

LMI FOR ALL

TECHNICAL, STAKEHOLDER ENGAGEMENT AND DATA DEVELOPMENT SERVICES

Final Report

2015-16

Project Director and Principal Investigators

Professor Jenny Bimrose, Dr Nick Sofroniou and Dr Sally-Anne Barnes
Institute for Employment Research, University of Warwick
Graham Attwell, Pontydysgu

Technical services

with Philipp Rustemeier and Dirk Stieglitz, Pontydysgu
Raymond Elferink, Raycom

Stakeholder engagement services

with Dr Deirdre Hughes OBE, Institute for Employment Research, University of
Warwick

Data development services

with Dr Yuxin Li, Dr David Owen, Professor Rob Wilson, Luke Bosworth and Peter
Millar, Institute for Employment Research, University of Warwick



Table of contents

Executive summary	i
1. Introduction	1
1.1. Project overview	1
1.2. Project aims and objectives	2
1.3. Report structure	3
2. Technical services	4
2.1. Introduction.....	4
2.2. Platform and database -maintenance and monitoring	4
2.3. Extract, Transform and Load system (ETL).....	6
2.4. Data security and data disclosure	8
2.5. Wiki for tracking project development	8
2.6. LMI for All web portal	8
2.7. Data cubes	11
2.8. Analytics / API Dashboard and usage data.....	11
2.9. Maintenance of the API and further development	12
2.10. API monitoring	12
2.11. Query Error Monitoring	12
2.11.1. Outage monitoring.....	13
2.12. Technical support and customer care	13
2.13. Blue sky gazing – future directions for development	13
2.13.1. Infrastructure	13
2.13.2. Potential uses of LMI for All.....	14
2.13.3. App directory	14
2.13.4. Widgets	14
2.13.5. Careers Chatbot and Machine Learning	15
2.13.6. Granularity of data.....	15
2.13.7. Job vacancy data	15
2.13.8. Employability	16
2.13.9. Subjects, Disciplines and Occupations	16
2.13.10. Social learning and LMI for All	16
3. Stakeholder engagement services.....	17
3.1. Introduction.....	17
3.2. Processes for designing, developing and implementing an application or widget to supported by LMI for All.....	17
3.3. Case studies.....	21
3.3.1. icould: Building LMI for All into an existing website.....	21
3.3.2. Prospects: Developing an app.....	25
3.3.3. Skills Match London: Building a website using the LMI for All API	28
3.3.4. KareerHub: Building an online LMI dashboard using LMI for All	31
3.4. Users of LMI for All	33
3.4.1. LMI for All web interfaces	33
3.4.2. Careerometer in practice	34
3.5. LMI for All usage statistics	35

3.5.1. Query volume.....	35
3.5.2. Unique users.....	36
3.5.3. Traffic by endpoint.....	37
3.5.4. Traffic by country.....	38
3.6. Engagement activities.....	39
3.6.1. Other engagement activities.....	42
3.7. Recommendations for future stakeholder engagement services.....	43
4. Data development services	45
4.1. Approach to providing data.....	45
4.2. Summary of data and indicators included in the portal.....	47
4.2.1. Core indicators.....	47
4.2.2. Other indicators.....	51
4.3. Data development summary.....	64
5. Future issues and potential resolutions.....	65
5.1. Enhancing the database: potential and additional data sources.....	65
5.1.1. General considerations	65
5.1.2. Vacancy data	66
5.1.3. Course information.....	66
5.1.4. Census of Population data	67
5.1.5. European data – the Cedefop database, EU Skills Panorama, ESCO and EURES 67	
5.2. Stakeholder engagement activities: maintaining momentum.....	69
5.2.1. General considerations	69
5.2.2. What works?	69
5.2.3. Overcoming challenges to implementation.....	69
References	71
Annex A: Users of Careerometer	73
Annex B: Engagement activities.....	75
Annex C: Core data sources included in LMI for All	81
Annex D: O*NET.....	117
Annex E: Other data considered for inclusion but rejected.....	128

List of tables and figures

Figure 2.1 Overview of LMI for All platform and database	5
Table 2.1 Overview of database servers.....	5
Figure 2.2 STAR model illustrated by LFS data	6
Figure 2.3 Screenshot ETL upload process – start	7
Figure 2.4 Screenshot ETL upload process – tracking progress	7
Figure 2.5 Page progression of visitors to LMI for All.....	10
Figure 3.1 Possible development processes for an LMI for All interface	20
Figure 3.2 Map of Careerometer users	35
Figure 3.3 LMI for All query volume, 2015-2016.....	36
Figure 3.4 LMI for All unique users, 2015-2016.....	36
Figure 3.5 LMI for All traffic by endpoint, 2015-2016	37
Figure 3.6 LMI for All traffic by endpoint and month, 2015-2016	38
Figure 3.7 LMI for All traffic by country, 2015-2016	39
Figure 3.8 Map of stakeholder engagement activity, 2015-2016	41
Figure 4.1 Data overview – LMI for All.....	49
Table C.1 Typical Earning Function Results	85
Table C.9.1 Broad Sectors (SIC2007).....	111
Table C.9.2 Industry Groups (SIC2007)	112
Table C.9.3 SOC2010 Major Groups and Sub-major Groups	113
Table C.9.4 Qualifications classified using Regulated Qualifications Framework (RQF)	114
Table D.1 Mapping from SOC 4-digit categories directly to O*NET.....	119
Table D.2 Alternative steps to improving the matching	120
Table D.3 Data layout of ‘Skills.txt’ and ‘Abilities.txt’	123
Table D.4 Abilities.txt.....	124
Table D.5 Skills.txt	125
Figure E.5.1 Labour market questions in 2011 Census of Population	144
Figure E.5.2 Journey-to-work questions in 2011 Census of Population	145
Table E.6.1 Mapping from ISCO08 to SOC2010	147
Table E.6.2 Map from ISCO 88 to SOC2010 at 2-digit level	149

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.
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	<i>Destinations of Leavers from Higher Education</i> 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.
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.
MySQL	MySQL is a type of database management system that enables data to be added, accessed and processed in a database. It is open source. MySQL is supported by Microsoft and Oracle.
NOMIS	Web-based database of labour market statistics from ONS, includes statistical information on the UK labour market (i.e. Employment, Unemployment, Earnings, Labour Force Survey and Jobcentre Plus vacancies).

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 is 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.

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

Executive summary

The purpose of this report is to document the progress of the LMI for All project during its fourth phase, detailing the data processing required to populate LMI for All, the technical development supporting the LMI for All service, identifying current data available and describing the stakeholder engagement processes used to raise the profile of the service offer. The report updates much of the material presented in the progress report for the project during its third phase (available at: http://www.lmiforall.org.uk/wp-content/uploads/2015/07/LMI_for_All_Final_report_v020715-for-website.pdf) and highlights new developments.

The development of LMI for All service has extended over a four year period (2012 – 2016). During this time, the feasibility of developing a comprehensive career LMI data tool that exploits open data sources that can be mainstreamed into service provision has been demonstrated. The overall aims of the project, specified by the UK Commission for Employment and Skills, have been:

- ❖ 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.

Data have been made available through a purpose built web portal and data Application Programming Interface (API). This has been built on four successive iterations (a prototype, phase 1, followed by pilot phases 2A and 2B and then most recently by phase 4), 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. The web portal is freely available for any third party developers wishing to harness its potential in their own particular operational context, who determine the purpose of the applications developed.

Operational work has been 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. The main web portal is available 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. A frequently asked questions section is also available.

So far, the development process for LMI for All has been successful in achieving the following three key goals:

- ❖ A comprehensive data offer;

- ❖ Implementation of robust, secure, fit-for-purpose technical infrastructure; and
- ❖ 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 other 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 exploitation of its potential, together with the likely processes of engagement, are understood and can be implemented within organisational contexts.

1. Introduction

1.1. Project overview

Despite a policy 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. The ultimate aim for 'LMI for All' is to provide a single access point for multiple, high quality sources of LMI, which is openly accessible, 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.

The LMI for All project has been funded and managed by the UK Commission for Employment and Skills. Operational work has been 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. 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. Four sequential, but distinct, research and development phases have ensured a successful outcome. A prototype stage (Phase 1, November 2012 to May 2013¹) 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., 2013², for a detailed account). The third stage, Phase 2B, which was largely based on recommendations from Phase 2A, detailed progress on implementing these recommendations and was published in 2015³.

The report that follows describes the activities undertaken in the fourth stage of the project.

¹ 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. [Summary available to download at: http://www2.warwick.ac.uk/fac/soc/ier/research/prototypecareerslmi/bimrose_et_al_2012_lmi4all_summary.pdf

² Bimrose, J., Wilson, R., Barnes, S-A., Owen, D., Green, A., Li, Y., Millar, P., Bosworth, L., Holden, A., Attwell, G., Rustemeier, P., Elferink, R., McGreggor, A., Wicks, D., Long, A. and Dickerson, A. (2013). 'LMI for All' Developing a Careers LMI Database Phase 2A report. London: UKCES. Available online: http://www2.warwick.ac.uk/fac/soc/ier/publications/2012/bimrose_et_al_2012_lmi4all_summary.pdf

³ Bimrose, J., Wilson, R., Barnes, S-A., Owen, D., Li, Y., Green, A., Bosworth, L., Millar, P., Holden, A., Attwell, G., Rustemeier, P., Elferink, R. and Higginbottom, J. (2015). LMI for All: Developing a Careers Database (Final report) . London: UK Commission for Employment and Skills. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/446332/LMI_for_All_Final_report_v020715_FW.pdf

1.2. Project aims and objectives

The overall aims of the LMI for All project have been 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.

As with phases 1 to 3, during phase 4 of the project, these were operationalised in three separate, but inter-related work strands, specified by the UK Commission for Employment and Skills. These are identified below, together with their related objectives, all of which have been fully met:

Technical services:

- ❖ The maintenance of a secure and robust infrastructure for the data, including test and production servers and provision for back up services.
- ❖ To implement processes to ensure that the data made available to developers and end-users meets required quality standards.
- ❖ To provide modern and flexible software tools to allow the querying of the database by external users.
- ❖ To provide software tools and spaces for documenting the process and allowing public access to the LMI for All database.

Stakeholder engagement services:

- ❖ To motivate, encourage and support the integration of the data tool into practice, with a range of stakeholders.
- ❖ To record, and make available on the website, lessons learned from these stakeholders relating to the process of integration into practice.
- ❖ To increase the awareness of LMI for All across the broad community of careers practice by a range of dissemination activities.

Data development services:

- ❖ To update existing datasets.
- ❖ To continue identification of other datasets that are useful for 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.

1.3. Report structure

This final project report focuses on the Phase 3 activity. As with previous progress reports for LMI for All, it deals with the three different work strands separately: technical services (section 2); stakeholder engagement services (section 3); and data development services (section 4). Future issues and potential resolutions are presented in section 5, to secure LMI for All going forward. Its overall purpose is to document progress of LMI for All during phase 4, 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.

2. Technical services

2.1. Introduction

The overall aim of the technical development work on LMI for All has been the provision of services to develop and maintain a robust technical infrastructure.

