of apps and websites. Since developers are unlikely to use the data spontaneously unless there is a clear way to profit from it, this group also includes people who commission careers websites.
- **Wider interest:** This group includes a wide range of different individuals and organisations that have an interest in the project because LMI is central to their role.
- **Technical experts:** Individuals and organisations that have knowledge and expertise about data, website development and/or IT development that can help us to better understand some of the technical issues and the wider agenda around open data.

4.2. Testing the database API

Hack and Modding Days were organised along similar lines to those organised for the pilot phase for the project and Phase 2A. The general 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. 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' in Phase 2B was organised for 23 June 2014. The objectives were to:

- Test, further, 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.

The corresponding 'modding day' was held on 10 September 2014 with the aim of taking the winning application from the hack day through a process of further technical development, towards becoming a marketable product. To ensure the application was useful, feedback from the hack day was used by the developers in a further iteration of the application. The overall aim of the hack and modding days was to produce a marketable development-application. More detailed information about the selection of developers, the stakeholders who participated in these days and the applications produced can be found in Appendix E.

Overall, the feedback from the careers stakeholders on all the applications was positive with many praising the developers for their innovative use and visualisation of the data in the LMI for All database. The applications raised some issues around the need to ensure that they were targeted as different information and data would appeal to those of different ages and stages of their career. Concerns were raised about individuals understanding and being able to recognise their skills in order to start career exploration through an application or web interface. A career narrative element to applications was proposed, whereby a user can explore career pathways. The different approaches were seen to add value at different stages of careers learning and transitions through the labour market for the end user.

Feedback from the developers was also positive. Suggestions were around the development of documenting the data and improvements to the LMI for All website.

4.3. Stakeholder engagement and communications

Engaging with the wider stakeholder community (defined as careers organisations, developers, schools, further education colleges, higher education institutions, recruitment agencies and jobsites) has been a key element of the project to ensure that the LMI for All data tool could be used by developers, support the work of careers professionals and career organisations, and users/customers/clients. The first element, as detailed above, has been the testing the LMI for All data portal and API with developments. The second element of this project has been dissemination and awareness raising activities with a broad range of stakeholders. This has also been key to gaining feedback to inform the final development of the data tool and explore how it can be used by careers organisations. The targeting of specific events to raise awareness of LMI for All has provided focused opportunities to gain feedback for particular groups of users, as well as the opportunity to explore the appetite to use the data and whether stakeholders see the value in the data tool as well as the value in linking this to their own work. This was framed against an assessment of client/customer LMI needs, such as the information needs of different groups, gaps in information, influences on and the process of career decision making, and understanding of LMI.

Various methods have been used to disseminate LMI for All to different stakeholder groups. Over the past 15 months, 851 participants have attended these events to learn about this innovation. Methods have included:

- Presentations at conferences (e.g. CDI and IAEVG), n=4;
- Invited presentations to targeted audiences (e.g. Universities UK), n=6;
- Invited keynote presentations (e.g. National Symposium, Republic of Ireland), n=10;
- Discussions (e.g. Education Services Australia, Association of Colleges; plotr), n=11;
- Article in professional journal (Career Matters, CDI Professional Journal);
- Hack and Modding Days (i.e. career stakeholders), n = 4.

Details of the stakeholder engagement and communications strategy, together with events and numbers of participants are presented next.

4.3.1. Stakeholder dissemination and communication strategy

Objective	What practical steps do we want them to take?	Contribution to KPIs	Dissemination activity
Schools			
Raise awareness among teachers of LMI for All as a source of intelligence to inform careers practice within schools	• Want teachers involved in provision of careers support to access LMI for All via existing websites (iCould etc) •	• Increase in unique visitors to API • Widen base of end-users	• Development of schools strategy development paper led by Sir John Holman. Discussion with Sir John Holman (27/08/14), stressing the importance of disseminating to schools. He followed up on 29/08/14, with an undertaking to 'give some thought' to the challenge of developing a school strategy paper. • Presentations to ASCL (15/06/15). • Special Schools and Academies Trust (SSAT): meeting on 28/11/14 introduced LMI for All – • SSAT workshop on 12/02/15 (n=50) • CEIAG Conference (David Andrews) – Keynote on 21/11/14 (n=45) • Inspiring Futures: Regional Directors' Forum on 15/12/14 – keynote on LMI for All (n=38) • Dissemination and promotion of Careerometer and publicly available websites using LMI for All data • Education Services Australia – skype focused on LMI for All initiative. Discussions on-going. • Education and Employer Taskforce – seminar presentation on 28/11/14 (n=58)
Schools to promote LMI for All as resource for pupils and their parents	• Schools to implement widget on their websites • Schools to refer pupils and parents to third-party websites and apps that make use of LMI for All • Schools to develop their own apps, on individual basis or as part of consortium	• Increase in number of apps using API • Increase in unique visitors to API	

Objective	What practical steps do we want them to take?	Contribution to KPIs	Dissemination activity
FE colleges			
Colleges to use LMI for All data to inform curriculum strategy and development	• Colleges to access existing websites including RCU data store and Skills Match (forthcoming product from Mime Consulting)	• Increase in unique visitors to API	• Presentations to Association of Employer and Learning Providers, Titan partnership Ltd. On 29/01/15 (n=16) • Association of Colleges, contact with Regional Representative for the West Midlands • Further Education Learning Technology Action Group (workshop and stand) 22/10/14 (n=48) • Dissemination to JISC and City and Guilds • Involvement of Gloucester FE College in the Hack and Modding Days
Colleges to offer LMI for All data to students to inform learning/careers decisions	• Colleges to install widget on their websites • Colleges to refer students to third-party websites and apps that make use of LMI for All • Colleges to develop their own apps, either individually or as part of consortium	• Increase in unique visitors to API • Increase in number of apps using API • Widen base of end-users	
Universities, HE institutes			
Colleges to offer LMI for All data to students to inform learning/careers decisions	• Colleges to install widget on their websites • Colleges to develop their own apps, either individually or as part of consortium	• Increase in number of apps using API • Increase in unique visitors to API • Widen base of end-users	• Presentations at AGCAS annual conference (opening address), AGCAS Heads of Service conference on 06/01/15 (n=39) • Dissemination to Universities UK – presentation on 19/02/15 (n=17) • Early development at University of Warwick with Student Services (05/03/15) • Open University – meeting with Head of Careers Service on 19/02/15 • Contact with Republic of Ireland, AHECS executive

Objective	What practical steps do we want them to take?	Contribution to KPIs	Dissemination activity
Recruitment agencies, jobsites			
Jobsites to offer access to LMI for All data as an additional information resource to support customers in exploring careers options	• Jobsites to install widget • Jobsites to develop their own dedicated apps using LMI for All	• Increase in number of apps using API • Increase in unique visitors to API • Widen base of end-users	• Presentation to The Recruitment & Employment Confederation research steering group on 12/12/13
Careers organisations			
Raise awareness of LMI for All among careers professionals in order to encourage them to use data to inform their careers practise	• Careers professionals to access websites that already offer LMI for All	• Increase in unique visitors to API	• CDI conference, 2013 keynote (n=120) • CDI conference 2014 keynote (n=120) • CDI 2014, workshop presentation (n=16) and stand • CDI student conference (2015) – keynote presentation (n=60) • Article published in 'Career Matters' October 2014 • Presentations at international IAEVG (2013, n=25; 2014, n=27) conferences • DWP – on-going discussions about app development for both employer engagement teams and training and development work coaches • NCS West Midlands Education & Training Sectors 25/03/15 (n=30) • National Careers Guidance Show, 04/03/15 Opening presentation at Breakfast Reception (n=40) • Republic of Ireland, National Symposium 10/10/15 Keynote presentation (n=70) • CDI South East regional meeting (n=28) • Hack and Modding days (n=22)
Careers organisations to draw on LMI for All data as part of their wider IAG offer to clients	• Careers organisations to develop their own dedicated apps using LMI for All • Careers organisations to install widget on their websites or link to existing resources (e.g. iCould)	• Increase in number of apps using API • Increase in unique visitors to API	

Objective	What practical steps do we want them to take?	Contribution to KPIs	Dissemination activity
Developers			
Raise awareness of LMI for All among developers as a data resource that can be incorporated into their offerings to commercial customers	• Use LMI for All as a source for their own app development • Review examples of existing apps on LMI for All website; re-use existing code as part of their own development work; promote potential of LMI for All to clients	• Increase in number of apps using API • Increase in unique visitors to API • Widen base of end-users	• Presentations x 2 and stand at Alt-C conference (1-3 September, 2014) • Hack days x 2 • Modding days x 2 • Plotr website – 2 meetings to explore potential

The team has also engaged with the media and social media to disseminate the LMI for All web portal. An infographic was produced to illustrate the type of data that are available in the database, which was used with the media by both IER and the University of Warwick. Twitter (@WarwickIER and @CareersResearch) has been extensively used by the team to promote activities and events, as well as report progress with the project. These have been retweeted by UKCES who manage the LMIforAll twitter account. Engagement with social media has been successful at promoting the project to a wider audience.

4.4. Future implications

The level of participation, and interest of participants attending dissemination events, has been consistently high. This has been gratifying and demonstrates a real appetite for the product. For instance, SSAT (The Schools Network) is a UK-based, independent educational membership organisation working with primary, secondary, special, free schools, academies and University Technical Colleges (UTCs). A session on LMI for All was delivered in workshop format to SSAT membership on 12th February, 2015. SSAT organised and hosted the workshop, circulating information about the content in advance and inviting expressions of interest. Over 50 indications of interest were received. After a presentation about the web portal and demonstrations of applications that could be developed, participants identified priority target groups for the development of applications which could present customised labour market information. These included: students from families experiencing intergenerational unemployment; parents and carers; subject teachers; disengaged young people (NEET: Not in Education, Employment or Training). The purposes of the applications designed for these target groups would be to: inform, inspire, motivate and educate. Barriers to integrating LMI for All in schools included: technology compatibility issues and language (students for whom English is not the first language). Advantages of harnessing the potential of the dataset were also identified. For example, access to high quality, reliable data about the labour market and the potential for an application enhancing students' e-portfolios.

A number of organisations have requested follow-up meetings, subsequent to initial presentations, to explore the potential next steps within their organisation. Other organisations and consortia have indicated an interest in, for example, implementing Careerometer, exploring organisational requirements and capacity to use LMI for All, reviewing existing app code and how it could be developed to meet organisational needs and, in one instance, exploring whether organisational data could be added to the LMI for All database. There have been progressed, where feasible. The success of the stakeholder and communications strategy, does, however, emphasise the importance of retaining the momentum of this activity, to ensure that the uptake of LMI for All is fully integrated in organisational practices.

4.5. Stakeholder engagement and communication summary

The LMI for All service was thoroughly and successfully tested through two separate iterations during the phases 2A and 2B. Hack and modding days were organised during the two phases, which enabled developers to explore and test the service. During these events a number of apps, widgets and websites aimed at individuals making learning and work decisions were design and developed. Careers stakeholders were able to judge the

developments and inform future iterations. Overall, the events proved that useful services could be developed using the LMI for All data.

An extensive stakeholder and communications engagement strategy has been pursued throughout, but with a particular emphasis during the final fifteen months of the pilot project, to consult and raise awareness in the key target groups. These have comprised: the broad community of careers and employment guidance practice; developers, technologists; further education, higher education; and schools. A variety of methods were used, including: keynote presentations at conferences; workshop presentations at conferences; exhibition stands; article features in professional journals; discussions with stakeholder interest groups; presentations to target audiences; and the use of social media. The UK Commission took the lead on dissemination to the policy audience. A range of promotional materials were also developed to support dissemination activities.

High levels of attendance at these events testify to the genuine interest in, and demand for the LMI for All product. However, there is a real danger that the impetus gained through this strand of work will be lost quickly, should the potential user community lose confidence in the longevity of the data portal, not least because investment decisions have to be made regarding the potential use of the dataset for particular operational contexts. The UK Commission for Employment and Skills has made a commitment to continue to support the portal into the longer term, though this commitment currently has no formality or visibility in the public domain.

5. Future issues and potential resolutions

5.1. Enhancing the database: potential and additional data sources

5.1.1. General considerations

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 supporting individuals make better decisions about learning and work.