There are four key objectives to providing and developing the technical infrastructure:

- ❖ The maintenance of a secure and robust infrastructure for the data, including test and production servers and provision for back up services. The hardware and software environment should be capable of being extended in relation to increased data provision and added numbers of users.
- ❖ To implement processes to ensure that the data made available to developers and end-users meets required quality standards and avoid disclosure. This involves programming at the API level to suppress data which are unreliable due to sample sizes and to provide warnings to developers. It also requires processes for checking data against expected returns and providing guidance to developers in the use of the API.
- ❖ To provide modern and flexible software tools to allow the querying of the database by external users. This entails the extension of the previous data tools to facilitate querying through data cubes.
- ❖ To provide software tools and spaces for documenting the process and allowing public access to the LMI for All database. In particular, this involves supporting developers in producing applications which utilise the API through resources made available on the LMI for All website.

This work has involved contributing to internal data management processes, including regular liaison with the data development team around documentation and Quality Assurance of data.

The following reports on the work undertaken in order to fulfil these objectives.

2.2. Platform and database -maintenance and monitoring

The management, monitoring and maintenance of the LMI for All database is designed to ensure maximum performance and ease of access and querying. Similarly the management of the cloud-based hosting environments (test and production), which includes server maintenance, security updates and daily backups is designed to ensure uninterrupted service.

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 2.1 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 2.1 Overview of database servers

<p>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)</p> <p>SQL server authenticated: lmi4all_api [pwd: lmi4all]</p> <p>Windows authenticated:</p> <p>PRODATA\Administrator NT SERVICE\SQLWriter NT SERVICE\Winmgmt NT Service\MSSQLSERVER NT AUTHORITY\SYSTEM NT SERVICE\SQLSERVERAGENT NT SERVICE\ReportServer</p>	<p>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)</p> <p>SQL server authenticated: lmi4all_api [pwd: lmi4all] data [pwd: eong)ah6U]</p> <p>Windows authenticated:</p> <p>PRODATA\Administrator NT SERVICE\SQLWriter NT SERVICE\Winmgmt NT Service\MSSQLSERVER NT AUTHORITY\SYSTEM NT SERVICE\SQLSERVERAGENT NT SERVICE\ReportServer</p>
---	--

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.

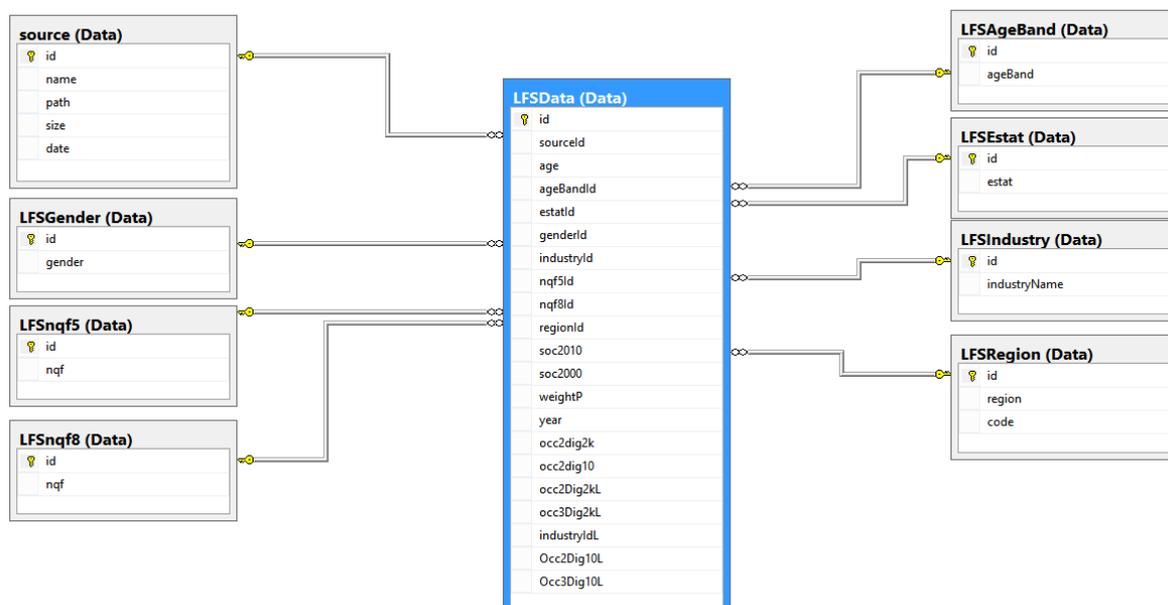
In each of the databases, the same data model that consists of two databases is hosted:

- ❖ 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 the LMI for All 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 2.2 STAR model illustrated by LFS data



2.3. Extract, Transform and Load system (ETL)

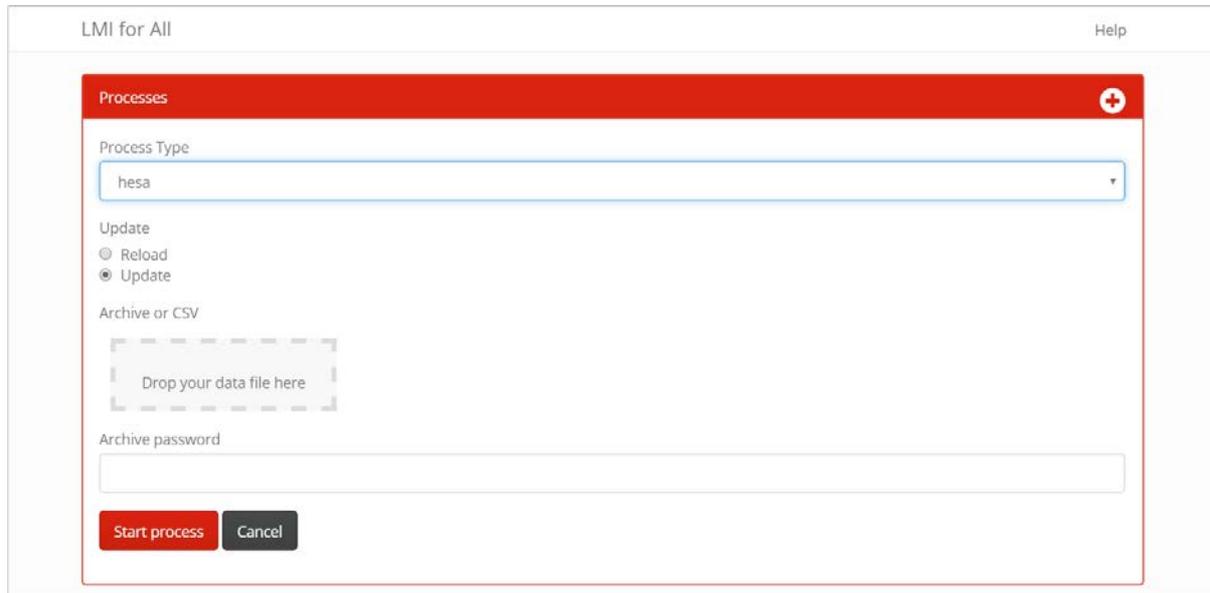
The development and implementation of ETL processes are designed to support the refreshment of existing datasets and integration of new datasets and to maintain the quality of the data.

SQL Server Integration Services (SSIS) packages to automatise 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.

This year a data administration dashboard has been created in beta that allows Data Administrators to upload raw data packages to start the automated ETL processes. The web-interface gives a live status report of the ETL process and prompts for user input when needed.

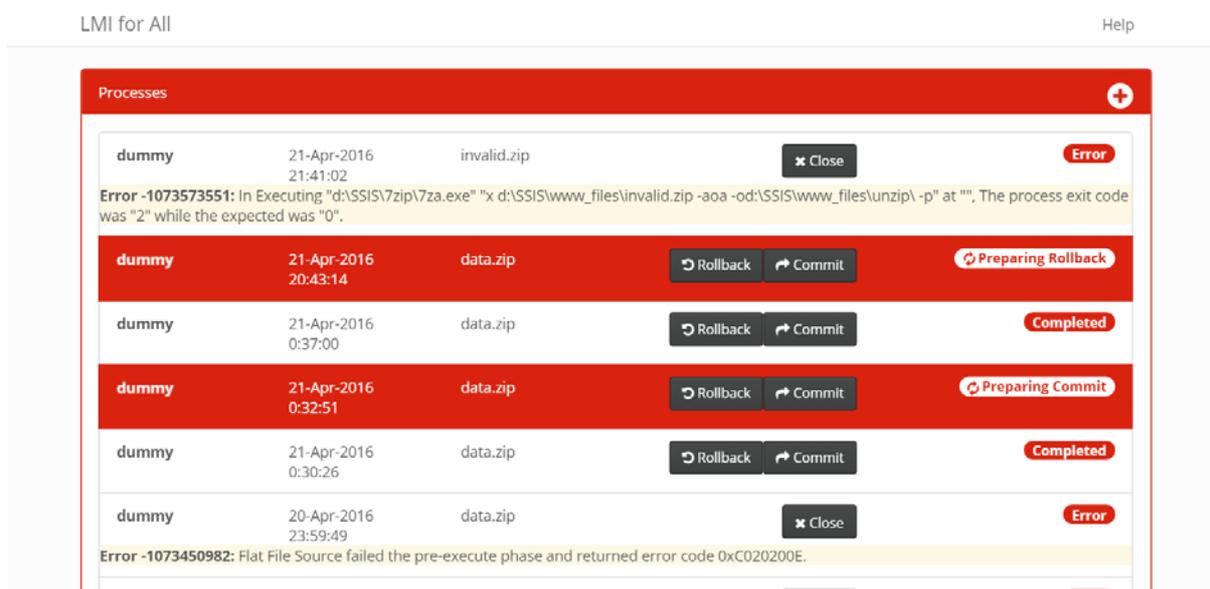
The first step is to upload the file and start the corresponding ETL process.

Figure 2.3 Screenshot ETL upload process – start



Then after the file has been uploaded to the server the process starts and the data administrator can follow the progress.

Figure 2.4 Screenshot ETL upload process – tracking progress



If any error occurs during the ETL the process will throw an error and the data will not be loaded. Errors can occur when the uploaded package is corrupted or when the raw data is not consistent with the preset rules.

When no errors occur the data is loaded into DEVDATA and the data administrator can test the data through the DEV LMIforAll interface. Depending on the results of this testing she can either choose to Commit or to Rollback the transaction. When a transaction is

committed a new option is presented that allows for the new data to be propagated to the PRODATA database, making it available for general use.

The automation of ETL processes needs to be done carefully so that we can catch any eventuality and avoid corruption of the live data. The current ETLs are now being prepared for full automation and we are using real data files to fine-tune the data administration dashboard and the ETLs themselves in parallel.

2.4. 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. This architecture ensures that the underlying data that can be accessed from the LMI for All database are not potentially disclosive or confidential.

2.5. Wiki for tracking project development

A new wiki has been developed, using hackpad software for both internal communication and internal and external data documentation. The wiki currently holds five collections: case studies, data timetable, ETL processes, issues and problems and stakeholder engagement.

2.6. LMI for All web portal

One of the activities under the current phase of work was the development, monitoring and maintenance of the LMI for All website, with the aim of facilitating better developer engagement. The website also uses Google Analytics allowing better analysis and understanding of its use.

The LMI for All website continues to be run on the Wordpress Content Management system, but has been extensively redesigned with a new theme and the enhanced use of graphics. It also allows access to external content based on social media including Twitter and YouTube.

Google Analytics provides a wide array of statistics and a full report is available on request. Here we highlight a number of figures relating to the web site access between May 1, 2015 and 30 March, 2016.

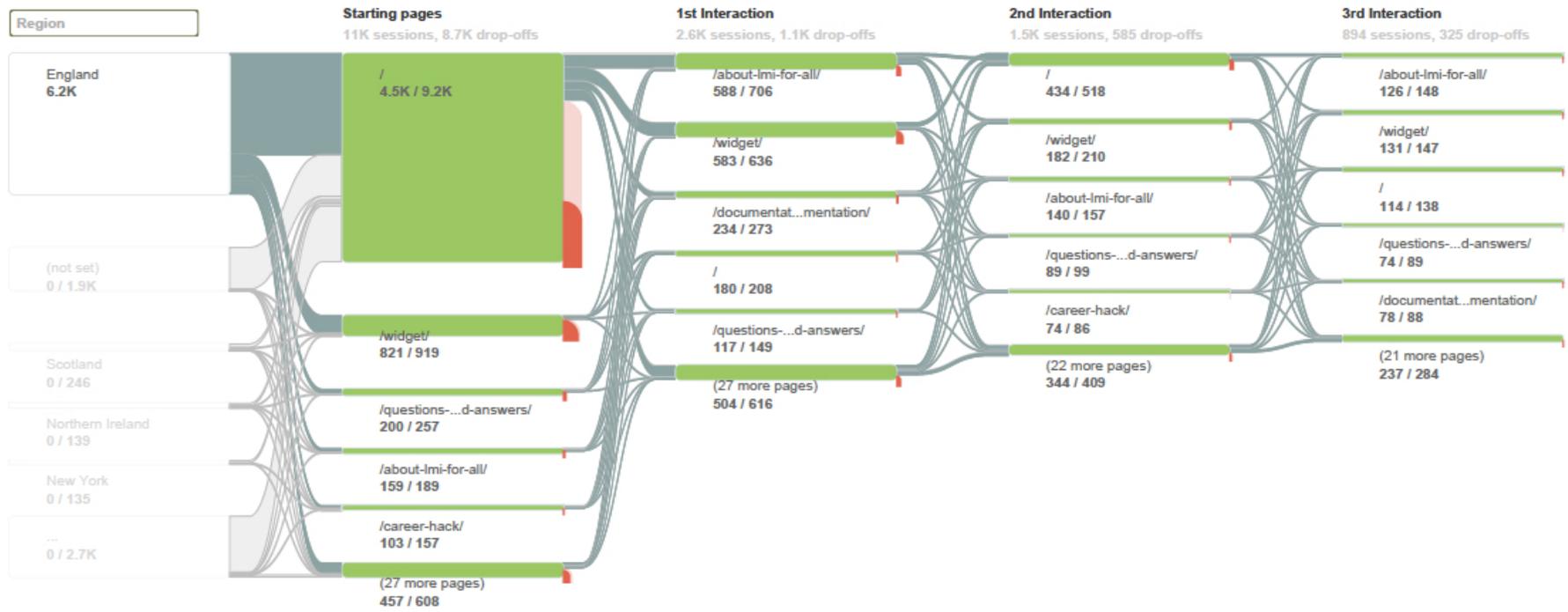
There were 11,411 sessions, of these 8,735 or 76.55 per cent were new users. The average pages per session were 1.76 and the average session duration 1.44 minutes with a bounce rate of 69.2 per cent. Bounce rate is an Internet marketing term used in web traffic analysis. It represents the percentage of visitors who enter the site and then leave rather than continuing on to view other pages within the same site. Whilst it can be difficult to interpret such figures it suggests a broad impact in terms of interest in LMI for All. In terms of monthly visits, with the exception of holiday periods session usage was remarkably consistent

throughout the year. Including the 8 per cent of users whose country of browser use is unknown, almost 60 per cent of access was from the UK and 13 per cent from the USA.

Of particular interest is the page progression of visitors to the site, illustrated in figure below. The diagram below shows entry routes by visitors to the site and their subsequent progression to other pages.

It is interesting to note that those visitors who stayed beyond just finding out about the project tended to go towards the more technical areas of the site including the data documentation and the widget. The report and videos of the hack days were also popular.

Figure 2.5 Page progression of visitors to LMI for All



2.7. Data cubes

A cube is a set of data that is usually constructed from a subset of a data warehouse and is organised and summarised into a multidimensional structure defined by a set of dimensions and measures. Currently, there are two cubes; 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.

ASHE Pay

This cube contains weekly pay estimate data for one year (the most recent that could be obtained). 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. 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 2013, 2014 and 2015. 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

2.8. Analytics / API Dashboard and usage data

Software infrastructure for real-time monitoring of the API has been developed, together with the new API key management system. The monitoring services collect data in a common format, which can be viewed and interpreted using a real-time dashboard (where real-time means the displayed data is accurate up to the most recent five minutes).

The dashboard has been used to generate analysis against key indicators and has proved extremely valuable in understanding usage. However, the level of detail of the collected data has started to reach the upper limits that the hardware resources of the API server can support. For proper and effective monitoring, server upgrades are needed. As of the time of this report, the real-time dashboard has been deactivated because generating the live reports takes an unacceptably long time and is starting to cut into resources originally reserved for the system's main task— API query processing.

In the meantime, offline analytics are still a viable alternative. Upgrades to the processing scripts have extended the period over which offline analytics may be performed back to over a year (at current volume of queries), at the loss of some details like city-level geomapping of requests. In light of the increasing number of applications relying on the API, **the technical team strongly suggests adding two more servers for the API; one for real-time analytics and one for failover and extra redundancy.**

2.9. Maintenance of the API and further development

Development and maintenance of the API has continued, shifting more towards maintenance as the functionality of the API is fairly stable by now. Notable developments have been:

- ❖ Integration of the Burning Glass online job posting information service;
- ❖ Development of the ASHE Median/Deciles query (still on-going);
- ❖ Overhaul of the foundational software and deployment environment (newest Java, Scala and Jetty server);
- ❖ Stability and speed improvements to backend connections (most notably requests to Universal JobMatch);
- ❖ Development of customised real-time monitoring services integrating into the API;
- ❖ Development of special access key system for certain endpoints only (Burning Glass).

2.10. API monitoring

API monitoring remains in place to generally very good results. Real-time API analytics have increased the depth and detail of monitoring errors (when they are available). Nevertheless, the previous API monitoring system also remains in place. Both email errors and severe irregularities during processing to the technical team as they happen.

2.11. Query Error Monitoring

API queries can produce a variety of results, which are standardised and classified by well-known HTTP status codes. 2xx Codes (200, 201, ...) indicate that generally the request was successful. 4xx codes indicate that there was some kind of user error, such as an incomplete request or missing authorisation. 5xx codes indicate a server-side error, i.e. something in the actual API code went wrong. The API monitoring tracks the status codes of all API results, and immediately emails 5xx-type codes to the development team. 2xx and 4xx codes are not alerted, but are stored in the analytics database for later review.

There have been no “serious” 5xx codes during the period covered by this report. Backend failures, for example UJM being down for maintenance, occasionally generate code 504 (which roughly corresponds to “service not reachable”) and get reported even though the problem is really with UJM and not with the API itself.

2.11.1. Outage monitoring

The API and the API Explorer websites are regularly monitored (at five-minute intervals) for outages by an outside service. If the API is found to be unresponsive, an instant alert is sent to the lead developer's mobile phone.

There was one major unexpected outage in the period covered by this report, leading to intermittent API service for roughly half a day. The underlying problem was file descriptor exhaustion on the API server due to stalling connections to Universal Jobmatch. The problem was identified and fixed, restoring service without further incident. All other API service interruptions were planned upgrades taking no longer than five minutes each.

2.12. Technical support and customer care

Support and customer care have been provided as usual during the period covered by this report. Several rounds of load testing the API were originally envisioned, however it was agreed between the tech team and project management at UKCES that the API is quite robust under load, and further tests would have been unnecessary during the reporting period. As such the resources for load testing have been shifted to the provision of additional technical support, especially in conjunction with integrating the Burning Glass and Av Live apprenticeship vacancy data sources.

2.13. Blue sky gazing – future directions for development

This section of the report reviews the technical development and the emerging patterns of use of the LMI for All database and considers future possible directions for development. Of course, it is acknowledged that such developments will have resource implications and that there may need to be some choices over priorities given resource limitations. It may also be that it is possible to seek external funding or some of these developments or to seek potential partnerships to take this work forward.

2.13.1. Infrastructure

The present infrastructure has emerged over time, inline with the availability of different datasets, and their associated calculations and preparation methods. Now that the range of available datasets is stabilising, the API may benefit from a gradual overhaul and the establishment of an overall data management framework. This would also go well in hand with the ETL capabilities the technical team is developing. While this sort of future-proofing development is very desirable, a more critical area is the hardware infrastructure. Increased requirements for API monitoring and access management are forcing the API systems to operate at high capacities and low redundancies. This can be expected to lead to service degradation or outages as the number of users increases.

The technical team strongly recommends investing in additional API resources in the form of servers and storage for API access management. Adding more redundancy via load balancing and fall-back servers for the actual API is also highly desirable to ensure smooth future operation and minimise downtime, e.g. during service upgrades. There are other

potential applications that require additional server capacity, such as generation and storage of long-term time series from third-party supplied data.

The technical team also recommends a cost/benefits of current hosting providers / solutions be performed. The API has been running on the same infrastructure for at least three years, while cloud hosting costs, relatively, have likely gone down. The current field of available hosting solutions should be evaluated with respect to the increased resource demands created by deep API access management, data storage requirements, as well as possible political requirements (e.g., could the API be hosted outside the UK proper, such as in a – likely cheaper – data centre in Ireland).

2.13.2. Potential uses of LMI for All

The LMI for All database and API has been designed for use in careers guidance, either directly by end users, mediated through applications, or by careers organisations and agencies working with clients. However, feedback from conversations at various events has suggested further potential uses. These include for economic and social planning, for planning courses and education and training provision and for use in labour market and social research. Of course, there is nothing to stop organisations developing applications now for these purposes and indeed the Department of Works and Pensions is collaborating with the European EmployID project in developing and implementing a Dashboard for use by Work Coaches in Job Centres as well as by Employment Advisers working with companies. However, in order to promote such uses, then this has implications for dissemination and promotion and for the further development of the LMI for All website and support materials. It might also have implications for future data provision, for instance through including travel-to-work data and other economic and social data sets.

Further thought could also be placed on organising hack days or competitions encouraging and promoting different uses of LMI for All for these different potential user groups.

2.13.3. App directory

Although it is understood that many applications being developed by different organisations and the studies undertaken through the dissemination of the project are producing reports on some of these, there is limited transparency around many of the applications. Some organisations are developing commercial uses of the data and may not wish to share details of this work. However, others may be prepared to share details. The development of an App directory would both allow organisations to promote their products, but at the same time provide ideas for others as to how they can best use LMI for All.