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 and 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 many problems of consistency. There is no obvious methodology for merging “real” and predicted values in a seamless fashion. This is likely to be a very demanding technical exercise, which would require detailed consultation with ONS, with no guarantee of reaching a successful and agreed outcome. This is probably too difficult and would raise too many new problems to make it worthwhile pursuing. In general, the authors of this report are of the view that we should use either:

- Statistical/survey estimates (where reasonably reliable information is available and the demand for detail is not that great); or
- Econometric (or similar) estimates (where the survey estimates cannot provide the level of detail required).

Not all data are 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 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). This requires 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.

A number of other sources might add information that could be of value to a broader audience than those concerned with the support of individuals making better decisions about learning and work. Once the database is fully 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.

From the perspective of supporting individuals make better decisions about learning and work, the two main areas that need to be enhanced in the short-medium term are:

- Provision of more detailed data on vacancies, properly coded to SOC2010; and
- Addition of more and better information on links between courses of study and job outcomes i.e. understanding what types of course are relevant to particular occupations and vice versa; and then providing access to information about specific course opportunities.

From a more general perspective the database has potential for many other uses, including local economic analysis and development. This calls for a much greater use of data sources such as the Census of Population as well as other ONS data, some of which may be available via NOMIS. Annex C provides a more detailed consideration.

The remainder of this section consider the main possibilities in a bit more detail.

5.1.2. Vacancy data

As noted above, in principle, the data on vacancies collected by Monster on behalf of DWP provides a key dataset for LMI for All. However, there is a need for vacancy metrics classified by SOC2010 in order to provide fuller integration with the rest of the LMI for All database.

It should be a priority to make these data available (or an equivalent dataset) for the LMI for All database.

5.1.3. Course information

Course information is particularly important for learners, as it enables the identification of learning opportunities that relevant to a chosen career path. However, two issues have been encountered in trying to locate and include course information and data into LMI for All. First, it is difficult to map occupation (defined by SOC) and subject classification and second, collating course information and data and classifying it to the subject classification. Information is variously available, sparse and provided in a range of formats. Consistency and quality are a concern.

Data on courses and training available across the UK are not held in any one central database. Discussions were held with various government departments and other relevant organisations to negotiate access to the information repositories, which are accessed through various search tools. From this it is evident that compiling a comprehensive list of further and higher education training and courses is very complex, mainly due to the number and range of courses available, as well as the variable quality of the data. Accessing the

data is 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[10], XCRI[11]). In order to include such detailed 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. Current problems with the quality of course data continues to be an issue, but it is improving. 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. 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.
- In **Scotland**, information about learning opportunities and careers in Scotland is collected and collated 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. It covers both further and higher education data, which are updated on an annual basis. This information is currently available through a range of online services (such as MappIT, MerIT, PlanIT Plus, WorkIT) and reference books. Course data are not freely available.
- In **Wales**, the Welsh Government and Careers Wales collect and update course information and vacancy data for Wales. Agreement was secured, in principle, that access to these data could be provided through an API, but has yet to be followed up.
- 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, but is no longer

[10] 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.

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

available and there are no plans to update or maintain this 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 with other organisations that collect course data included HESA and their data are now included in the database.

More generally, these discussions confirmed that compiling a comprehensive list of course information and data is too complex and time-consuming to be incorporated into LMI for All as a matter of course. In particular there are problems in ensuring that data can be automatically updated on an annual basis.

Another possibility that was considered was the use of LFS data to provide some insight into subjects of learning that are relevant for entry into particular occupations. In principle, this data source offers some potential as survey respondents are asked questions about their formal qualifications acquired and hence course of study followed. In practice, problems of limited sample size mean that this is not much use in the context of LMI for All. Queries at the level of detail that are meaningful from the perspective of supporting individuals make better decisions about learning and work (which would require 4-digit occupations and a detailed breakdown by both level and field of study) would return zero entries in the vast majority of cases. Aggregation up to higher levels by occupation and across qualification categories eases such problems but at the expense of the detail required. The main employment indicators in the LMI for All database, which provides information on occupation by 4-digit occupation and broad level of qualification exploits that data to its limits. Further details on these issues are provided in Annex C.

It is clear that there is no central database of course data available, but various stakeholders have partial information, which makes the link between course subject and information. For instance:

- Some further education courses have been mapped using XCRI (XCRI stands for eXchanging Course Related Information. It is the UK standard for describing course information). However, this is limited and regional coverage is varied. For instance, Nottinghamshire has the majority of further education information coded and available.
- Higher education courses are available on the UCAS (University and Colleges Admissions Service) website, which covers the UK. These courses are mapped to JACS.

Discussions have been 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. These data would supplement and extend the HESA data already included in the database, as well as provide an additional source of information.

Discussions so far have confirmed that compiling a more comprehensive list of course information and data will be complex, time-consuming and likely to be resource intensive.

In the medium to longer term, it is likely that course data will need to be carefully mapped and expectations managed as data will not be automatically updated on an annual basis. Manual input will be required unless it is possible to access external APIs to dynamically update data.

Overall, accessing course 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[12], XCRI[13]). A basis for mapping higher education course subjects (JACS) to occupation using the HESA data is available. However, there is gap in mapping courses to occupations in further education sector. Ideally, a common classification would be preferable particularly as JACS does not take account of subjects that are relevant to lower skilled occupations.In order to include course data in the LMI for All database, a comprehensive mapping of courses to occupational codes would be needed and commissioning such work should be considered.

To explore possibilities to map XCRI and JACS, and JACS and SOC. Discussions are underway with those who led on the Salami project in Nottingham as they have been developing a method of coding SOC to JACS through a thesaurus, which they may be willing to share. This mapping would enable data from the XCRI API feed of higher and further education courses to be included in the API. However, this would not be a complete directory of further education courses.

It will be necessary to follow up discussions with Student Information Services Limited to explore the API.

It seems unlikely that the disparate national sources of course data can be pulled together to create a complete dataset.

5.1.4. Census of Population data

The Census of Population provides very geographically detailed information on the location of employment and the characteristics of workers in 2011. The LMI4All database includes a number of variables from the Census, which can be used as the basis of indicators, which detail the spatial pattern of labour demand and the geographical distribution of workers.

Future developments that might enhance the database could be focused more on a local economic development perspective rather than the careers support angle.

From a local economic development perspective, the main value of the Census data is to provide a detailed geographical breakdown of the availability of workers of different skill levels. The sort of variables which could be derived include:

---

[12] 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.

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

- Number of workers at a given skill level (defined in terms of SOC major groups) within certain distance bands of a location of interest;
- The percentage of workers at various skill levels within a locality being considered for industrial development;
- Identification of areas in which employment of particular occupations is concentrated.

Further details of what is available can be found in Annex C.

5.1.5. European data – the Cedefop database and EU Skills Panorama

Cedefop projections and related data

Over the past 10 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 2014/15 will use ISCO08. 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 for LMI for All, given the lack of occupational detail and the difficulties in making a simple mapping of occupational categories.

The European data is also being expanded by Cedefop to populate the EU Skills Panorama. The latter is a new website/portal aiming to provide a comprehensive one stop shop for LMI at a pan-European level. This is still under development by Cedefop. The current version can be found at: http://euskillspanorama.cedefop.europa.eu/

Other European sources

A range of other European sources has also been considered for inclusion in LMI for All. These include 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. They 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.

5.1.6. Stakeholder impact and future viability

A high level of interest has been generated in the product through different dissemination activities, delivered as part of the stakeholder and communications strand (see section 4.3, above). Stakeholder activities were designed to target representative bodies to ensure the effective use of resources and target large numbers of stakeholders. Technical skills and resourcing have been particular issues arising from the dissemination activities. Developing innovative ideas on what is useful and could be developed from the LMI for All service has been unproblematic for stakeholders. The LMI for All service is seen as having the potential to make significant impact in helping individuals make learning and career decisions.

A frequently asked question at events has related to the future prospects for the data portal. Investment decisions in app development by different stakeholder organisations clearly hinge, in many cases, on evidence that the future of the portal is secure. It was not possible to convey this level of assurance during the final months of phase 2B. At the time, there was concern that the momentum gained through the intensive stakeholder activities was at risk of being lost due to the project's uncertainty. In the final part of phase 2B, the UKCES commissioners approved the continuation of the LMI for All service for an indefinite period. Throughout the project, a reoccurring question has been the uncertainty about future viability, which could not be determined or communicated until the later stages of the project. Organisations were, understandably, hesitant about using resources to develop an app containing data that may not have been updated.

5.2. Future implications for costing

5.2.1. General considerations

As discussed in the previous sections there are many technical problems linked to the fact that these sources were not designed with the particular purpose of providing data suitable for supporting individuals make better decisions about learning and work. For this reason, contingency planning for time required to deal with issues related to each dataset have been included in estimates below, based on experience to date. Additionally, the costs of maintaining and updating the database from a data perspective are therefore much more significant than if it were possible simply to tap into a relevant API for each of the main data sources involved.

The LMI for All project has demonstrated that adequate data are available to populate a rich 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 core 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).

As noted in the previous section, 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 would 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.

5.2.2. Employment

There is a need to update *Working Futures*:

- Employment (historical time series 2000-12);
- Projected employment (2012-22);
- Future job openings (replacement needs).

Because the data available directly from the official sources are not sufficiently detailed to provide data for 4-digit occupations cross-classified by other dimensions of interest, it is necessary to generate estimates using econometric and other methods. This has been characterised as the Working Futures employment database.

Updating of the employment estimates therefore requires that the full Working Futures database is updated. This is a major project that has typically been let by competitive tender once every 3 years or so. The budget required depends on the precise specification set out in the tender, but is likely to be well into 6 figures (i.e. £100-200K).

This excludes any time required by the Data Team to process the Working Futures data and by the Technical Team to upload the processed data to the LMI for All portal. Assuming the specification for any update to Working Futures builds in a requirement to produce data compatible with LMI for All, this should be quite modest.

Based on the experience in LMI for All Phase 2. This is expected to involve around:

- 1 day of senior research time (SRT) to manage and supervise the process;
- 2 days of research support time (RST) to process the results and upload them for the Technical Team;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

5.2.3. Pay and Hours

Need to update econometric and related analysis of LFS and ASHE data, which realistically could be undertaken on an annual basis:

- Mean Weekly Pay;
- Medians and deciles;

- Estimates by age;
- Annual changes in pay;
- Weekly Hours.

Because the data available directly from the LFS and ASHE are not sufficiently detailed to provide data for 4-digit occupations cross-classified by other dimensions of interest, it is necessary to generate estimates using econometric and other related analysis.

Based on the experience in LMI for All Phase 2. This is expected to involve around:

- 3 days of senior research time (SRT) to manage and supervise the process;
- 20 days of research time (RT) to manage and supervise the process and to conduct the relevant econometric analysis, (some of which needs to be carried out in the Secure Data System run by ONS);
- 20 days of research support time (RST) to process the results including updating the RAS processes and generate the new estimates;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

5.2.4. Occupational descriptions and skills

ONS descriptions:

- Nothing to be done until SOC is revised (No date for this process has been published by the ONS, but it is expected to be 2020).

O*NET Skills required (based on US O*NET skills information):

- Redo any mapping to US occupations
- Identify collate and make available relevant data files on skills

Based on the experience in LMI for All Phase 2. This is likely to involves around:

- 2 days of senior research time (SRT) to manage a supervise the process;
- 20 days of research support time (RST) to update any mapping using CASCOT and to process the results and upload them for the Technical Team;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

5.2.5. Unemployment and Vacancies

LFS analysis of unemployment rates

- This involves interrogating the LFS and extracting the relevant data on Unemployment rates.

Based on the experience in LMI for All Phase 2. This is estimated to involve around:

- 1 day of senior research time (SRT) to manage and supervise the process;
- 2 days of research support time (RST) to process the results and update the Wiki;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

ESS Vacancies

- This involves obtaining the relevant vacancies data from the survey company responsible for conducting ESS and processing the data for use in the database.