2.13.4. Widgets

Although the strategy of developing and providing access to an open API has proved successful in attracting a range of organisations to develop applications and the project has been able to offer help to developers when required, our research suggests that application development is a challenge for small organisations lacking in resources and competence. This is particularly so for many organisations offering careers in England.

There has been considerable take up of the Careerometer widget applications, providing access to data around pay (see section 3.4.2). The present application requires redevelopment to overcome a number of limitations and bugs. Consideration should be given to the development of a more advanced modular widget allowing the selection of different data to be displayed and the graphic rendering of such data.

2.13.5. Careers Chatbot and Machine Learning

There is presently considerable interest, investment and development in the computer industry in chatbots. A chatbot is a computer program, which conducts conversation either through audio or text. With rapid advances in Artificial Intelligence and machine learning, chatbots are seen as especially useful for customer services and for information acquisition. Such advances have led to the release of software development kits, making the development of new chatbots far less resource intensive.

Young people, in particular, are heavy users of text chat applications and the LMI for All database is well suited to query by text. The development of a Careers Chatbot based on LMI for All could have a major impact in allowing young people access to quality LMI for use in planning careers.

2.13.6. Granularity of data

One of the limitations of LMI for All which has been raised by a number of users is the granularity of data, particularly with respect to geographical locations. At present most of the data is available at the finest level to regions. This is simply due to sample sizes. Although in some applications, for instance the DWP dashboard, finer grained data is available through a mashup with ONS local data, this data is classified under Standard Industrial Classification (SIC), which is of limited value in careers and job planning.

Consideration could be given to developing and testing systems for 'crowdsourcing' more local data, for instance by encouraging DWP Employer Liaison staff to annotate and add local intelligence about future job opportunities in their locality to be mashed with the LMI for All 'official' data. Such a system could provide more engagement with LMI for All and enhance the data, although of course has quality implications which would need research.

2.13.7. Job vacancy data

One of the most used datasets is job vacancy data, which is taken from the Monster API. These data are, however, somewhat messy and limited. For instance, as jobs are registered to the head offices of companies, there is no ability to specify geographical range of data. Furthermore, there is no requirement to register vacancies with Monster and research suggests that the data over reflects low paid, low skilled jobs and public sector vacancies. At present we have no long-term job vacancy data.

Consideration should be given to either licensing data from external companies, such as Burning Glass, or to experimenting with scraping data from local websites.

2.13.8. Employability

Within education and training, there is an increasing focus on employability. This requires referencing to both the formal and technical skills and social or soft skills required in an occupation as well as an understanding of the transferability of competences and skills between different occupations. Obviously, in part because of rapid change in technology but also work organisation, these competences and skills change over time. LMI for All provides access to very detailed data on competences and skills through the O*NET data set. O*NET has had only limited take up in present applications which may well be due to lack of understanding of the potential uses of this data. More documented use cases and possible events (see below) around this data, together with consultations with those concerned with employability could help uptake and the development of new applications.

2.13.9. Subjects, Disciplines and Occupations

One issue which has arisen frequently in presentations of LMI for All is qualification routes and pathways between academic courses – both at school and Higher Education levels to different occupations. At present the only indications provided by LMI for All is through tracking of occupations of graduate students after finishing their degree. The main difficulty is in coding systems with degree subjects and disciplines coded under JACS3 and occupations under SOC2010. There may be a number of ways forward in this regard. Neil Bachelor from Omnifolio has already developed mappings from US SOC to JACS3, which he is prepared to share. Some of the occupational bodies already map possible courses to occupations and these could be computer scraped or manually added to a database. There may be value in further discussions with the Skills Funding Agency over access to their data. The XCRI standard for course classifications is useful but has only limited implementation in Further Education colleges. However, Alan Paull, who developed the standard whilst working with JISC, has also worked on technologies for mapping courses to occupations. This area should be reconsidered as a priority for further development.

2.13.10. Social learning and LMI for All

Even with very good visualisations and detailed data, turning labour market information into labour market intelligence requires skills and competences in interpreting and sense making. Although this can be done through face-to-face workshops, careers organisations are presently short of resources and may lack the funding, time and skills to organise such events.

The DWP has experimented with Online Social Learning Courses, which incorporate the use of their Dashboard (based on LMI for All) as part of the courses. The courses and the Dashboard received an overwhelmingly positive evaluation. LMI for All could consider developing its own online social learning programme. This might include Massive Open Online Courses, online seminars or online learning materials for asynchronous use. It is anticipated that such courses would promote the take up of LMI for All plus increase the added value of use of the data.

3. Stakeholder engagement services

3.1. Introduction

The overall aim for LMI for All during the first three phases of the project (2012-2015) was to create a data tool that developers would use to create products to support individuals in making better decisions about learning and work. It was clear that work should therefore not focus solely on the development of the data tool, but also on raising demand and publicising the completed data tool. The development of strong partnership arrangements across a range of different categories of stakeholders and partners was regarded, therefore, as essential to the ultimate success of the LMI for All service. This has been the approach pursued during the first three phases of the project and continued during the fourth phase (2015-2016). It has been particularly important during the latest phase of the project to maintain the momentum of stakeholder dissemination and engagement, as well as increase active use of the service.

The particular aim of the stakeholder engagement services has been to raise awareness and understanding of the LMI for All resource among key audiences. This has been achieved through a range of activities, specifically:

- ❖ Motivating, engaging and supporting third parties to develop applications or enhance existing applications in order to increase the reach of LMI for All among end-users;
- ❖ Motivating and supporting third-parties to implement the LMI for All “widget”;
- ❖ Driving end-user traffic to existing applications that offer access to LMI for All data;
- ❖ Foster positive and supportive attitudes towards the LMI for All service among key groups, including policymakers;
- ❖ Developing the LMI for All website content with four case studies, illustrating the development of LMI for All interfaces;
- ❖ Disseminating the LMI for All service, current users and widget through various dissemination activities (e.g. conference presentations, training workshops, use of social media and publications).

Together, these activities have provided a clear and consistent focus on end-user need and represent a coherent engagement strategy with third parties most likely to be the primary data users. All activities have been UK-wide, including organisational representatives from England and the devolved nations. The LMI for All service has also been promoted internationally.

3.2. Processes for designing, developing and implementing an application or widget to supported by LMI for All

The case studies presented in section 3.3 detail the processes that four organisations went through to design, develop and implement an LMI for All interface to enhance their service provision and delivery. Throughout this phase of the project, the LMI for All team worked with a total of 17 organisations and consortia that were interested in developing an interface, as

well as four who were in the process of developing and redeveloping their interface. Issues and challenges reported by these organisations were systematically recorded.

Changing and evolving context – Potential users of the service noted organisational changes and the evolving context of careers in terms of funding challenges to overcome so that they could engage with the LMI for All resource. The operational environments impacted in different ways. These ranged from whether developing an interface or application was viable in the current context at all, to determining whether applications could be used by clients/customers on their own, or within a careers guidance support session. Linking the development of an interface to organisational strategic objectives emerged as helpful in initiating the process of implementation.

Generating buy-in and the need for a lead/steering group – The majority of potential users recognised the value and potential of LMI for All for providing high quality LMI and enhance their service provision, so gaining ‘buy-in’ from management was not problematic. Some suggested that LMI for All would add credibility to their current LMI provision. It was clear that organisations needed an individual, or a steering group, to lead on the development. Examples of existing interfaces were valued for demonstrating the potential of LMI for All.

Resourcing an interface – Resourcing, both in terms of understanding how much an interface would cost and how to fund it were consistently identified as issues. The costs of developing an interface were found to vary by organisation. It was very much dependent on the availability of in-house technical skills, requirements and complexity of the interface and whether external support was needed both in terms of technical skills and a platform to locate the interface.

Understanding the development process and requirements – For many, a significant challenge was achieving an understanding of the design and development process. Whilst recognising the need for a plan, budget and lead, many were unsure where to start – finding and defining a budget, the type of interface to be developed, target audience(s), etc. Guiding questions were found to be helpful, such as: what questions do your advisers and/or clients/customers ask most frequently?; who would your target audience be?; and how do clients/customers access your service (face-to-face, on the web, or through a mobile device)?

Technical knowledge and skills – For careers organisations, sourcing the required technical skills and know-how was identified as an issue. For some, this related to developing new skills. Others tested out existing tools and platforms to develop an interface (such as Appy Pie (<http://www.appypie.com/>) and a jQuery plugin called Chartinator⁴) and yet others outsourced the development.

⁴ For more information on Chartinator, see <http://js-tutorial.com/chartinator-transforming-html-tables-into-charts-1482>.

Identifying a target audience and needs – The majority of potential users had a number of ideas about target audiences and how LMI needs could be met by the development of an LMI for All interface. Refining these ideas into a practicable product was sometimes a challenge. Some organisations refined their ideas by working with a steering group or surveying staff to identify the greatest LMI need, whilst others made decisions based on funding streams and organisational strategic objectives.

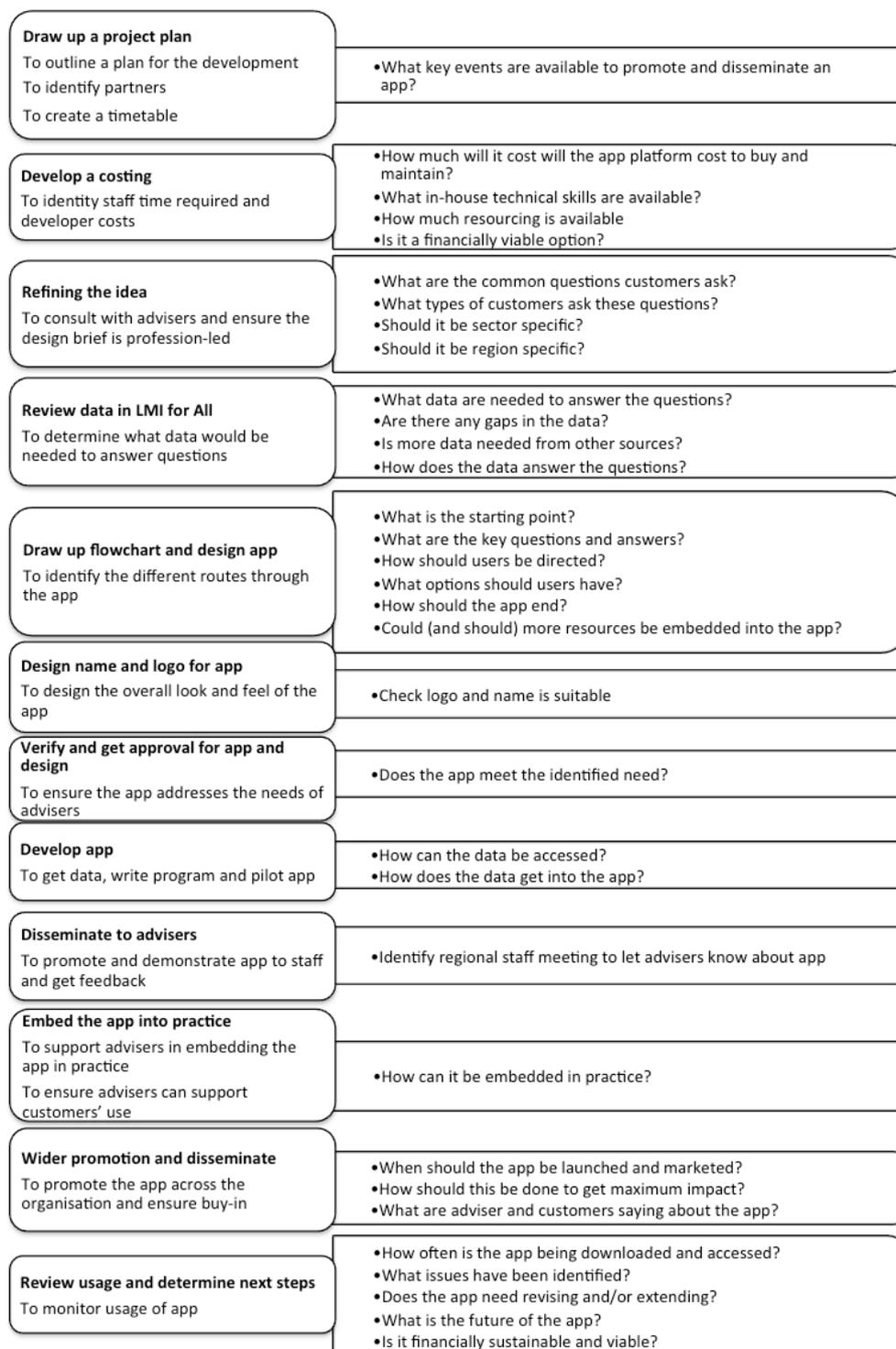
Data queries and concerns – Potential users of the service had a number of questions regarding the data. Whilst all recognised that LMI for All contained high quality data, some concerns were also raised about how to fill missing data gaps, such as course and vacancy data, and how to combine (or ‘mash’) with other data sources. Some questions were about the currency of the data and how often it would be updated. The long-term sustainability and availability of the resource was a recurrent concern.

Dissemination and marketing – This was considered to be a key phase of the development process, but was mainly recognised by those who had experience of developing a product in the past. Some identified future events in which a product could be launched to ensure the most impact. Buy-in from staff to both use and promote the product was considered to be vital.

Time – For many, the length of time anticipated for the development process has been underestimated. Early estimations of the time required for a development were optimistic and then quickly revised, as different stages of the process took longer than expected. For example, identifying a developer and getting a costing can take several weeks as the developer briefing is refined.

Figure 3.1 Possible development processes for an LMI for All interface

The following process plan was drawn up, with the help of Prospects, to illustrate a possible development trajectory for an app. At each stage, some of the questions and issues that need to be considered are indicated.



This is also available on the LMI for All website at: http://www.lmiforall.org.uk/case-study_intro/processes.

3.3. Case studies

As part of this phase of the project, case studies were developed that provide an overview of the context-specific processes involved in designing, developing and implementing an interface using the LMI for All service. The aim of the case studies has been to demonstrate the potential of LMI for All in practice and encourage third party interest in developing their own application or widget.

To create the case studies, a number of organisations and consortia were identified in consultation with the UK Commission and contacted. These organisations were categorised into:

- ❖ Higher and further education institutions (n=5);
- ❖ Career guidance organisations or community of practice (n=5);
- ❖ Schools and young people (n=5);
- ❖ Developer community (n=4); and
- ❖ Other organisations (n=2).

In total, the project team were in direct communication with 17 organisations and consortia interested in developing an interface using the LMI for All service and with four who were in the process of developing and redeveloping their interface. The project team provided support for those wishing to engage with the service. Support ranged from help with understanding the process and what would be required in terms of funding, designing, developing and implementing LMI for All, through to providing support with understanding and integrating the data. In some instances, guidance on the process of user-led design and software development was offered. As part of the development, the project team recorded the whole process. In four cases, the development process was documented as a case study, which were agreed in consultation with participating organisations and disseminated through the LMI for All website. Case studies developed and included on the LMI for All website, include: icould; Prospects; London Skills Match; and Northern Ireland's KareerHub.

The following case studies detail the decisions and processes that organisations have engaged with to use LMI for All. Each tell different stories of the design, development and implementation of the LMI for All service. They are also available on the LMI for All website, which also includes images of the developments (see http://www.lmiforall.org.uk/case-study_intro).

3.3.1. icould: Building LMI for All into an existing website

About icould

'icould is based on the idea of career exploration, encouraging young people to discover career possibilities, including options they may not have considered or didn't know about, rather than trying to match them to a particular job type. We recognise the many influences that can affect career choices and see careers very much as a continuous process as opposed to a one-off decision.'

icould (<http://icould.com>) uses real voices to help young people learn more about the world of work. Using video interviews, the free-to-use website offers unique insight into the day-to-day roles and career pathways of people working in a range of jobs.

icould has over 140,000 unique users per month, with users spending an average of seven minutes per visit. It is used by a wide range of individuals, mainly in the UK, in settings such as classrooms, job centres, careers services and in educational resources such as the Frog learning platform used in schools. Selected users include Careers Wales, Skills Development Scotland, UCAS, CASCAiD and the Discovery Education Platform used in schools.

icould is a small charity funded by a number of sources, including philanthropy.

What is on the icould website?

icould includes over 1,000 careers video stories that provide a real view of what it is like to work in particular jobs. Articles on a range of topics are available from people sharing their experiences, together with advice from industry experts. For example, information for young people and parents on key educational decision points and career-related topics (such as making job applications, CV writing, interviews and networking). There are also articles on a range of sectors, which describe potential future opportunities. A popular feature of icould is the Buzz Quiz, which is based on Myers Briggs. By answering a series of multiple choice questions, users find out their personality type linked to animal and an idea of the careers they may be best suited to. This is used within schools as part of careers education lessons, as well as people on an individual basis.

How does icould use LMI for All?

icould was an early adopter of LMI for All data, using it alongside its career videos, to visualise useful job information using charts and icons.

For each video, occupational information is provided, including average salary, average weekly hours, past unemployment, predicted employment, top 10 industries for the job, employment status of workforce, occupational description, qualifications required for job, tasks, employment by region, gender profile of the workforce and the skills important for the job.

The LMI for All data are also being used as part of the regional and national webpages to show the most popular jobs and the jobs with the highest number of vacancies.

Providing labour market information in a user friendly way significantly adds to the careers story featured in each video and is a unique selling point of icould.

Why LMI for All?

'By combining inspirational, real-life stories from the videos with detailed robust job information, a user could get a lot to go on in terms of informing career choices. LMI for All

was a 'natural fit' for us having already developed video content – it helped to strengthen our offer and connected well with our intention to present a balanced view of careers.'

The LMI for All data are considered a high quality source, as it comes from official national and regional statistics. Using these data addressed feedback from users that could be improved by including hard evidence to support career choices.

'Regional, occupational and vacancy data were particularly in high demand. LMI for All was considered the only good source of reliable LMI that integrated a number of sources into the database, which would address our needs.'

LMI for All, as a freely accessible source, has enabled icould to make these data available at minimal cost covered by the existing staff budget.

How was the LMI for All data used to develop the web interface?

The main aim was to combine real-life stories with detailed job information. The starting point was to look at the different data sets available through LMI for All and select those that would be most useful and relevant. The team thought through common questions young people ask, such as:

- ❖ How much will I get paid?
- ❖ What qualifications do I need?
- ❖ What exactly does the job involve?

The icould team were also interested in the more strategic information, such as working patterns, employment levels and the *Working Futures* data. Within 3 months of accessing LMI for All data, the icould web interface went live. A number of data sets in LMI for All were considered for use alongside the video stories on the icould website. These data included:

- ❖ Employment (historical, projected);
- ❖ Pay;
- ❖ Hours;
- ❖ Unemployment rates;
- ❖ Skills and qualifications;
- ❖ Abilities and interests;
- ❖ Occupational descriptions.

The icould developer began by designing some layout examples in powerpoint, which were the basis of the data visualisation. Feedback was sought and visualisations refined.

The next stage of the icould development was to apply Standard Occupation Classification (SOC) (this is a common way in which to classify occupations) codes to the back catalogue of videos. This was a manual exercise. In cases where there was not an exact match to the SOC code, the best match or multiple SOC codes were selected. All new videos are now

given a SOC code when they are uploaded to the website. An online tool was created to make it easier to look up and assign SOC codes to videos.

What is the impact of using LMI for All?

'The data have prompted questions from young people on the nature of work and a working week. Users appreciate the value of the data both on its own terms and also as a prompt for wider discussion on careers now and in the future.'

By adopting LMI for All and integrating this into their policies and practices, icould have gained an added-value labour market information resource to complement and extend their existing resources. The integration of the LMI for All data into the icould website has enhanced the offer icould can make to those that subscribe to its API. icould has continued to attract new partners and subscribers since its launch.

What advice would you offer?

Some issues encountered along the way were around occupational coding and styling to ensure data was understandable to the audience as well as being visually attractive.

Occupational coding

In order to map the LMI for All data, organised by SOC code, each of the video stories had to be manually assigned an occupational classification code; this took time. The developer created online tools to help this mapping, by displaying potential SOC code titles and descriptions. The developer was able to select the code that matched or edited the search term in order to find a match. One issue was in searching for particular occupations. For example, in order to get returns for lawyer an additional search on 'law' or 'legal' also needed to be undertaken. icould had technical in-house expertise available to inform and support the occupational coding process.

LMI webpages were created for each SOC linking to the video stories⁵. However, there were concerns as to whether site visitors would understand SOC codes and whether job type may be a better solution. It was decided to continue using the icould classification, as it was tried and tested.

Styling

Styling the data was a big issue and was important to get right, along with meeting accessibility requirements. Colour blocks and more colour were added after the initial visualisation drafts. Table styling was also developed. Work was undertaken on ensuring that the videos and data 'gelled'. This resulted in the data and charts being moved around to look less dense and placing the more appealing data first. Using rough rankings of skill levels were thought to be bland, so decimals were incorporated instead. Initial visualisations

⁵ Recent developments in the LMI for All services would mean that this is now easier, as the full index of job titles have been mapped to SOC within the LMI for All system.

used a spider diagram to display the skills data, but this was changed to the windrose chart as it was thought to be easier to understand.

What next for icould and LMI for All?

'We see this as an important part of enhancing our offer. As long as the LMI for All data continues to be updated, it will remain part of the icould website. The website is maintained and developed using funding from a range of sources, so as long as the website is funded then there will be resourcing in place to continue using LMI for All.'

Making use of icould: learning from practice (2015) reported on the usefulness of the LMI facility for the icould website. In 2015, the icould website and LMI visualisations were redesigned to improve the user experience and ensure it continues to be meaningful. As part of the 2016 site developments, and in response to user feedback, average statistics are being introduced into certain fields. This will improve data interpretation by providing context and giving users a benchmark, allowing for easy comparison between the average salary for a particular job type or average weekly hours, with the national average.

icould have recently set up detailed user metrics to show how the site is used and qualitative user testing is also in the pipeline, which will provide further evaluation into the use of LMI data on the site. icould, with its partners, will continue to evaluate how the data are used to inform further refinements. icould plan to continue using the LMI for All data in the long term.