Based on the experience in LMI for All Phase 2, this is estimated to involve around:

- 3 days of research time to manage and supervise the process;
- 1 days of research support time to process the results and update the Wiki.
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

DWP vacancy data (classified by standard occupations and regions)

Assuming DWP/Monster make the data available via an API this should be a relatively straightforward task. However as noted this does not seem likely without a major new investment by DWP or some other organisation. Unless this happens this will not progress. Currently the Technical team have implemented a temporary solution based on the data Monster have made available and fuzzy matching.

Based on the experience in LMI for All Phase 2 for other similar data sets, this is estimated to involve around:

- 2 days of research time to manage and supervise the process;
- 2 days of research support time to process the results and update the Wiki;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;

- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

5.2.6. Other indicators

Census data

There is nothing to be updated unless new indicators are included. There are some possible new indicators to be added (see Annex C). If these are to be added then the marginal costs will depend on precisely what is included.

Assuming a single indicator, based on the experience in LMI for All Phase 2, this is estimated to involve around:

- 5 days of research time to manage and supervise the process;
- 1 day of research support time to process the results and update the Wiki;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

HESA course data

First destinations of graduates – the main tasks are:

- Obtain updated information from HESA;
- Update documentation;
- Add to database.

Based on the experience in LMI for All Phase 2, this is likely to involve around:

- 1 day of research time to manage and supervise the process;
- 2 days of research support time to process the results and update the Wiki;
- 4 days of Technical Team time (TTT) to upload the new data for testing;
- 3 days of Technical Team time (TTT) to move data from testing to production;
- 2 days of Technical Team time (TTT) to adapt the API, as necessary, for the dataset;
- 3 days of Technical Team time (TTT) for contingencies – responding to unanticipated challenges with the data.

5.2.7. Technical improvements indicated

Because of the complexity of data, prebuilt packages for loading could not be used for most datasets. Almost every data update in the final iteration of the project had their format

changed or were completely new. Consequently, it will be necessary to build and deploy permanent SSIS packages, with the following technical improvements required:

- Database: clean-up and optimization is necessary, with maybe slight architecture changes (depending on patterns of usage emerging).
- Cubes: whilst SSAS is up and running, dimensions/attributes were selected in the absence of usage data. Changes will have to be applied as patterns become clearer.

For an auto import solution to be implemented, a format checking app would be needed, to get the file directly to db. Mandatory option could be: update or a complete reload. For this, some kind of an option for syncing DEVDATA would be necessary, after checks with PRODATA.

One issue that has emerged in discussions with employment advisers is the desire for more local and geocoded data. It is possible to 'mash' data from LMI for All with NOMIS local data. However, there remains the problem that as data becomes more local, the sample size becomes smaller and, thus, the analysis we can offer becomes less fine-grained in terms of occupational and other categorisations. One potential answer may be 'crowdsourcing' to all employment and careers advisers and other end users to add local 'intelligence' to the LMI provided at national and regional levels. There is nothing to stop application developers adding these features themselves. But there may be benefit in including such intelligence within the national database for scaling purposes. A further possibility would be to develop scrapers to collect data from for example local newspaper websites to add to the 'official LMI. These options would require a significant level of development and resourcing whilst highlighting issues of data quality and up-to-datedness.

5.2.8. Stakeholder dissemination and communications

A number of activities have been initiated during Phase 2B of the project that require follow up. The nature of the follow up would need to be discussed and negotiated with the UK Commission. However, stakeholder groups in the schools, higher education and careers guidance community have demonstrated real commitment to taking this initiative forward in their own organisational contexts. It is clear that lessons need to be learned regarding the level of support necessary to enable these stakeholders to grasp the necessary steps needed to embed practice that integrates the full potential of LMI for All.

Table 5.2 Summary of updating Data Costs[14]

Data source	Indicators in LMI for All database	Variables	Updating costs (resources required,days of SRT, RT, RST & TTT)			
			SRT	RT	RST	TTT
Working Futures (combination of LFS and BRES)	Total number of jobs by detailed type (historical time series)	Where possible all data available at SOC2010 4-digit occupations. Also covers: Industry; region; gender; employment status; and highest qualification held.	1	0	2	16
Working Futures (combination of LFS and BRES)	Projected employment (2002-2012)					
Working Futures (combination of LFS and BRES)	Expected replacement needs (total job openings 2002-12)					
ASHE/LFS	Typical pay (mean weekly pay) extended to include medians, deciles, part-time pay		5	30	30	16
ASHE	Typical hours (mean weekly hours)					
ASHE	Changes in pay 2012-13					
ONS Standard Occupational Classifications 2010 – Structure and descriptions	Occupational descriptions	SOC2010 4-digit occupations	0	0	0	0
US O*NET	Skills and abilities		5	5	30	16
LFS	Unemployment rates		1	0	5	16
UKCES ESS	Current vacancies		3	0	1	16
DWP Monster – UJM	Time series of vacancies (DWP JCP replacement)		2	0	2	16
DWP Monster	Types of vacancies (Fuzzy matching “patch”)		0	0	0	0
Census of Population 2011	Location of jobs, workers by occupation, jobs by industry, travel-to-work distances (per new indicator)		5	0	1	16
HESA	Graduates first destinations		2	0	5	16
Total	All indicators		24	35	76	128

[14] Excludes update of *Working Futures* database – (£100-200K) and development of a replacement to DQP JCP vacancy series

Annex A: Core data sources included in LMI for All

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 Employment, Pay and Hours, 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). Further, more technical, details are contained in Li and Wilson (2015).

The remainder of this Annex is structured as follows:

- The remainder of this section sets out the rationale for the general approach and describes the information placed into the public domain.
- Section A.2 summarises the case for making detailed data on Employment, pay and Hours 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 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.6 describes the Checking Algorithm used to avoid publishing unreliable estimates of 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.
- Section A.9 concludes by providing a summary of the main data on employment, pay and hours provided in the LMI for All database.

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 personal and 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 SOC2010, 369 Occupational categories);[15]
- Sector (up to the 2-digit level of SIC2007, about 80 industry categories);
- Geographical area (12 English regions and constituent countries of the UK);[16],[17]
- 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.

---

[15] 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.

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

[17] 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.[18] 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:

$$Ln(E) = a + b*A + c*A2 + D*X + u$$

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)[19];
- Occupation (default is Chief executives and senior officials, plus 368 other 4-digit SOC2010 categories);
- Qualification (default is a degree or equivalent and 5 other qualification categories[20] (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 can be extended to include other indicators (such as the *median* or *quartiles*). During Phase 2A, the focus was on mean pay only. In Phase 2B this was extended to provide median and decile estimates based on an assumption of the distribution of pay being log-normally distributed

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

[18] 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.
[19] 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.
[20] 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 below).

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. 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.

Table A.1 Typical Earning Function Results

Variable	Coefficient
Age (continuous variable)	0.06
Age squared	-0.001
(default =male)	
female	-0.10
(default =London)	
North East	-0.08
North West	-0.10
Yorkshire & Humberside	-0.17
East Midlands	-0.16
West Midlands	-0.16
Eastern	-0.19
South East	-0.18
South West	-0.20
Wales	-0.21
Scotland	-0.14
Northern Ireland	-0.21
(default =degree or equivalent)	
Higher education	-0.11
GCE A Level or equivalent	-0.18
GCSE grades A-C or equivalent	-0.24
Other qualifications	-0.28
No qualification	-0.33
(default=Agriculture, etc.)	
02 Coal, oil & gas	0.62
03 Other mining and quarrying	0.13
04 Mining support	0.08
05 Food products	0.01
06 Beverages & tobacco	-0.03
07 Textiles	0.06
08-75…….etc, etc	*
(default =Chief executives and senior officials)	
1120 'Elected officers and representatives'	-1.08
1121 'Production managers and directors in manufacturing'	-0.67
1122 'Production managers and directors in construction'	-0.43
1123 'Production managers and directors in mining and energy	-0.10
1131 'Financial managers and directors'	-0.39
1132 'Marketing and sales directors'	-0.33
1133 'Purchasing managers and directors'	-0.25
1134-9274……etc, etc	*
9275 'Leisure and theme park attendants'	-1.77
9279 'Other elementary services occupations n.e.c.'	-1.30
constant	5.86

Notes: LFS 2013 full-time regression results. 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,[21] 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?

[21] In practice, the *Working Futures* database does not use the BRES data as such, but makes use of the various sectoral employment time 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.[22] The *Working Futures* database is the result of a complex combination of datasets, models and assumptions (including various iterative procedures).[23,24]

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).[25]

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).

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. It has been agreed with ONS that such data is, therefore, not disclosive. The aggregation of those few industries into the 75 slightly broader categories mean that ***NONE*** of the *Working Futures* data is regarded by ONS as disclosive.

---

[22] 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 (75 categories); occupation (25 2-digit SOC categories, 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); 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.

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

[24] 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.

[25] 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.

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.[26] It was agreed that it would be helpful in future 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.

ONS agreed to release data at a more detailed level into the public domain (at the level of the 75 industries aggregated up from 79 as discussed above). This only required a modest increase in the level of detail made available.

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

A.5 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 further development work as and when such data become available. 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 to medium-term, it is recommended that the database continues to be based on predicted values throughout.

A.6 Checking Algorithm to avoid publishing unreliable estimates

A checking algorithm is built-in to the API to avoid 'publishing' estimates that might be regarded as unreliable. This algorithm checks roughly whether or not the employment numbers concerned would be likely to be regarded as 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 involved in generating the weekly pay and hours worked estimates in the LMI for All portal. The data used are taken from the UK Labour Force Survey (LFS) and the Annual Survey of Hours and Earnings (ASHE). The analysis adopts the 2010 Standard Occupational Classification (SOC2010). The same approach has been applied to the LFS and ASHE data wherever possible.

The use of "raw" data from the LFS or ASHE in the LMI for All data portal is limited due to sample size and concerns about confidentiality. Reliance on the "raw" data would result in huge gaps in the information available to be presented in the portal. To get around these limitations the portal uses "predicted pay" estimates, based on an econometric analysis of the ASHE and LFS data sets.

In order to provide additional details by age, as well as features of the distribution of pay such as deciles, supplementary equations are used.

The discussion in this section describes the specification and estimation of the earnings functions. It also describes 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 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 into 6 sub-sections. This first sub-section (A.7.1) provides a brief introduction. Section A.7.2 explains how the LFS and ASHE databases are constructed and introduces the definition of earnings and other variables used in the analysis. Section A.7.3 discusses the specification of the earnings functions and how the estimated results are used for predicting pay. Section A.7.4 explains some supplementary analysis focusing on mean pay which is used to generate prediction by age 'on the fly' in the LMI for ALL API. Section A.7.5 compares ASHE and LFS, outlining the advantages and limitations of both datasets. Section A.7.6 concludes this discussion.

A.7.2 Data and definitions

Pooled samples from the UK Labour Force Survey (LFS) and the Annual Survey of Hours and Earnings (ASHE) are created to derive the pay estimates for the construction of the career database. LFS and ASHE complement each other in various aspects.

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 2012 and 2013 are pooled together to form an aggregated sample of 288,937 different individuals covering two years. For the purpose of this exercise, the pooled sample is further constrained to employees aged 16 and over, leaving 86,828 full-time employees and 33,608 part-time employees for the 2013 pay estimation.

ASHE originated from the New Earnings Survey (NES) which was started in 1970 and carried out each year subsequently. It is the most comprehensive source of 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). It covers 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. ASHE currently has a sample size of around 180,000 employees in the UK. The selected sample covers about one per cent of the whole working population in the UK.

Compared to the LFS, ASHE has the advantages in that it has more reliable pay information which is provided by employers rather than individuals and it has a larger sample size than LFS. However, information on individual characteristics is limited in ASHE and it does not have any information on education or qualification. In order to get around these problems, the LMI for All database is based on a set of estimates/predictions of pay using data from both ASHE and LFS. In addition, 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). A pooled sample for pay estimation is constructed by including the 2012 and 2013 waves of ASHE, constraining to employees aged 16+. It has 237,117 full-time workers and 110,810 part-time workers in the core research sample.

Gross weekly pay is used in all the pay estimations based on LFS and ASHE. Here the term Pay is generally used, although following standard conventions the term “earnings equation” is used to refer to the econometric equation estimated to predict pay. The main earnings equations estimate follows the well-established tradition pioneered by Jacob Mincer (Mincer, 1974). The econometric analysis adopts the standard “Mincerian” earnings function or earnings equation. This is the “main” earnings equation as described in section A.7.3. below.

The pay variable used in the LFS is “GRSSWK”. It is the gross weekly pay before deductions in an individual’s main job. It applies to employees and those on a government scheme but not those employed on New Deal, in the voluntary sector, or the environmental task force. Information on components of gross wage and the contribution of each component is not available in LFS. The pay variable used in ASHE is “GPAY”. It is the average gross weekly earnings in the reference period from either the main job or another job. Its main components are basic gross weekly earnings and allowances. The other components include overtime payments, incentive/bonus payments that relates to this pay period, and additional premium payments during the pay period for shift work and night or weekend work not treated as overtime.

The predicted mean weekly pay estimates in the LMI for All database are generated using the main earnings equation. These initial predictions are then adjusted using an iterative RAS procedure to match the published pay figures from ASHE and the LFS across each of

the main dimensions/characteristics (gender, region, industry, occupation and qualification).[27] This process is described in more detail in Section A.7.6.

In order to generate predictions of pay by age in the database supplementary age equations are estimated. These results are then use to predict pay by age based on the mean value for all ages. This is done “on the fly” in the LMI for All API.[28]

Similarly, predicted median and decile pay levels are based on parametric methods and the assumption that pay is log-normally distributed.

Note that there is no pay data available for the occupation 'Armed Forces' in the LMI for All database.

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 earnings equation and how this has been estimated. 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 earnings function has been run using the log of gross weekly wage as the dependent variable. The independent variables included are as far as possible identical in the LFS and ASHE earnings functions. Their definitions, are as follows:

- Age: a continuous variable ranged 16 to 84 in LFS and 16-93 in ASHE;
- Age squared: continuous variable;
- Gender: male and female, 1 dummy variable for male (base category: female) (same in LFS and ASHE);
- Region: 12 government official regions of England or devolved countries within the UK, 11 dummy variables in the regression (base category: London) (same in LFS and ASHE);
- Highest qualification: 8 QCF levels of qualifications, QCF1-8, and no qualification. and 8 dummy variables in the regression (base category: QCF8). Information on highest qualification is only available in LFS, thus regressions conducted based on ASHE are without any education measures.
- Industry: standard 75 categories as used in Working Futures, 74 dummy variables in the regression (base category: Agriculture, etc.) (same on LFS and ASHE);
- Occupation: 4-digit SOC2010, 369 categories and 368 dummy variables in the regression (base category: 115 Chief executives and senior officials) (same on LFS and ASHE).

---

[27] RAS is an iterative process used to reconcile row and column totals of a two dimensional data array with some target figures. See McMenamin et al. (1974), Toh, (1998), Miller and Blair (2009) and Lahr and Mesnard (2004) for a general discussion of RAS methods.

[28] API stands for Application programming interface.

Interactive terms have also been included to detect heterogeneity across different groups (these are the same in LFS and ASHE):

- 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 2012 and 2013 It is interacted with industries to control time trend differences within each industry. The base groups are industries in 2012.
- 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 2012.

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 working in the Agriculture sector and are Chief executives or senior officials in 2012. The log expected wage for an individual with these default characteristics at 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 generate predictions of pay by age and provide an indication of how far pay of each age varies from mean pay for all ages, “supplementary age equations” and ratios between pay of a particular age category and mean pay of all ages are developed which enable calculation of variations of pay by age ‘on the fly’.[29] These supplementary age equations and age ratios reflect how age affects the deviation of pay from the mean pay in groups with different combination of characteristics. Supplementary age equations were

[29] 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.

performed to derive the estimated pay at each age in a particular combination defined by four dimensions including occupation, gender, full-time or part-time working and the highest level of qualification. To ensure a reasonably large sample size of each combination, occupation has been defined for this purpose at the broad 1-digit level (covering 9 categories). To provide information of all possible aggregates, an extra category for all occupations has also been included. The occupational categories are as follows:

- Managers and senior officials;
- Professional occupations;
- Associate professional and technical occupations;
- Administrative and secretarial occupations;
- Skilled trades occupations;
- Personal service occupations;
- Sales and customer service occupations;
- Process, plant and machine operatives;
- Elementary occupation;
- All occupations.

For the same reason, the highest level of qualification held has been classified into three broad groups plus an aggregated group for all qualifications:

- High: QCF Levels 4-8;
- Medium: QCF Levels 1-3;
- Low: no qualifications;
- All qualifications including “High”, “Medium” and “Low” qualifications.

Gender and full-time or part-time workers both have two categories and an aggregated total. Across all the four dimensions this gives a total of 360 combinations. Industry is not included here because the sample size tends to get very small once industry is considered. It is assumed that patterns by age are common across industry one these other dimensions have been taken into account.

The main objective of the supplementary age equations is to provide a descriptive summary of how pay varies by age (all else equal). Thus the factors included in the linear supplementary equations are much more limited and only include average gross weekly pay of each age group (the dependent variable), age and age squared to generate the age coefficients that depict the age curve in each group. (Note: the most common finding in the literature is that the relationship between age and pay is an inverted U-shaped with pay peaking in middle age and declining smoothly thereafter).

The supplementary age equation is estimated using the ordinary least square method and is based on the pooled 2013 LFS data. The estimated coefficients of the independent variables and the constant term are used to derive the expected wage for an individual at a particular

age. The age equation is performed for each of the 360 combinations to derive the age coefficients and constant term. The expected pay at each age from 20 to 65 is calculated subsequently by applying the estimated coefficients to the value of age and age squared. (Note: some ages are missing in some combinations, the estimated coefficients are applied to those ages to derive their expected pay).

To provide an indication of how the expected pay at each age between 20 and 65 is distributed around the mean pay of all ages in each combination, the mean pay of all ages in each of the 360 combination is calculated from the LFS data. A ratio between the predicted pay of a particular age and mean pay of all ages in a combination is derived to indicate the distance of pay of an age from mean pay. The ratios calculated for each age then enable a prediction of pay by age around the mean pay to be made.

A.7.5 Median and deciles

Median and deciles are used to describe the distribution of pay. For a normal distribution, median and other deciles can be predicted using mean and standard deviation. The pay distribution in ASHE and the LFS is not normally distributed, however the natural log of pay tends to follow a normal distribution. Consequently, by converting pay to log pay it is possible to use the log normal distribution of pay to predict the median and deciles in the log-normal distribution. In order to generate predictions of pay for medians and deciles in LMI for All, supplementary "distribution equations" are used, based on analysis of LFS data.

The median and deciles analysis is to show how median pay (and other deciles) typically vary around mean pay of a selected dimension (for example, mean pay of an industry, or mean pay of an occupation, etc.). This assumes that the pay distributions are otherwise the same across the other main dimensions such as gender, region, etc.

The formula used to compute median and other deciles pay follows the property of the log-normal distribution. For a selected dimension, the median or deciles of log pay equals to the mean of log pay plus the relevant z score times the standard deviation. z scores measure how far away the decile or median of interest is located from the mean in a normal distribution, (or in another words, how many standard deviations it is away from the mean).They are known for any specified deciles or median and can be obtained from the standard normal cumulative probability table. They are fixed values in the normal distribution and are the same for any selected dimensions with a normal distributed log pay measure. Given mean log pay and the standard deviation of a selected dimension and z scores, the median and deciles of log pay can be predicted. Exponentiation is needed to convert the log pay back to Pay.

However, the mean of log pay of a selected dimension is normally not available directly. Given that median equals to mean in a normal distribution and median log pay equals log median pay in a log normal distribution, the median level of pay for a category can be estimated by assuming the ratios of median to mean are common to a small subset of categories chosen arbitrarily. Using ASHE published figures on median and mean pay for 2013, the ratios of median to mean are calculated. The ratios are applied to the mean pay of a select dimension to generate the median pay of this dimension by assuming same ratios

apply across all other dimensions of the database. The median log pay are calculated subsequently for prediction of log pay at other deciles.

Ideally estimates of mean and standard deviation are needed for all the main dimensions, but limitations of sample size in both LFS and ASHE imply this is impossible for all possible permutations and combinations. Inspection of the data suggests that variations are greatest by status (FT/PT), industry and occupation. Log mean pay and values of σ have therefore been estimated across FT/PT, industry and occupation and similar patterns are assumed to apply across all other dimensions for the purpose of this calculation. Typical values are assumed, based on variations across the main dimensions of interest (but not all possible cross dimensions).

A.7.6 Concluding remarks on pay predictions

This section of Annex A 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 2013 results are based on the UK LFS and ASHE. The same methods and analysis are applied to LFS and ASHE. Although ASHE has a number of advantages compared to LFS, it does not provide any information on education, thus it will not be possible to include the same highest qualification variable as in LFS. Thus the estimated coefficients derived for other variables using ASHE are overestimated because they are taking account of education effects (omitted variable bias). The estimates from ASHE therefore are not 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 original project plan included pay, hours and employment, broken down into as much detail as possible by:

- Occupation (up to the 4-digit level of SOC2010, 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 it soon became clear that 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 indicators used have therefore been estimated, using data from *Working Futures* and using econometric analysis (earning functions, etc, as described in Section A.7).

This section sets out details for the algorithm used to constrain the estimates to match official ”headline” published figures. 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.

Employment RAS processes

Employment data at the 2-digit level are published in the *Working Futures* (WF) database (See Wilson and Homenidou, 2012a, 2012b). This dataset has been expanded from the 25

[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).

2-digit occupations in the WF dataset to 369 4-digit categories for LMI for All database. 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. In practice, this refinement was not made.[31]

In the longer-term, it is also necessary to think about how these patterns change over time and how to extend the projections to 2022 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 SOC2010. 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).

[31] As noted above, initial attempts to produce projections ran into problems for the category chefs. This occupation was part of a larger 2-digit occupational grouping for which employment was projected to decline sharply. It was not possible to generate plausible employment projections for chefs within that total. An amended set of 2-digit occupational projections was therefore produced for LMI for All which differed slightly from the original published *Working Futures* estimates.

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 RAS processes

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).

   A long-term aim might be to 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 (2013).[32]

[32] Other problems associated with the level of detail being attempted also mean that variation over time is difficult to capture precisely. Seethe discussion of the "change in pay" indicator elsewhere in this report for further details.

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 main 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 discussion in Section A.7.4).

Details of the Main RAS process for Pay

1. The starting point for the main RAS process for Pay is:
    a. A detailed database (based on the econometric analysis of ASHE and LFS data [33]) showing predicted pay across the various dimensions (sector, occupation, gender, region, qualification, status (but excluding 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 (weekly "pay bills" (pay*employment)) are consistent. This is analogous to the same way this is done for employment in the *Working Futures* employment database. In this case the RAS targets are set in terms of wages & salaries ("weekly pay bills", millions of pounds) rather than employment (thousands of people).
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

[33] ASHE is the preferred data set. However, because ASHE does not include qualifications some part of the regression analysis needs to be based on the LFS.

with ASHE, so this has been built in as a key feature. However, ASHE excludes qualification, so a second stage involving LFS data is also needed.

5. The main stages are as follows:

    a. Estimate the earnings equation (excluding qualifications) using ASHE data in the SDS;

    b. Estimate the earnings equation (including qualifications) using LFS data;

    c. Using the extracted parameters from 5a and 5b generate predictions of weekly pay for all the dimensions set out in 1a. In principle, this can be done for all possible ages. For the purpose of the RAS process pay is estimated for the average age for the group concerned;

    d. Based on the data assembled in 1b, generate a set of consistent 'RAS targets' for wages & salaries to constrain the dataset in 5c. These are based on mean pay data multiplied by the relevant *Working Futures* employment numbers (these are used as weights rather than the original sample weights.[34]

    e. An adjustment to these wages and salaries (weekly pay bill) estimates is needed to ensure that they add up consistently to the same totals when summed across different dimensions set out in 1a. This is done using a preliminary RAS process (discussed in more detail in Section A.8.3 below).

    f. Using these targets the data in 5c can be RAS'd to be consistent with those from 5d and 5e. This required the development of a new programme analogous to the employment one used for *Working Futures*.

    g. A key issue was how to deal with age, which is not a dimension in the *Working Futures* employment database. In principle, to adopt the same methodology would require the *Working Futures* database to be extended to cover every single year age group. However, there is no reliable source of data to provide such information cross-classified by all the other dimensions simultaneously. Even if such data were available the limits of the current Python programme have been reached. It was decided therefore to focus on 'all ages' for this step.

    h. The method developed therefore ignores age in the RAS process and imposes an average age for each detailed category based on LFS data.

    i. In addition to the coefficients for pay relating to the other dimensions a file of average age by 4-digit Occupation and gender is produced for this purpose. The predicted pay at this average age is what is input into the RAS process in steps 1a and 5c. The predicted pay at step 5c is made therefore for the average age for that particular combination of sector, occupation, gender, region, qualification and status.