3.3.2. Prospects: Developing an app

About Prospects

'We inspire the people we work with to develop their full potential through education, employment and training. We will grow Prospects to benefit the individuals we support, our colleagues and our partners by delivering more expert and innovative services to more people each year.'

The Prospects group (<http://www.prospects.co.uk>) delivers a range of services to over 500,000 young people and adults each year. Its user base is wide ranging from young people in schools, vulnerable young people and those not in education, employment or training (NEET), plus adults and offenders.

Prospects is a private sector organisation providing services that include:

- ❖ The National Careers Service in London, the South West and the West Midlands, as well as a partner in Career Yorkshire and the Humber;
- ❖ The Work Programme in the South West and leading subcontractor in London and Bristol;
- ❖ Targeted support for NEET and vulnerable young people;
- ❖ Ofsted Early Years Inspection Services;
- ❖ The Youth Contract;

- ❖ School improvement services and educational resources;
- ❖ Events, business support and networking.

Prospects work with government and local authorities, businesses, the Third Sector, schools and academies, plus further and higher education. It employs 1,400 people and has an annual turnover of over £80 million.

Why LMI for All for an app development?

Prospects were keen to explore ways of providing LMI for All data via an app to raise awareness about career opportunities in London, particularly in the construction sector. They saw real value in LMI for All and the access to up-to-date labour market data.

The Native app⁶, named 'Help Build London', has been developed to complement campaigns and initiatives that aims to attract, inform and retain a talented workforce for the construction and built environment sector, such as the *Go Construct* campaign (see <https://www.goconstruct.org>) and *Build London* website.

Drawing upon their experience of developing a previous app called 'What if I', Prospects were confident that they have the right knowledge and expertise. Research undertaken by Prospects had found that 90 per cent of their customer base has access to a mobile phone; most of which are smart phones.

The app has been developed using in-house developer and graphic design skills.

What is the aim of the app?

The 'Help Build London' app is aimed at raising awareness of current and future opportunities in the construction sector in London. It also encourages users who are wedded to one occupation to consider others that may fit their skills or that have better employment prospects or are in high demand in London.

It is not intended that the app will be used on its own. Instead, Prospects' career advisers will support people to access and use the app during a careers session, as well as respond to careers and learning questions arising from its use. Telephone advisers will also encourage customers to download the app and support them in exploring the data.

The app is also aimed at other stakeholders, including CITB (see <http://www.citb.co.uk/>), Construction Youth Trust (see <http://www.constructionyouth.org.uk/>), colleges, training providers and employers. It is helping Prospects link together a number of useful London-focussed campaigns and tools in a coherent way.

⁶ A Native app was chosen as it enables a user to download it from a public or private app store and store it on their mobile device.

How was the app designed?

Some issues were encountered during the development of the 'Help Build London' app such as determining the focus of the app, refining initial ideas, resourcing technical skills and managing the development within time and budgetary constraints.

An initial challenge was determining the focus of the app and the challenge of refining their initial ideas. Both the construction and creative sectors were initially considered, but eventually, the construction sector was agreed to be the focus of the app

Construction sector – Why?

- Important in London, for the Local Enterprise Partnership and in terms of employment
- Great potential for workforce to move into the sector with the right information
- Need to inspire young people to consider construction as a career
- Option not normally considered by customers, but data show that there is a demand in London

Creative sector – Why?

- Interesting sector in London, but not easy to find information on
- More potential partners in the development of the app and therefore the possibility of creating something more visually interesting
- Receive many queries regarding the sector so there is a need for more information for customers and advisers

Sourcing, tendering and resourcing a developer within the timescale of the development was difficult. It was decided the app would be developed using Appy Pie (see <http://www.appypie.com/>) using in-house skills and knowledge. However, early developments using Appy Pie revealed that it did not have the required functionality, such as being able to select two options. This then led Prospects to building the app in-house to ensure functionality.

The 'Help Build London' app was designed for installation directly onto a mobile device.

How does the app use LMI for All data?

The LMI for All datasets were selected, as they are key to answering some of the questions users will have about the construction sector, such as:

- ❖ Will there be future job opportunities in the construction sector?
- ❖ What skills do I need to work in construction?
- ❖ What jobs are there in construction?

Ideally, Prospects would have liked data on course and vacancy data, as well as data at a local level. These data are not available in LMI for All due to limitations in the current available data. Data included in the 'Help Build London' app includes:

- ❖ Employment (historical, projected and replacement demand);
- ❖ Skills, Abilities, interests data from O*NET.

What does the app do?

A client who is interested in the construction sector in London, enters the app and selects two trades (such as painters and decorators, production managers, electricians, etc.) in construction to compare. These options are based on a selection of skills. Data on how these trades are growing across London is displayed. A chart displaying the Working Futures data on projected employment and employment (replacement demands) is shown. A user can also access more information via a link to the National Careers Service website.

What next for Prospects and LMI for All?

The app is being pilot by a group of Prospects advisers. The app has been promoted at a recent event where potential users could sign up for updates. The app is currently available to use on Android phones, but IOS and Windows versions are being developed.

3.3.3. Skills Match London: Building a website using the LMI for All API

About Skills Match

'Skills Match is an interactive website for different types of people and organisations interested in connecting young Londoners to London's jobs. It is an online resource that allows an exploration of the dynamic between skills supply and employer demand (at Level 3 and below) in London.'

Skills Match (<http://skillsmatch.intelligentlondon.org.uk>) was created in partnership by London Councils (an organisation which represents the 32 London borough councils and the City of London) and education data specialists MIME Consulting.

Skills Match was funded through a successful bid to the Open Data Breakthrough Fund from the Department for Business, Innovation and Skills and the Cabinet Office. The development of Skills Match was brought about by the need to address specific issues, namely: youth unemployment; skills shortages/deficits; skills mismatches; and the lack of good, local skills supply data.

The target audience for Skills Match is careers advisers, schools, colleges, planners, policy makers and employers.

What is on the Skills Match website?

Skills Match provides analysis of skills gaps (expansion and replacement demand) and trends (projected jobs over time) in a visually engaging way. The Skills Match website provides analysis of:

- ❖ Skills gaps: Forecasts of learners successfully completing courses in different subject areas, compared with the vacancies linked to those subject areas, for courses and jobs at Level 3 and below.
- ❖ Projected vacancy trends: Forecasts of new vacancies to 2020 linked to related subject area, for courses and jobs at Level 3 and below.

- ❖ Supply and demand maps: Maps which compare future skills supply and projected vacancies in each London local authority, for courses and jobs at Level 3 and below.
- ❖ Provision mapping: Maps that show providers in London that offer different subjects at Level 3, split by type of qualification.

How does Skills Match use LMI for All?

Skills Match uses LMI for All (principally *Working Futures*) to drive the analysis of likely trends in the London labour market by occupation and qualification level. Other datasets that contribute to Skills Match are:

- ❖ The National Pupil Database (NPD) – for schools provision.
- ❖ The Individualised Learner Record (ILR) – for further education and Apprenticeship provision.
- ❖ Greater London Authority Economics (GLA) data – for overall job projections and replacement demand.
- ❖ The 2011 Census – for local authority distribution of jobs.

Why LMI for All?

As Skills Match was funded through the Open Data Breakthrough Fund (made available to enable open data release and ease of access to open data), Skills Match was developed largely with open data sources, including LMI for All.

Critical to the development of Skills Match was linking and standardising skills and labour market datasets. LMI for All had already connected and standardised existing sources of labour market information, so it was an ideal dataset to contribute to the development of Skills Match.

Additionally, LMI for All provided access to an extensive set of labour market information and was specifically geared towards applications designers that were attempting to address transitions issues into and through the labour market.

How was the LMI for All data used to develop the web interface?

The project team set out the questions that needed to be tackled, that is:

- ❖ What skills are coming through the secondary and tertiary education system?
- ❖ What skills are employers demanding, now and in the future?
- ❖ Where are skills being developed – do we have underlying skills infrastructure issues?
- ❖ Which occupational areas are growing?

Based on the above, MIME Consulting built a multi-levelled theoretical data model then looked at different data sources to match the model. The team initially explored two different, existing jobs data feeds, but the way the data was codified was not useful, and there were data quality issues.

With LMI for All, the team found that the way the data was codified could be more easily matched to the skills data that was being used from the National Pupil Database and the Individual Learner Record, although it was still a significant task to make the match.

What advice would you offer in using LMI for All?

Coding

As there was no pre-existing mapping of subjects to occupation codes, the team had to produce its own mapping of Sector Subject Area (SSA) 3 course codes to Standard Occupational Classification (SOC) 4 occupation codes. This was done on the basis of word similarities, as well as research into the nature of particular courses.

Since there are 369 SOC codes and only 130 SSA3 codes, most subjects mapped to multiple occupations. Where one job code mapped to more than one subject area, the demand for jobs was apportioned evenly between subject areas. In practice this will be an oversimplification, but there is no hard data available showing exactly how the codes should be apportioned.

Granularity

When exploring LMI for All, the team found that 'region' was the lowest level of available data, yet local authority level data was key to the Skills Match specification. To address this, the team used zone information available in the 2011 census to create regional data.

Geographical context

GLA Economics data (for overall job projections and replacement demand) was also used as it provides London level analysis that takes account of the London context (for example, housing and population growth), which does not feature in nationally produced datasets. The team also needed to ensure that Skills Match complemented the London economy 'story' that the GLA Economics analysis sets out.

What is the impact of using LMI for All?

Forecasting analysis tools have to tread the line between art and science with the skill of a mountain goat, so data credibility is essential.

LMI for All provided the team with comprehensive, robust and nationally recognised data sources, which in turn provided Skills Match with credibility.

What next for Skills Match and LMI for All?

The team would like to apply the next wave of UKCES (*Working Futures*) modelled data to the site (and update the latest skills data from the Department for Education and the Skills Funding Agency).

The team is also interested in modelling higher level data for jobs at Level 4 and above using Higher Education data in addition to LMI for All.

3.3.4. KareerHub: Building an online LMI dashboard using LMI for All

About Kore Education Systems

Kore Education Systems is a division of Kore Management Systems LLP (KMS) – a Northern Ireland business established in 2011. Its roots stem originally from Project Information Management Systems for Engineering and Construction projects. The first Cloud-based implementation was in 2004. Clients have included British Airport Authority (BAA), Prudential Pensions, Merrill Lynch, Price Waterhouse-Cooper (PwC) for major projects in the UK and Europe.

Two years ago, KMS were approached by a leading Northern Ireland education expert to support the development of online learning resources for careers education, information, advice and guidance (CEIAG) in schools and colleges. Having scoped a possible ‘cloud’ based CEIAG delivery project, followed by preliminary testing, Kore Management Systems applied for a TechstartNI (see <http://www.techstartni.com/>) ‘Proof of Concept’ grant to further develop ‘KareerHub’. This was successful in May 2015.

The system was initially piloted in 32 post-primary schools across Northern Ireland. Feedback responses indicated high demand for greater access to high quality, reliable labour market information and intelligence (LMI) from teachers and from interaction with pupils in classroom sessions.

KareerHub

KareerHub is still under development, but you can register your interest on the Kareerhub website (<http://kareerhub.co.uk/>).