[34] The *Working Futures* employment numbers provide a complete and consistent set of weights based on LFS and other sources as described elsewhere. There are some differences in these patterns and those used in ASHE, which warrant further investigation.

j. To respond to queries about pay by age in LMI for All, the API generates an estimate “on the fly ”based on the mean pay estimate for all ages and a ‘typical’ age earnings profile for each category. This is based on supplementary data, which shows how pay varies by age, all else equal. In particular this allows such patterns to vary by occupation. This allows a prediction of pay by age ‘on the fly’ in the API (based on the predicted mean pay from the main earnings equation and the age in the query).

Hours RAS processes

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 2013. 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.

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 suggest that such a methodology could deliver robust estimates.[35] A linear regression for working hours was estimated using both ASHE and 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. 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. The working hours equation was analysed for full-time workers and part-time workers separately.

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 Preliminary RAS processes – Generating the pay and hours ‘targets’

The basic data are taken from 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 themselves 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’.

[35] In practice a simpler method was adopted as described above.

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.

In the main LMI for All database estimates of pay are generated for full-time and part-time employees separately. The estimates cover just a single year (currently 2013).

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).

Data sources and methods for the Preliminary 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 were used. Ideally, the aim would be 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 SOC2000 is used by ASHE. More years are available for SIC2007, which is available from 2008 onwards. However even if there are no changes in classification the construction of a times series for all the very detailed categories is problematic given the statistical noise in the data. The final estimates are therefore made available for just a single recent year (currently 2013) with a separate indicator showing typical changes over the past 12 months for broad categories only.

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.

From the ASHE pay data, it was noted that for some specific occupations (such as florists) this method resulted in implausible pay levels for some categories. These anomalies occurred when there is an occupation that has no published figure. In this case the UK All value (full time plus part-time) was initially imposed. However in certain instances this value could be inappropriate. To overcome this problem the numbers were taken from ALL full-time results instead which resulted in fewer gaps and more plausible estimates.

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 can 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 cover all the main dimensions (369 occupations at the 4-digit level, 75 industries, 12 countries/English regions, gender, status), for historical and projected years. The projections were refined slightly to reflect problems encountered in developing consistent projections at the 4-digit occupational level. Estimates of replacement demand (RD) are generated 'on the fly' in the API, based on an assumption of common RD outflow rates (the same rates for all the 4-digit categories within a particular 2-digit category). Relevant data and instructions are provided in the Wiki.
2. ***Pay***: the detailed mean pay estimates again cover all the main dimensions but just for a single year (2013). They are based on a combination of ASHE and LFS data. The detailed estimates are supplemented by additional information, which provides the parameters necessary to generate estimates of pay by age as well as medians and deciles. These estimates are created 'on the fly' within the API. Some limited information on changes in pay between 2012 and 2013 is also provided.
3. ***Hours***: the hours database again covers all the main dimensions but just for a single year (2013). It is based on ASHE data. It 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 are for an N-dimensional data array, with the following dimensions

- Year – 2000-2022;
- Gender – 2;
- Status – 3;
- Industry – 75;
- Occupation – 369;
- Geography – 12;
- Qualification – 9; and
- Weight – number of people employed.

The first column is the 'year', which runs from 2000 to 2022. 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. Many cells have fewer than 1,000. In these cases the API flags this up or suppress the numbers replacing them by more aggregate information as set out below. There are two main possibilities:

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 SOC2010.

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 can be aggregated to a six fold one in which the higher levels are combined (this is equivalent to the 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

If the numbers in a cell are too small to provide reliable information the API:

1. Generates the sub-totals; and/or
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;

Any query relating to Replacement Demands is dealt with analogously to employment[36]:

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" levels of granularity in the data, but if that query returns "no reliable data available" it offers the option of searching on a more "coarse" level of granularity. It is, of course, possible to pre-set the query to obtain 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 relate to just a single year (2013). The file for the N-dimensional data array, includes the following dimensions:

- Year – 2013
- Gender – 2 (male, female);
- Status – 2 (full-time and part-time employees);
- Industry – 75 standard industries (see classification and aggregation below);
- Occupation – 369 4-digit SOC2010 categories;
- Geography – 12 (UK countries and English regions); and
- Qualification – 9 National Qualification Framework levels.

The first column is the 'year', currently just for 2013. The second to seventh columns, from 'gender' to 'qualification', indicate the characteristics of people covered by the dataset.

[36] Queries to Replacement Demand through the API were, at the time of writing, to be imminently implemented.

Two next columns have the information needed to calculate mean pay:

- Employment - the relevant total employment number that should be used for weighting (based on Working Futures);
- PayBill - The Total (weekly) Pay bill for the category concerned (Pay*Employment).

The ratio of the latter to the former represents the average weekly pay for the year specified in the first column and with the characteristics in columns two to seven.

In order to assess reliability of the estimate, the API checks the corresponding employment 'weight' from (based on data from the *Working Futures* Employment dataset). The weight 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 flags this up or supresses the pay or hours estimates, replacing them by more aggregate information as set out below. As for employment there are two main possibilities:

1. Replacement by a *sub-total* across one (and/or, if necessary more than one) 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 solution for hours and pay defaults to:

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 from the mean value and the selected age as well as medians and deciles. See section A.7.5 for more details.

Annex B: O*NET

B.1 Introduction

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 was further explored. This Annex describes that process in more detail, including developments made in Phase 2B.

B.2 Initial Approach

The initial approach followed the one developed by Dickerson and Wilson (2012). They 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.[37] Dickerson and Wilson (2012) 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.[38] 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.

The method involved the following steps:

---

[37] CASCOT (Computer Assisted Structured COding Tool) is a piece of specialist software, originally developed by the Institute for Employment Research at the University of Warwick, designed to classify occupational title into Standard Occupational Classification (SOC) categories.

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

- 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.

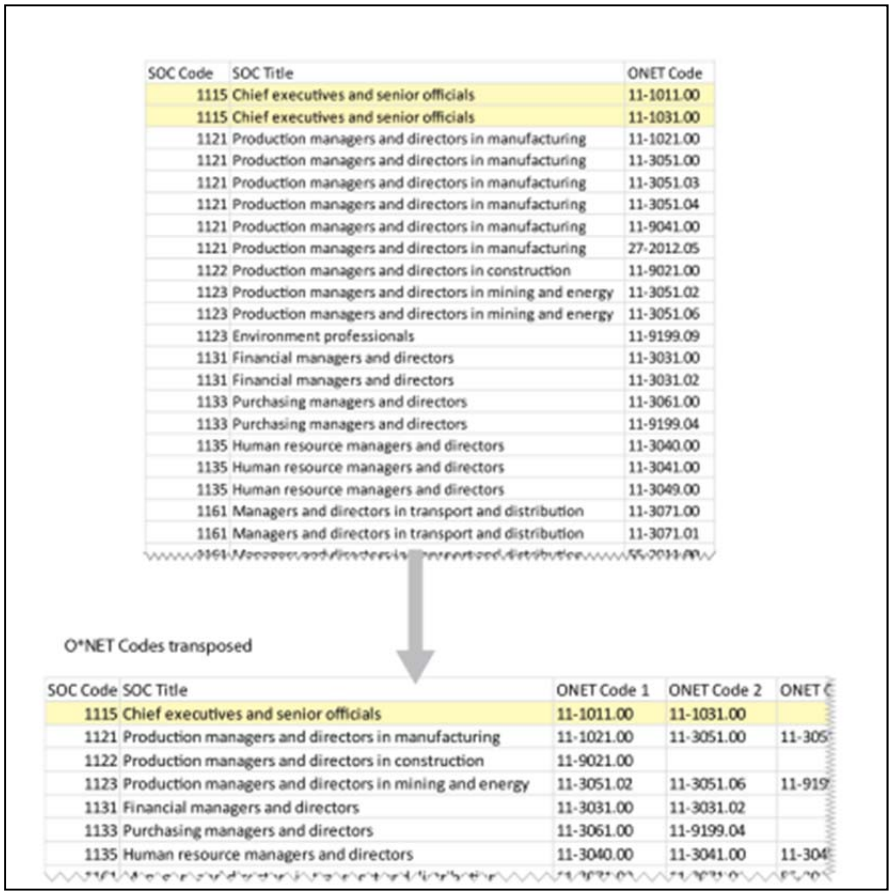
Rather than creating a SOC 4-digit O*Net database (369 categories) the emphasis therefore shifted to one of linking directly from the 369 SOC 4-digit categories (and the underlying 28,000 SOC occupational titles (effectively SOC 6 digit)), recognising that there may be no unique mapping, but links to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

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.

Table B.2 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.

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).

ONET 2010 SOC

Initially the mapping between US SOC to UK SOC was completed using database 15.1 (US SOC2009). Due to the availability of a newer ONET database (19.0), which includes US SOC2010, the mapping was subsequently updated. This was based on an update of the initial mapping to SOC2009. The main steps were as follows:

1. Worksheet containing mapping (UK SOC2009 to UK SOC2010) was used as the basis, which was then updated to the latest US SOC (2010).
    a. Using an automatic matching method ("=INDEX(MATCH())"), the US SOC was updated to SOC2010.
    b. This was further refined and checked by hand.
2. These new mappings to the latest SOC2010 were then checked and further occupation groups were added as additional options.
3. The data was then tidied to give a one SOC to many O*NET (SOC2010) occupation code.

O*NET 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.

Initially a decision was taken to focus on 'ONET 15' in Phase 2A of the LMI for All project to maintain consistency with what had been done previously. This was subsequently updated to ONET 19.0 (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. Rather developers are simply provided with the full O*NET skills database linked to US SOC2010 plus a look up table that goes from one UK SOC2010 4-digit category to one or more US O*NET SOC2010 categories. This means that there is not a simple one to one mapping from UK 4-digit occupational categories to a corresponding US one. In order to find information about relevant skills associated with a particular occupation, developers may need to consider the skills in one or more US occupations.

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.3 to B.5), 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.3.

Table B.3 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 for All database and could be omitted. For the moment the O*NET file is included in its entirety.

Table B.4 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).

Similarly, Table B.5 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 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

Table B.5 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

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 was extended to cover the full range as shown in Tables B.3 and B.4.

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 can potentially provide useful information relevant for the database (e.g. required training, etc.).

The full LMI for All database now includes:

- **Abilities.txt** These data come from the O*NET website. The information shows the level of abilities required and the importance of these abilities for the occupation concerned.
- **Skills.txt** These data come from the O*NET website The information shows both the levels of skill required and the importance of these skills for the occupation concerned.
- **Interests.txt** These data contain the Interest data associated with each O*NET-SOC occupation.
- **Education, Training, and Experience Categories.txt** **t**hese data contain the categories associated with the Education, Training, and Experience content area.
- **Education, Training, and Experience.txt** These data contain percent frequency data associated with Education, Training and Experience Content Model elements associated with each O*NET-SOC occupation.
- **Job Zone Reference.txt** These data contain the Job Zone, Name, Experience, Education, Job Training, Examples, and SVP Range.
- **Job Zones.txt** These data comes from the O*NET website within 'db_15_0.zip', the file contains each O*NET-SOC code and its corresponding job zone number.
- **Knowledge.txt** These data contain the Knowledge data associated with each O*NET-SOC occupation.
- **Occupation Data.txt** These data contain each O*NET SOC code, occupational title, and definition/description.
- **Occupation Level Metadata.txt** These data contain the Occupation Level Metadata associated with each O*NET-SOC occupation.
- **Task Categories.txt** These data contain the categories associated with the Task content area. Categories for the scale Frequency of Task (FT) are included.
- **Task Ratings.txt** These data contain the Task Ratings data associated with each O*NET-SOC occupation.
- **Task Statements.txt** These data contain the Task Statements data associated with each O*NET-SOC occupation.