During the initial development phases of KareerHub, the Department for Employment and Learning (DEL) in conjunction with the University of Ulster published a new Northern Ireland Skills Barometer (see <https://www.delni.gov.uk/publications/ni-skills-barometer>). This contains an excellent analysis of LMI in Northern Ireland, but is a fairly dense dataset. The challenge was to find ways of integrating this into the KareerHub database and to develop several ‘Web Quests’. These directed students, teachers, parents/carers and careers advisers to relevant areas of the ‘Skills in Demand’ document.

KareerHub and LMI for All

Whilst early thinking for KareerHub was centred on the Northern Ireland Skills Barometer, LMI for All offered potential to expand the KareerHub database. LMI for All makes a significant contribution to the aim of KareerHub, which is to make LMI accessible both within and outside of the classroom. For the first phase of development, Careerometer was embedded in the KareerHub portal. The next phase will involve full integration of the LMI for All Application Programme Interface (API) into the KareerHub database.

Why LMI for All?

The KareerHub team consider LMI for All as offering a unique way to access a number of official datasets. It was considered a high quality source.

The KareerHub team were initially unsure about UK LMI sources, their scope and reliability. However, LMI for All has addressed these uncertainties and the team believe they have a reliable platform as a result.

The KareerHub project team have a clear understanding of the everyday problems faced by teachers and careers advisers in accessing high quality LMI. Working with practitioners, the demand for LMI was clear and requirements were identified. LMI for All data helped meet these demands.

What was the process for developing KareerHub using LMI for All?

The KareerHub team designed a development plan including timescales. Technologies that could be utilised for the platform were researched first. The team engaged with ElasticSearch, Apache Solr and other technologies.

The second development phase involved researching the relevant datasets available through LMI for All and looking at how others are using LMI for All, namely Skills Match and Plotr.

The KareerHub team have looked at a number of datasets from LMI for All and established how best to connect to the NI Skills Barometer database. Datasets considered including:

- ❖ Employment, projected employment and replacement demands from Working Futures;
- ❖ Pay and earnings based on the Annual Survey of Hours and Earnings and the Labour Force Survey;
- ❖ Hours based on the Annual Survey of Hours and Earnings;
- ❖ Unemployment rates based on based on the Labour Force Survey;
- ❖ Skills shortage vacancies based on the Employer Skills Survey;
- ❖ Skills, Abilities and Interests based on the US O*NET system;
- ❖ Occupational descriptions from the ONS Standard Occupational Classifications;
- ❖ Current vacancies available from Universal JobMatch;
- ❖ Higher education student destinations data from HESA.

Internal ICT and management expertise has been drawn up. An external developer and the LMI for All technical team have also supported the development process. A career education and guidance specialist has provided UK and international research findings on the use of technologies and LMI.

What is the impact of using LMI for All?

The LMI for All portal is a real value-added function in our KareerHub portal. So far, a significant number of the Northern Ireland post-primary schools have used KareerHub. The first phase of the project has demonstrated the technology works in the classroom.

Importantly, students, teachers and careers advisers welcome the concept of a Dashboard, which can present CEIAG and LMI sources in different formats.

What advice would you offer in using LMI for All?

Access to the LMI for All Developers toolkit has significantly reduced the development time and cost of the dashboard. This has enabled KareerHub to test and validate the early development ahead of schedule. This was considered to add credibility to the project with participating schools and colleges. The KareerHub team consider it crucial in ICT projects to be able to feed users with a programme of continuous improvement updates to retain their engagement and confidence.

What next for KareerHub?

A team of established education sector market testers, including careers and employability advisers in Northern Ireland post-primary schools, have tested the early dashboard development. The next phase, underway, will utilise the LMI for All API, other existing APIs and the Northern Ireland Skills Barometer for full integration into the KareerHub database. Integrated data will be displayed on the KareerHub dashboard. The first version of commercial release is planned for September 2016.

3.4. Users of LMI for All

3.4.1. LMI for All web interfaces

There are a number of web interfaces that have been developed throughout the lifetime of the project and are available to careers professionals and clients/customers. These have been promoted as part of the engagement activities detailed in section 3.5, below. The LMI for All system does not allow us to easily identify third party users, so the following list of selected users is likely to be incomplete. The imminent implementation of the API key system will, however, enable better knowledge of LMI for All third party users. Selected users of LMI for All include:

- ❖ Active informatics;
- ❖ Adviza;
- ❖ KoreEduSys (KES);
- ❖ National Careers Service;
- ❖ Plotr;
- ❖ RCU Ltd.;
- ❖ SACU;
- ❖ Skills Development Scotland
- ❖ Skills Match;
- ❖ SkillsPlanner;

- ❖ UExplore.

Two applications have been developed, by LogToMobile and Prospects, for users to access LMI for All from mobile devices.

3.4.2. Careerometer in practice

As part of the LMI for All service, the UKCES commissioned the development of a customisable widget, called Careerometer (see <http://www.lmiforall.org.uk/widget/>). This has been made freely available for third parties to embed in their website as an easy interface to LMI for All. The widget has been provided to encourage and support third parties to engage with the LMI for All service, who may not have the resourcing to develop their own interface. In the Careerometer widget, the user types in the title of the job they are interested in and the widget provides a series of options from which the user can select the most relevant. It then provides access to a selection of headline data relating to pay, weekly hours of work and a visualisation of future employment prospects; all organised by occupation. The widget can be used to explore and compare key information about occupations to support the process of identifying potential careers.

Careerometer has been embedded in an estimated 32 websites, including:

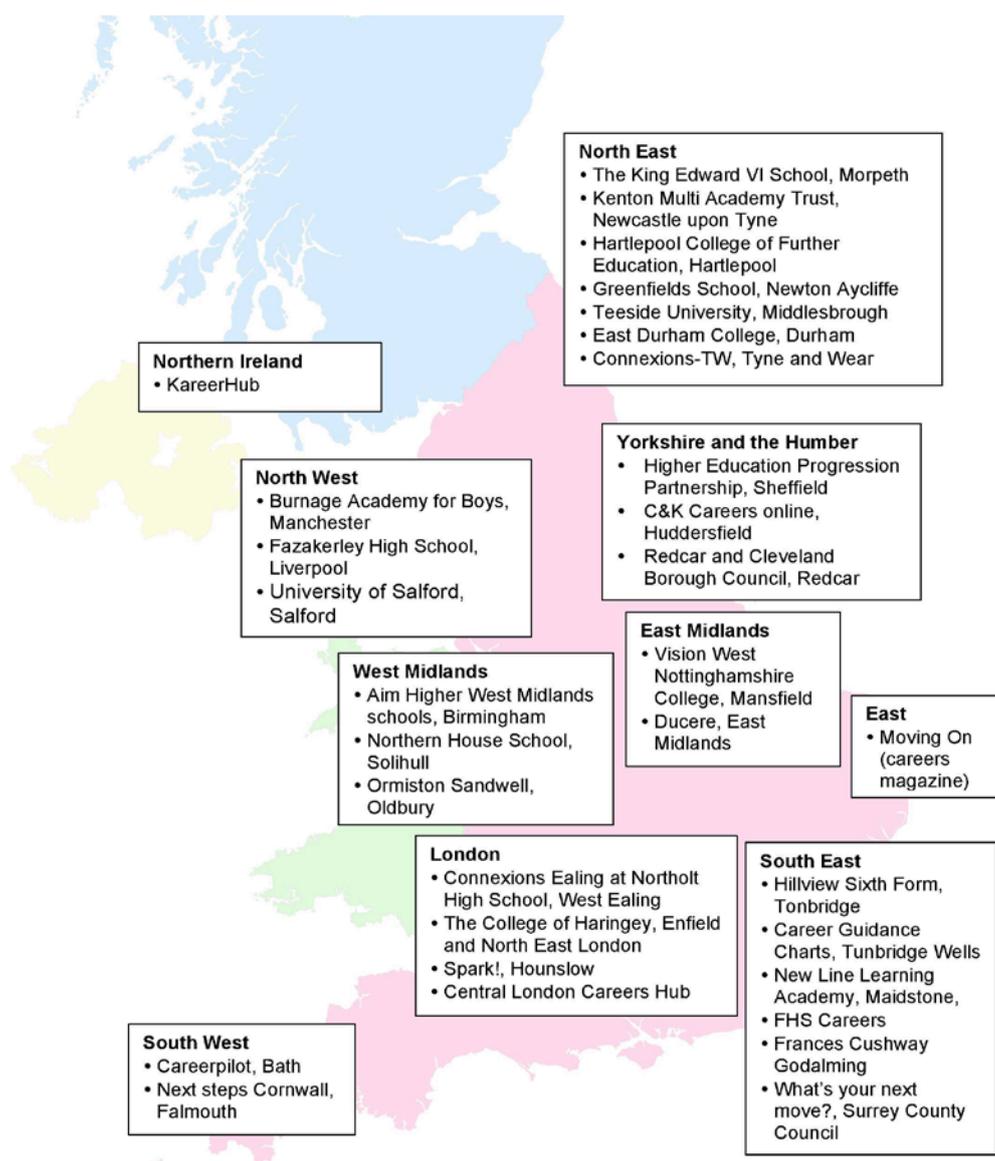
- ❖ One online careers magazine (a number of other websites point to Careerometer on this resource);
- ❖ One campaign to promote an occupation;
- ❖ Three higher education institutions, schools, academies and colleges consortia websites;
- ❖ Nine schools and academies;
- ❖ Five further education colleges and sixth forms;
- ❖ Eleven career organisations and services; and
- ❖ Two higher education institutions.

Moving On magazine (a print and online careers magazine) has embedded Careerometer in its website, see: <http://movingonmagazine.co.uk/careerometer/>. Moving On is aimed at 14-18 year olds in English schools, colleges and sixth forms and has an estimated readership of 250,000 plus and a circulation of 25,000 plus. The aim is to provide young people and parents a range of information regarding post-compulsory education choices. A number of organisations (particularly schools) point to this resource on their own websites.

The Gatsby Technicians make it happen campaign have also embedded Careerometer in their website (see: <http://technicians.org.uk/make-it-happen/>).

A detailed list of websites is in Annex A, but it should be noted that this may not be a comprehensive list as users are not required to register their use of the widget. The following map provides an overview of the geographic coverage of users of Careerometer.

Figure 3.2 Map of Careerometer users



This is also available on the LMI for All website at: <http://www.lmiforall.org.uk/widget/careerometer-in-practice/>.