- **Work Activities.txt** These data contain the Content Model Work Activity data associated with each O*NET-SOC occupation.
- **Work Context Categories.txt** These data contain the categories associated with the Work Context content area.
- **Work Context.txt** These data contain the Work Context data associated with each O*NET-SOC occupation.
- **Work Styles.txt** These data contain the Content Model Work Styles data associated with each O*NET-SOC occupation.
- **Work Values.txt** These data contain the Content Model Work Values data associated with each O*NET-SOC occupation.

Annex C: Other data considered for inclusion but rejected

C.1 Introduction

This annex briefly summarises a number of potential sources which it was considered might enrich the LMI for All database, but which for one reason or another it has been decided NOT to proceed. 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 (other indicators)
C.6 Cedefop – pan-European employment projections
C.7 Other European datasets
C.8 Course information

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 was 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), 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 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, cross-tabulations of variables from APS microdata can yield a little more 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 limited so LFS data only are used.

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 was narrowed down to providing unemployment rates. However, the extra value of the APS in this regard is 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.

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) can 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 always clear whether this would allow the freedom to distribute such data publicly by including it in the database. This requires further negotiations with the UKDA and ONS. The marginal benefits of doing this (in terms of value added to the database) are modest and the marginal costs quite high. It was therefore decided not to proceed further with this data set.

C.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[39]). 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[40].

Available data include:

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

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.

[39] 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.

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

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 SOC2010 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.

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 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 (SOC2000, 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.

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) 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 (SOC2000, 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.

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. Proportion of the unemployed in each occupational category.

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 available from NOMIS (as 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

Some data from the Census of Population are already included in the LMI for All database, but this only scratches the surface of the full potential the Census offers. However, much of the information available is of more interest to users other than those concerned with careers. In particular, there are many indicators relevant to those with interest in local economic development and local labour market issues. The discussion in this sub-section explores this potential in more detail.

The Census of Population provides the most complete source of information on the characteristics of the population of the UK at the time at which it is conducted. It covers around 95 per cent of the population, and the data when published is adjusted to provide a complete count of the population.

The most recent UK Census of Population was undertaken on March 27th 2011, but the final data sets from the Census were only published in spring 2015. The Census is undertaken by the three statistical offices of the UK – the Office for National Statistics in England and Wales, the General Register Office and National Records of 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 between the four nations of the UK 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 schedules for release of data vary between countries and only a subset of the data is harmonised and released at a UK level. This time, Scotland has been slower than the rest of the UK in publishing data from the Census, with its final outputs produced in May 2015. Moreover, the publication schedule for all parts of the UK has been greatly slowed by the demands of statistical disclosure control, under which outline tables have been announced, but later modified because of the risk of disclosing confidential information. This is a particular issue, because the Census yields data for very small geographical areas (called Output Areas) which are designed to represent relatively socio-economically homogeneous neighbourhoods (of around 200 households) and hence tables for these areas may be based on very small numbers of people (and there is a possibility that tabular information might be recognised as representing identifiable individuals).

The strengths of the Census are its very high response rate 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 main drawbacks of the Census are that it is only collected once every ten years and that it takes nearly two years before final results start to become available (because of the amount of work involved in processing the data). Hence, it 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 Census includes questions on the characteristics of the job held by an individual, including where it is located. It yields information on both the employment profile of people living in an area and the breakdown of employment located in an area (using a slightly different "workplace geography" which is designed to preserve the confidentiality of employers). The types of analysis which are possible with Census data include calculating a profile of employment within a given radius of a job-seeker's location, or alternatively, calculating the number of workers living within a given distance of a concentration of employment. It is also possible to analyse commuting behaviour and identify differences in distances travelled to work by type of job and type of worker. Typically, higher status (and better paid) occupations tend to have more geographically extensive local labour market areas.

Labour market indicators yielded by the Census include employment by industry (section) and occupation (major or sub-major groups), unemployment by industry and occupation, and the qualification profile of the population. Employment is broken down by gender, age group and ethnic group. The Phase 2A report contains an extensive list of indicators, which can be derived from the Census.

There are three main issues concerning the integration of data from the Census into the LMI for All database.

The first is the delay in availability of data and the differences in publishing schedules for the three Census Offices. The bulk of labour market information from the Census was only published in summer 2014 for England, Wales and Northern Ireland and publication of data for Scotland will only be completed in Spring 2015. UK-wide outputs cannot be made available until data on a topic is published by all three Census Offices and hence the publication of UK-wide tables will be completed in mid-2015.

Second, data on mode of travel-to-work (i.e. by car, public transport or other) was published by late 2014, but the publication of data on travel-to-work patterns was affected by concerns over confidentiality. Less detailed information was published via NOMIS in summer 2014, but the most geographically detailed information can only be accessed in a secure ONS environment and it is unlikely that data for small areas can be incorporated into LMI4All because of the risk of disclosure of confidential information.

The third issue is the level of detail for which Census data is made available. Most of the data on employment or unemployment by occupation is for SOC major or sub-major groups (though occupation is cross-tabulated by a number of dimensions, including age, ethnic group, highest qualification and age). Frequency counts on employment for the 3-digit level of the SOC has been published, but cross-tabulations by other variables and tables involving more occupational detail would require the commissioning of bespoke tables. A cost is involved, related to the amount of staff time involved in producing the table. Tables would have to be commissioned separately from each of the Census Offices.

Though it is becoming ever more outdated, the Census still provides the most detailed information on the geographical location of jobs and workers by occupation and is the only source of information on how the labour market matches workers with employment opportunities within cities. It is, therefore, worthwhile seeking to include Census information.

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

- Labour market and employment data from the Census;
- Commuting and workplace data.

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 were published over the next 3 years. The first results on the characteristics of the population were 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 2012 and February 2013 in England and Wales, in January 2013 in Northern Ireland and March 2013 in Scotland.

More detailed two and three-dimensional tabulations were published in the form of Local Characteristics and Detailed Characteristics. Local characteristics tables were mainly based upon those produced for the 2001 Census. The design of Detailed Characteristics tables changed during the output process as the need to reduce detail in order to preserve confidentiality became apparent. The publication schedule for the 2011 Census has experienced many more changes than usual and there have been revisions to a number of tables, which have had to be re-released

In England and Wales, publication of Detailed Characteristics tables started at local authority level in the third release of Census data starting in May 2013, and data for Middle Super Output Areas and electoral wards followed. Publication of Local Characteristics tables started 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 was 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 commenced in Summer 2013,

while Local Characteristics tables were be released from Autumn 2013, and publication of Detailed Characteristics tables started in Winter 2013.

The Census Offices commenced production of flow matrices for journey-to-work and migration, microdata and UK-wide tables in mid-2014. The flow tables provide considerable geographical detail, but their topic detail is limited. 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. These data sets make it possible to cross-tabulate any variable by any other variable. However, the detail of some variables (notably geography) is limited. The level of detail made available is much reduced in the most accessible versions of the datasets. 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 and the amount of detail in cross-tabulations is limited because their statistical reliability declines with sample size. The Census Offices will also produce bespoke tables commissioned by users of the Census, drawing upon the entire Census dataset. However, 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 were published:

- Employment characteristics of people resident in an area;
- Characteristics of people working in an area;
- Information on travel patterns of people in work. From this it is 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. The simplest tables produced are the Key and Quick Statistics, available for all geographical levels across the UK. The more detailed Local and Detailed Characteristics tables are 2 and 3 dimensional, and include breakdowns by age and gender. This information is available for electoral wards (above a specified population size threshold) and larger geographical areas.

The *indicators* which can be calculated relate to the labour market characteristics of the population, the nature of employment for working people living in an area and the characteristics of jobs located in an area.

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.

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

The possible *indicators* which can be derived include:

- 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

Commuting and workplace data

The bulk of data on the employment characteristics of workplaces will become available from the Workplace Population and journey-to-work tables, which are produced toward the end of the publication process (from mid-2014). A set of tables documents 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. 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 Census data on employment by workplace became available during 2014. The Workplace Population tables detail the characteristics of people working in "workplace

zones" and standard geographical areas. The new 2011 workplace geography reflects the geography of employment, enabling much more detail on the characteristics of employment to be released. The workforce is presented by age, gender, family status, mode of travel, SOC major group, NS-SEC, industry sector and ethnic group. The list below summarises the information available.

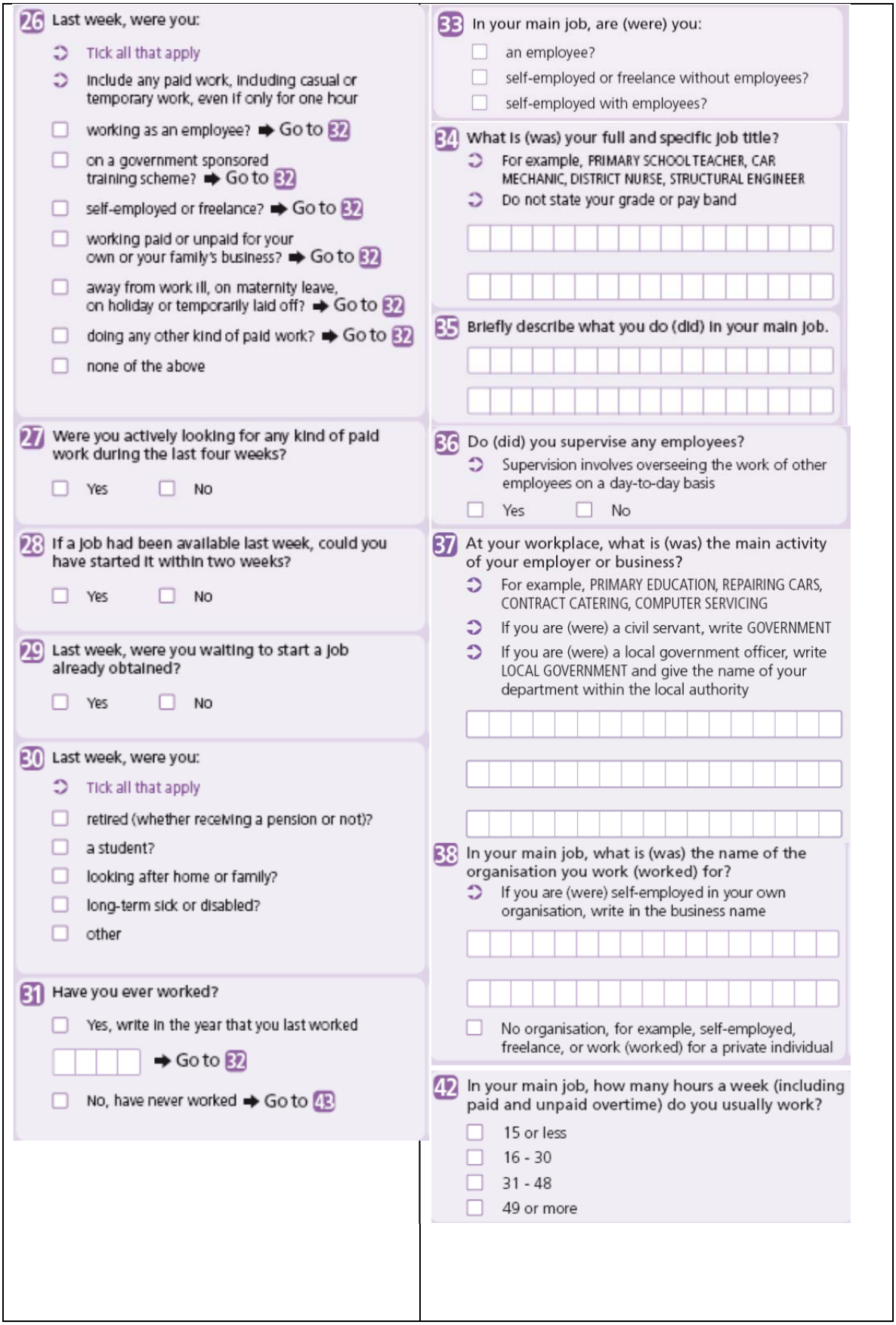
- Occupational profile (at least SOC2010 major group in all areas and potentially 3-digit SOC for England and Wales) of employment located in an area
- Industrial profile (at least SIC 2007 section) of employment located in an area
- Qualification (NQF) profile of employment located in an area
- Distance 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

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

The flow tables present the breakdown of workers involved in each commuting flow between their area of residence and work. 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. The flow tables, which are publicly available, only provided data on commuting by age and gender, with more detail available for flows between local authority districts than between small areas. The level of detail available is severely constrained because of fears over confidentiality, because of the small number of jobs located in many residential areas. Data on commuting by occupation and industry is only available via an ONS Virtual Microdata Laboratory and cannot be made public. .

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?

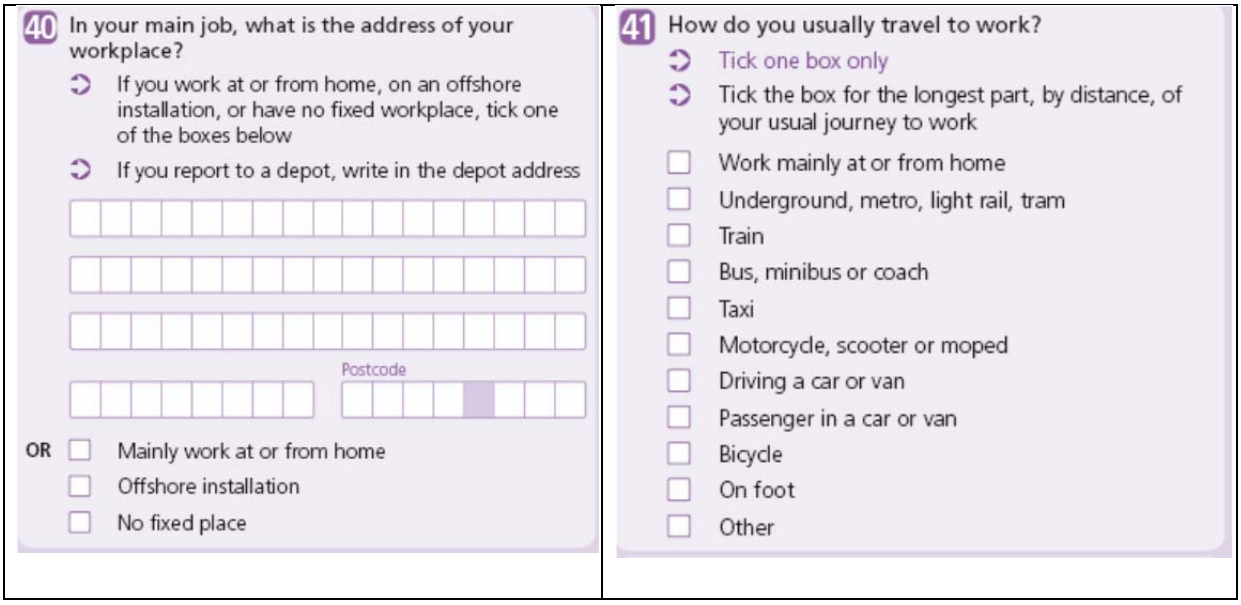
- If you are (were) self-employed in your own organisation, write in the business name

- ☐ 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



C.6 Cedefop database

Cedefop publish a range of skills demand and supply projections that are available in the public domain.[41] 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.

In principle, 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*.

However, the Cedefop data are presented using ISCO88 2-digit categories. In Phase 2A the team 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 is able to supply it in a user friendly form for the *LMI for All* project.

In principle, the data can also 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 SOC2010 (although a broad brush mapping can be derived (see note below)).
- The data to be published in early 2015 will use ISCO08. This is broadly compatible with SOC2010. 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

- Given the lack of 4-digit information and the limitations of mapping to SOC2010 this should not be a high priority for inclusion in the LMI for All database;

---

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

- If it were to be added in future, it would be best to use currently available 2-digit information, based on ISCO08, adopting a broad brush mapping to SOC2010 2-digit categories in the short term.

Note: on SOC2010 – ISCO08 Mapping

Full details of current ONS thoughts on mapping between SOC2010 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 SOC2010 (IER/ONS are working on this, see below).

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.

IER have done some work with ONS on a more detailed mapping. Table C.1 below shows an example of this work. It is clear that the mapping process is challenging and that a simple solution is not likely in the foreseeable future.

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	x	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.7 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, three candidate countries and three 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.8 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 requires a comprehensive mapping of courses to occupational codes. These issues are discussed in more detail in Section 2.4.3 of the main report.

An attempt was made to see if LFS data could be exploited to provide such information. In principle, this data source offers some potential insight. Survey respondents are asked questions about their formal qualifications acquired and hence course of study followed. To explore this possibility LFS data were extracted by 4-digit occupation cross classified by level of qualification held and field of study. Further cross- classification by other dimensions of interest such as geographical area (countries within the UK and English regions were also considered.

This exercise confirmed that in practice there are many problems and pitfalls with using such data that make it impractical to incorporate then within the LMI for All database. The most significant issue relates to problems of limited sample size. These mean that the data array at the level of detail of interest is very sparse. This means that it is not possible to provide meaningful responses to the vast majority of possible queries about what qualifications are associated with particular jobs (as defined by 4-digit occupational categories). This severely limits the value of such information in the context of LMI for All.

Queries at the level of detail that is meaningful from a careers guidance perspective (4-digit occupations combined with a detailed breakdown by both level and field of study) return zero entries in the vast majority of cases. Aggregation up to higher levels by occupation and across qualification categories eases such problems but at the expense of the detail required. The main employment indicators, which provides information on occupation by 4-digit occupation and broad level of qualification exploits that data to its limits.

Other problems also caution against reliance on this kind of information. The data (where the sample sizes are adequate) show the average qualification patterns for people of all ages. This may be very different form the qualification requirements for new entrants.

Annex D: Careers stakeholder preparatory questionnaire

Name	
Email	
Job title	
Organisation	
Please could you supply us with a brief written scenario of the type of questions and information a client/customer/claimant may ask in a typical one-to-one-session. Please also list some 'real world' questions.	
Currently, what type of labour market information do you most commonly use with your clients/customers/claimants?	
What are the gaps in labour market information you need for your business?	
Please specify the particular target group of clients/customers/claimants with whom you would want to use this application. What would be your priority for an application for this target group using the LMI for All database?	

Annex E: Hack and modding day feedback and developments

The developers

Twelve developers were selected for the second hack day, comprising: five developers who participated in the first LMI for All hack day; five from the UKCES careerhack competition; and two further developers who contribute to widening the skill set of the developers. There were eleven male developers and one female. Developers were selected based on their skill set to comprise teams in order to progress Phase 2A hacks. Developers are variously involved in: accessibility and open data; UX front-end and back-end development; product management; IOS developments; social and mobile apps development; RS 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.

The careers stakeholders

Fourteen stakeholders and experts in the use of data in careers guidance attended the hack and modding days. Careers stakeholders represented a range of sectors, including: education; charity; and the public and private sector. 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 careers websites and media developers.

Prior to the hack day, the careers stakeholders were invited to complete a pre-event questionnaire (a copy is included in Annex D). The anonymised responses were sent to the developers as background information. The careers stakeholders provided the identified typical career guidance questions, including:

Learning

What type of course/qualification would be best for me?

Is there funding available for the course I need, where can I find this, how do I apply?

What is the difference between a grant, bursary and loan?

Careers

I do not know what career to choose, I just know I want a change, what do you think I should do?

Where can I find voluntary work relevant to the sector I want to work in?

How can I earn more?

How do I understand my value as an employee?

Job search

Where can I find a job that I will enjoy?

What are competencies? How do you explain an example in an interview?

How do I know what type of CV to use, what does tailoring your CV involve?

What is a cover letter?

What type of recruitment agency should I approach?

How do I explain dismissal/tribunal, criminal record, illness, redundancy, disability/learning difficulty?

Job information

What does a job entail and what hours will I be working?

How do I get into that job and how hard is it to get into?

What qualifications are required? Do I need to go to university to do this?

Are their many jobs in this field? If I move to X, will there still be a demand for Y job?

What else could I do that's similar?

How much would I earn?

What sectors are growing/declining?

Skills

Can you help me understand my skills?

How do I sell myself and my skills?

How do I know if my skills meet the job criteria?

Social media

How will using Social Media benefit me in gaining employment?

Is using social media safe?

Can I contact anyone on LinkedIn?

Where can I network to build up my contacts?

The most common type of LMI used and required by stakeholders included:

- Mostly job profile data – career titles, alternative titles, work activities, personal qualities and skills, salaries, entry routes and qualifications;
- Data on growth and shortage areas;
- Number of jobs in a particular field;
- Predictions for the future;
- Data disaggregated in terms of gender, salary and skills;
- Geographical information on where specific occupations are centered;
- Concept of local, national and international labour markets;
- Where to research industry-specific information.

Stakeholders reported that LMI was collected from a range of sources including: National Careers Service website; careers websites such as icould; RCU Ltd; UKCES reports; ONS; Sector Skills Councils; news websites, both local and national; and employer survey data. However, the following gaps in LMI that is needed for business were noted as:

- Regional information, especially salary data;
- Vacancy data, such as how many and location;
- Competition for jobs, such as how many vacancies, number of people graduating, and how many are unemployed;
- Skill Shortages, including long and short-term skill deficits; and
- Skills and qualifications needs of local employers disaggregated by industry and occupational level.

Interestingly much of LMI data identified were considered crucial to planning future career and learning goals, but that it is hard to find and that it often lacks relevance to an individual needing careers support. It was also noted that the availability of data varies by sector and LMI source.

When asked to think about potential LMI for All applications, the careers stakeholders suggested priority target groups to be: secondary schools; 14-19 year olds; prospective year 10 and 11 students; and parents. Priorities for applications based on LMI for All should:

- Include projected trends in opportunities;
- Be easy to use and engaging with data presented in a range of visual formats (i.e. heat maps, graphs and charts, etc.);
- Support the development and refinement of customer research skills, expand their thinking/understanding, and enable them to take ownership of their career;
- Include data that links to similar or related work areas and points to a range of progression opportunities; and
- Enable the user to interrogate data at a detailed level (either at occupational or geographical level).

One career stakeholder suggested that applications needed to be targeted and have a good idea of who is using it rather than just responding to a generic query of the data. That is, customised to individual circumstances and needs.

By attending these events, the careers stakeholders were able to get an update on the LMI for All project and learn about the progress over the year. It was their role to judge the applications developed during the hack day to help guide the development of a marketable application. The stakeholders provided feedback on the overall project as well as the applications developed during the day.

Key features of the hack and modding days

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, as well as information on previous hacks and the UKCES competition entrants and winners. Career stakeholder survey responses were also sent to the developers to provide an understanding of the different careers scenarios in which an LMI for All application could be used. The developers were also able to participate in google hangout to exchange initial ideas before the hack day. Prior to the hack day developers were put into teams. It was agreed that they would work on developing two application-prototypes, based on:

- **Job Quest v2**, which allowed users to create a character and give them skill points into various attributes. The player can then increase attributes and skills, and present the most appropriate job role.
- **Linda**, which used LMI and other data to help students choose careers. Through this development there was opportunity to expand with additional datasets and questions,

and use APIs to change the presented advice when circumstances change. It was thought that this could be merged with data and logic from On Demand, which pivots the API data around the roles which employers are finding hardest to fill, disaggregated by region.

- **JobBungee**, which helped people explore career plans by providing smart access to LMI data. It also displayed real live data from JobCentre Plus and sample CV data from LinkedIn.

The LMI for All project team were available throughout the day to respond to queries and fix any errors identified by the developers.

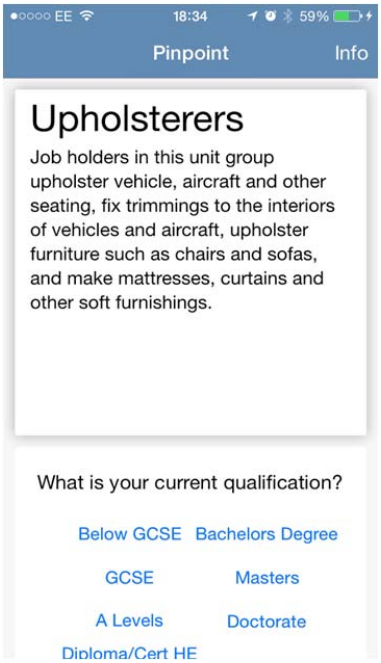
The following applications were developed and refined during the hack and modding days.

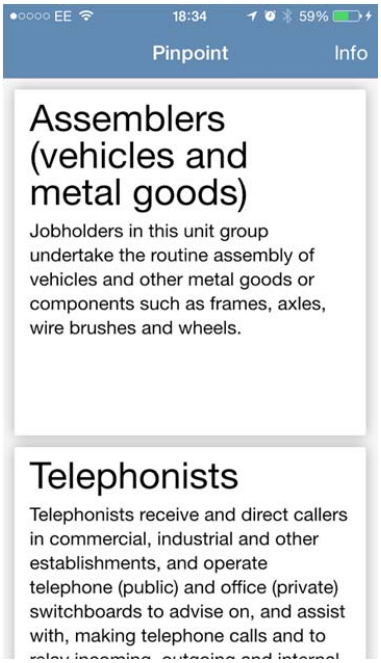
Pinpoint

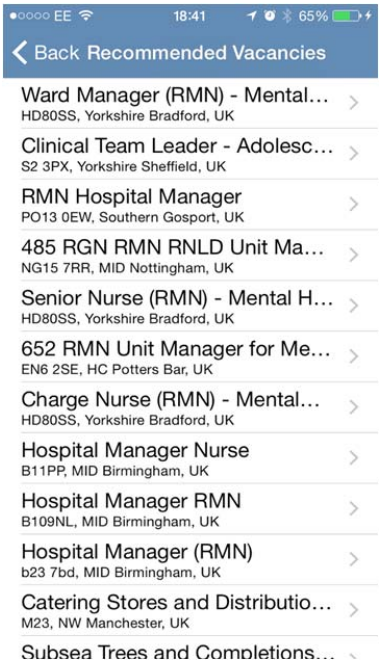
Pinpoint is a mobile app developed by a 17 year old student. It is intended for use prior to meeting with a careers advisor. Developers identified that most people going to see a careers adviser do not know what they want to do. By swiping on small snippets of information (left for 'no' and right for 'yes'), users automatically build a profile, which then suggests jobs. Users can then either apply for jobs in their area, or take that information to an adviser. It has a card-based interface and implements Tinder style swiping of screen to like/dislike information on card. Cards included: career card; pay by region; *Working Futures* data; unemployment data card; and qualifications card.

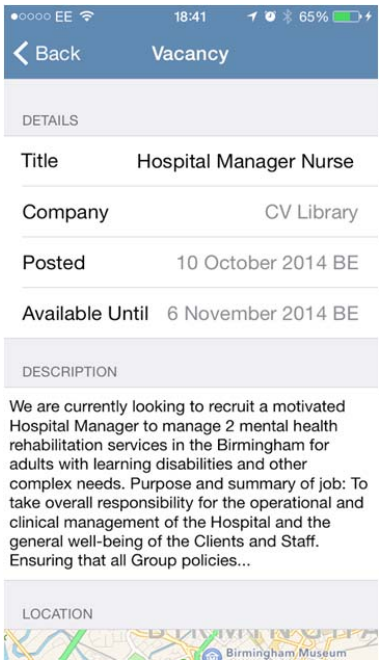
Feedback from the careers stakeholders was very positive with many suggesting that it was a good learning tool. It was also considered a good approach for an education environment targeted at younger users. The rapid response and results was also considered a key feature of the application. The only suggestion was the possible inclusion of an 'unlike button' to ensure users could refine choices as they learnt more.

Screenshots of Pinpoint









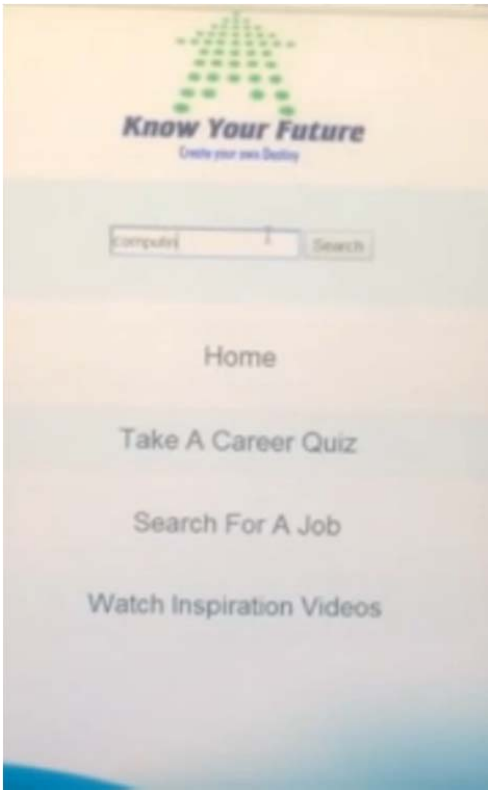
Code available at: https://github.com/datagovuk-hackcode/lmi-mod-pinpoint

Know Your In Demand Future

Know Your In Demand Future utilised the *Working Futures* data in the LMI for All API. Users could enter their qualifications and desired roles, and the website would return the likelihood of getting a job in that field in the future. It then identifies what steps could be taken to improve the chance of getting a job in that field. By selecting a specific job, the job description, qualification requirements, tasks, and estimated pay and hours are presented. The user is also able to search for vacancies by job and postcode. Data are presented visually to help users easily understand trends in the labour market. The overall aim of the app is to provide a range of tools to help individuals with their career pathway.

Feedback from the careers stakeholders was again positive. The website interface was commended for its clear presentation and navigation. It was also noted that the data were also clearly presented and useful. However, it was suggested that data could be presented as percentage increase or decreases. Overall the website was liked as it was considered easy to type in a job and get a lot of information in return, which meant the site was accessible to those with clear career plans and those exploring their options.

Screenshot of Know Your Future



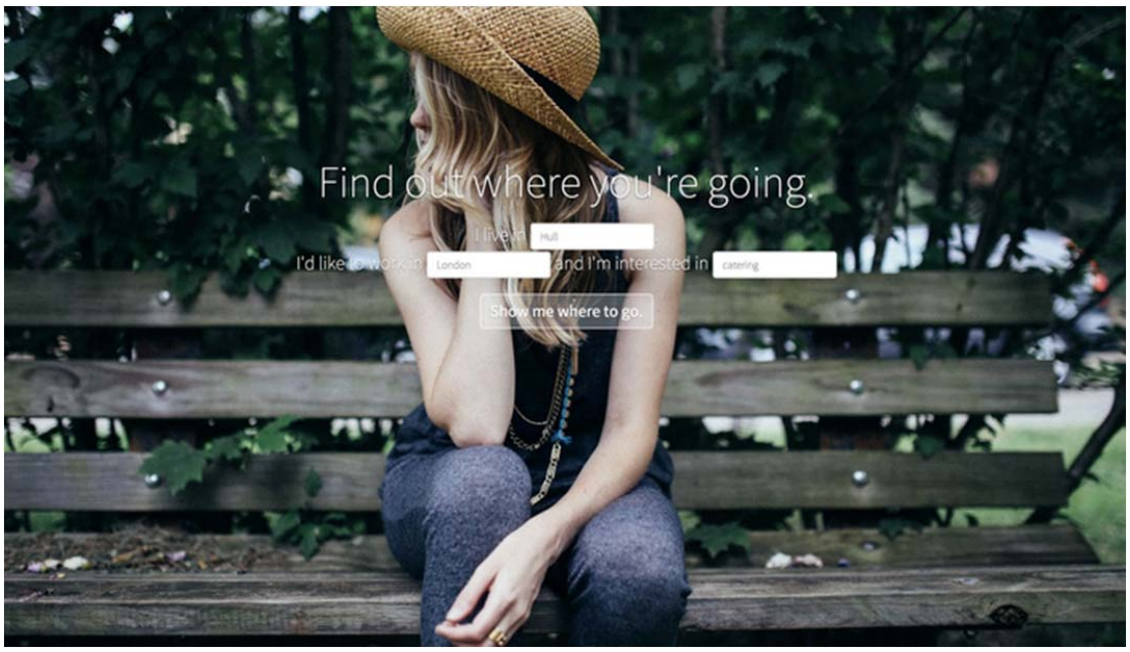
Code available at: https://github.com/datagovuk-hackcode/lmi-mod-know-your-in-demand-future

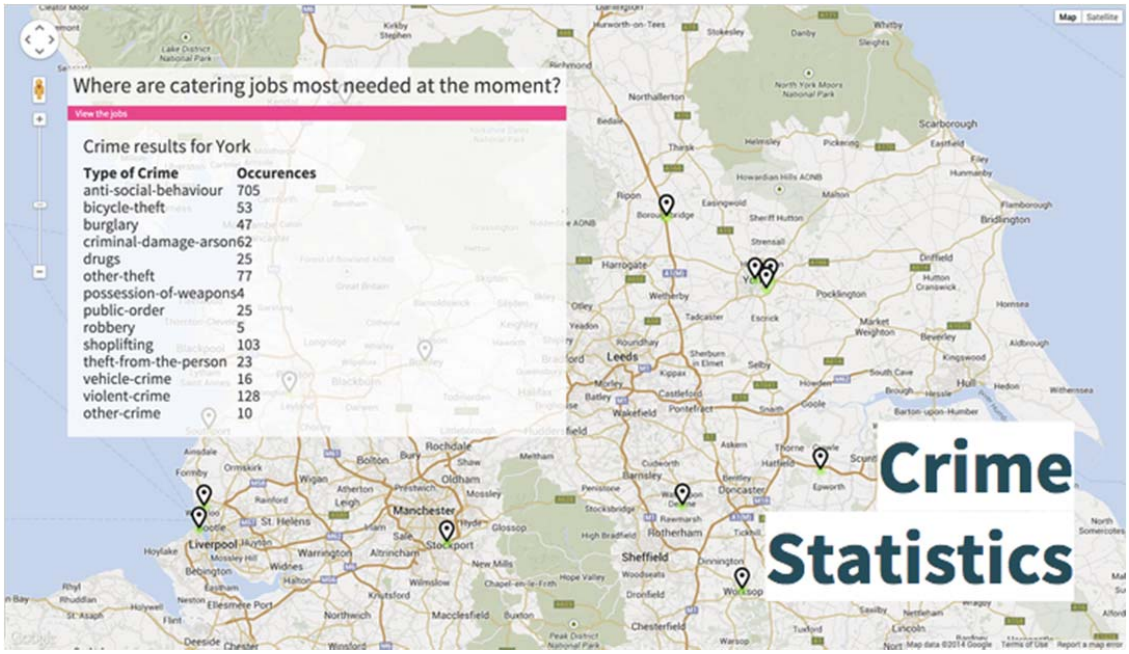
Hot Jobs

Hot Jobs allowed those looking for career guidance to visualise the density of job vacancies for a specific role or industry in the UK. Live data were pulled through the app to provide more context about the potential locations of the searches results. Crime statistics and property price information were also included. In addition, data from other APIs were embedded in the app to include information on property rental prices, cost-of-living and other contextual information.

Feedback on Hot Jobs was very positive and it was considered innovative in its use of data from other sources. The heat maps showing density of vacancies was particularly liked.

Screenshots of Hot Jobs





Code available at: https://github.com/datagovuk-hackcode/lmi-mod-hot-jobs

Job Hub

Job Hub was not a standalone project, but focused on bringing together the other applications and web interfaces in to one consistently designed interface. Instead of having to use three different applications, this let the users use one application, but still gather the relevant information in small, digestible chunks. This app had less development time, so feedback was more about how it could be developed and refined further such as providing more context to the data and ensuring there was a ‘call to action’. The concept was, however, commended.

Code available at: https://github.com/datagovuk-hackcode/lmi-mod-job-hub

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 Treasury and 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

Li , Y and R A Wilson (2015) *Technical Report on Generating Detailed Estimates of Pay and Hours*. Institute for Employment Research University of Warwick: Coventry

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.

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.

Wilson, R. A., and Homenidou, K. (2012a). *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