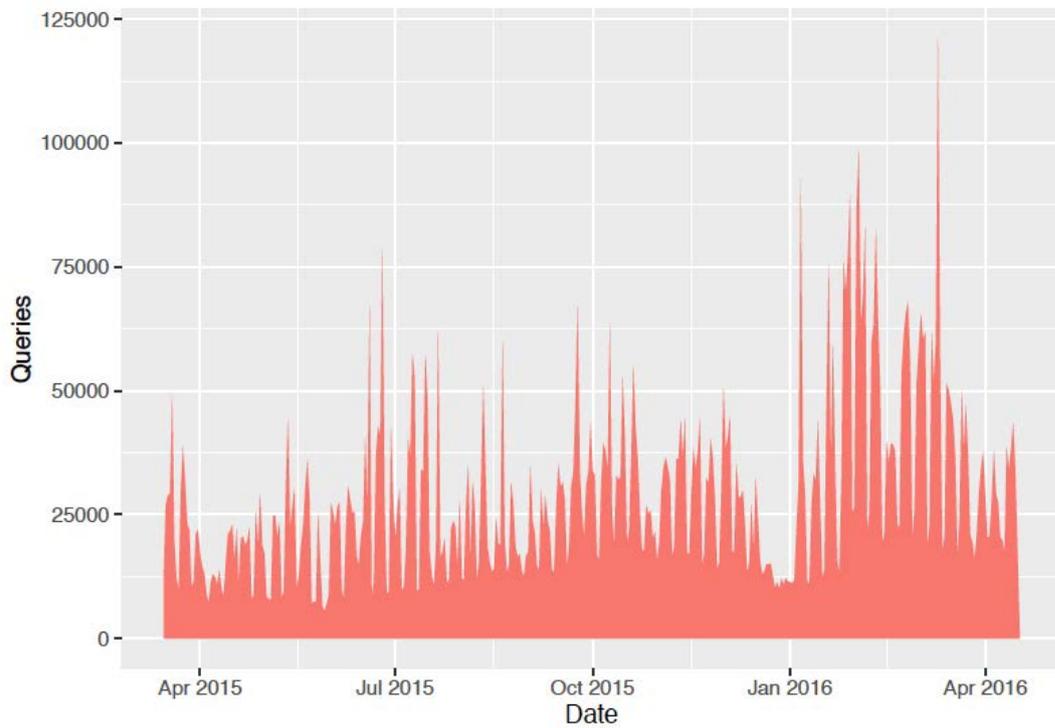
3.5. LMI for All usage statistics

The following reports on LMI for All usage from April 2015 to April 2016. During this reporting period, 24,369,145 requests were made to the LMI For All API in total.

3.5.1. Query volume

It is important to note that the data for the following charts have been 'cleaned' of mass downloaders skewing the results. There were occasional mass downloading from a site in France for a few months during 2015 causing extremely high O*NET traffic. Service updates now prevent this kind of abuse.

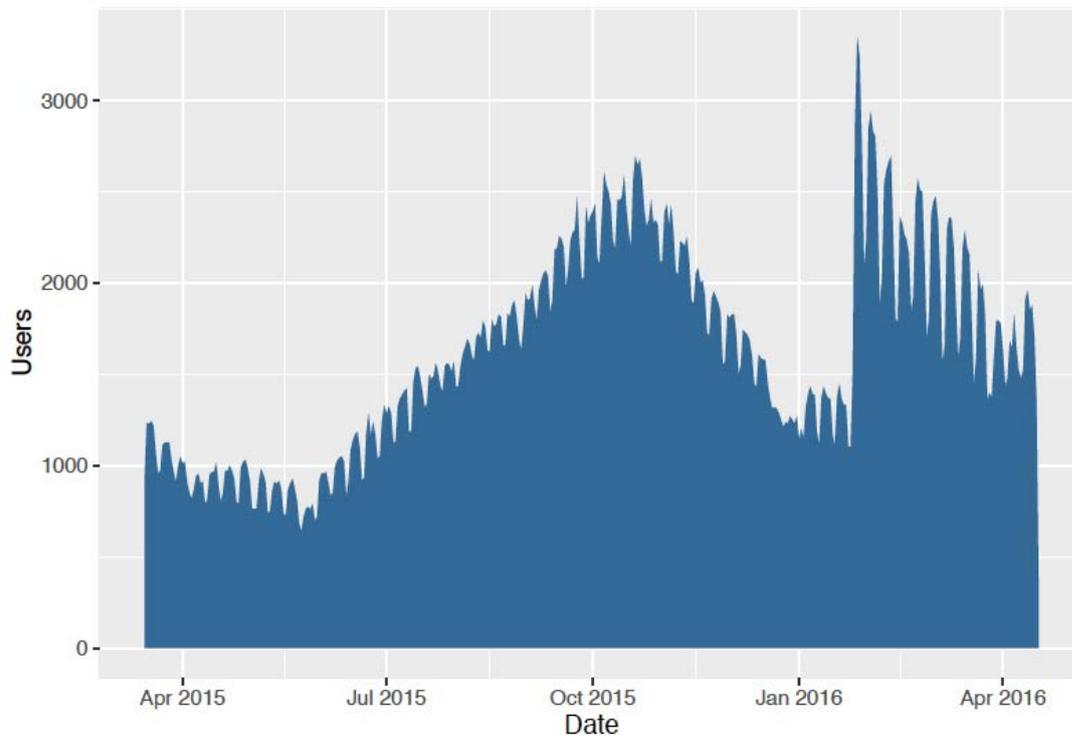
Figure 3.3 LMI for All query volume, 2015-2016



3.5.2. Unique users

Aggregating requests onto the same IP per day gives us a rough picture of the behaviour of regular users. Note the characteristic ‘weekly’ bumps.

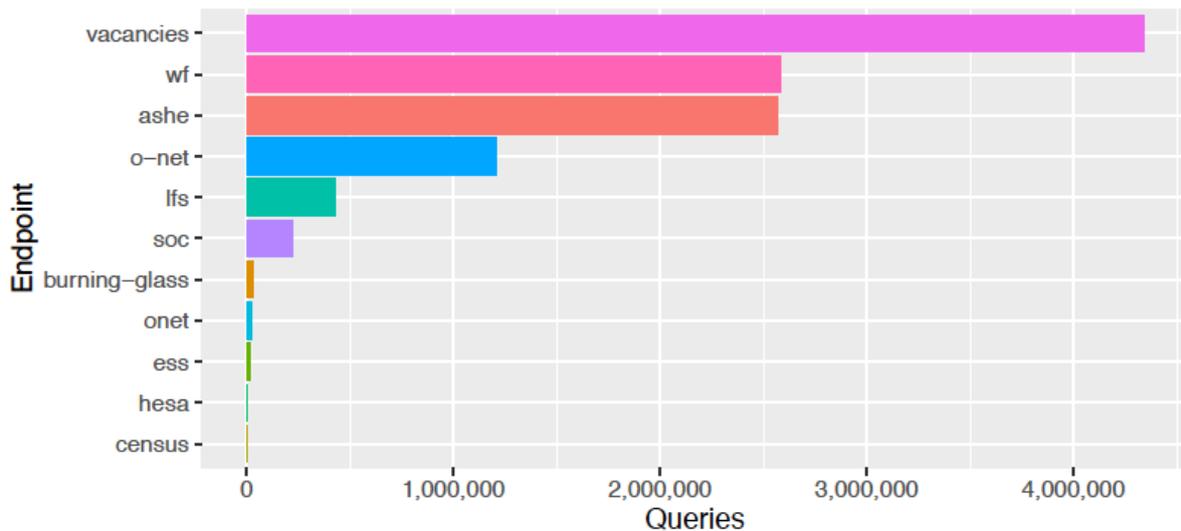
Figure 3.4 LMI for All unique users, 2015-2016



We have not been able to pinpoint The cause of the slow drop off during December 2015 and the sudden spike in January 2016 can not be pinpointed. However, it is assumed that the first is due to the holiday period in general, possibly extending into January, and the spike is the release of a new app. The new API key system is now operational and more reliable than the previous version, which should help to identify traffic sources with much more precision.

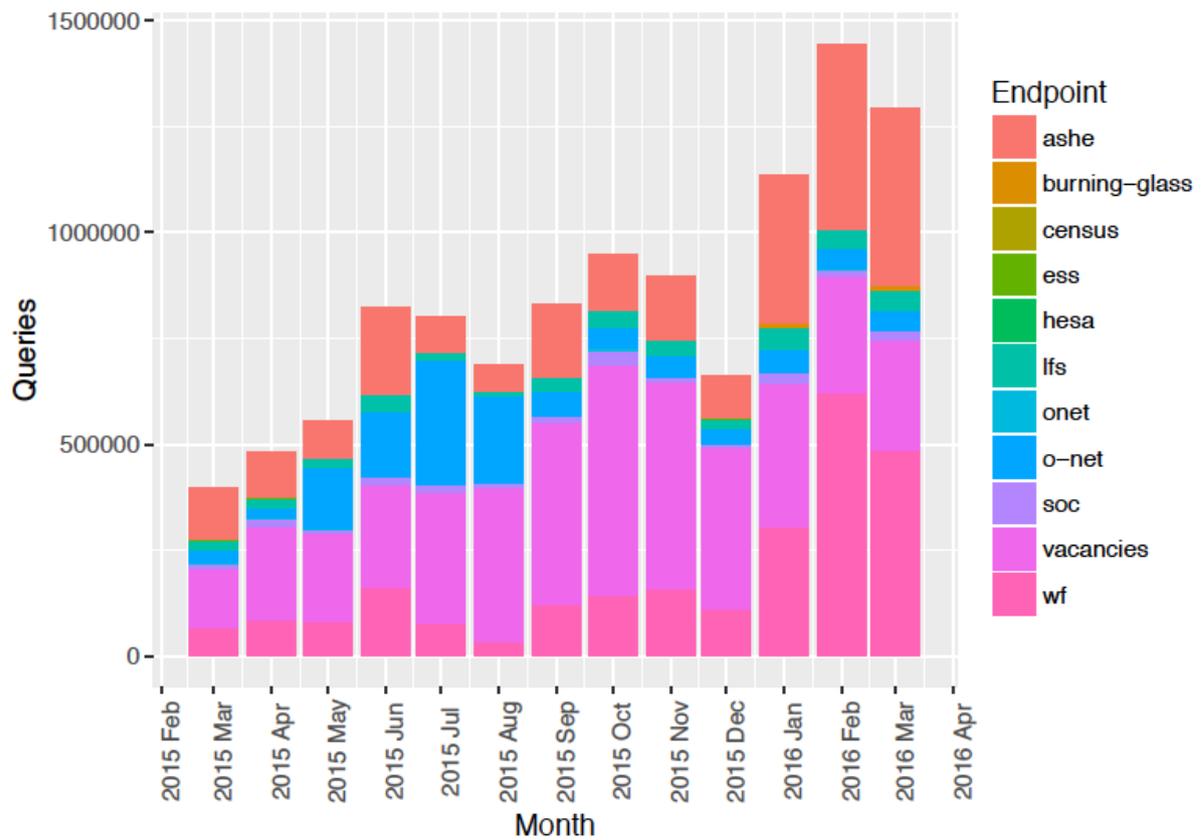
3.5.3. Traffic by endpoint

Figure 3.5 LMI for All traffic by endpoint, 2015-2016



Vacancies continue to be the most requested data available from the LMI for All API. This holds mostly true month-to-month as well, with spiking interest in the Annual Survey of Hours and Earnings (ASHE) data, Working Futures and O*NET. Recently queries to Working Futures seem to outpace queries to Universal JobMatch. This is likely due to some frequently used app (possibly eclips-online.co.uk) adding deeper information based on Working Futures or ASHE data. The new metrics system to come online shortly will mean that fairly accurate app usage profiles that break down endpoint usage by individual apps can be generated.

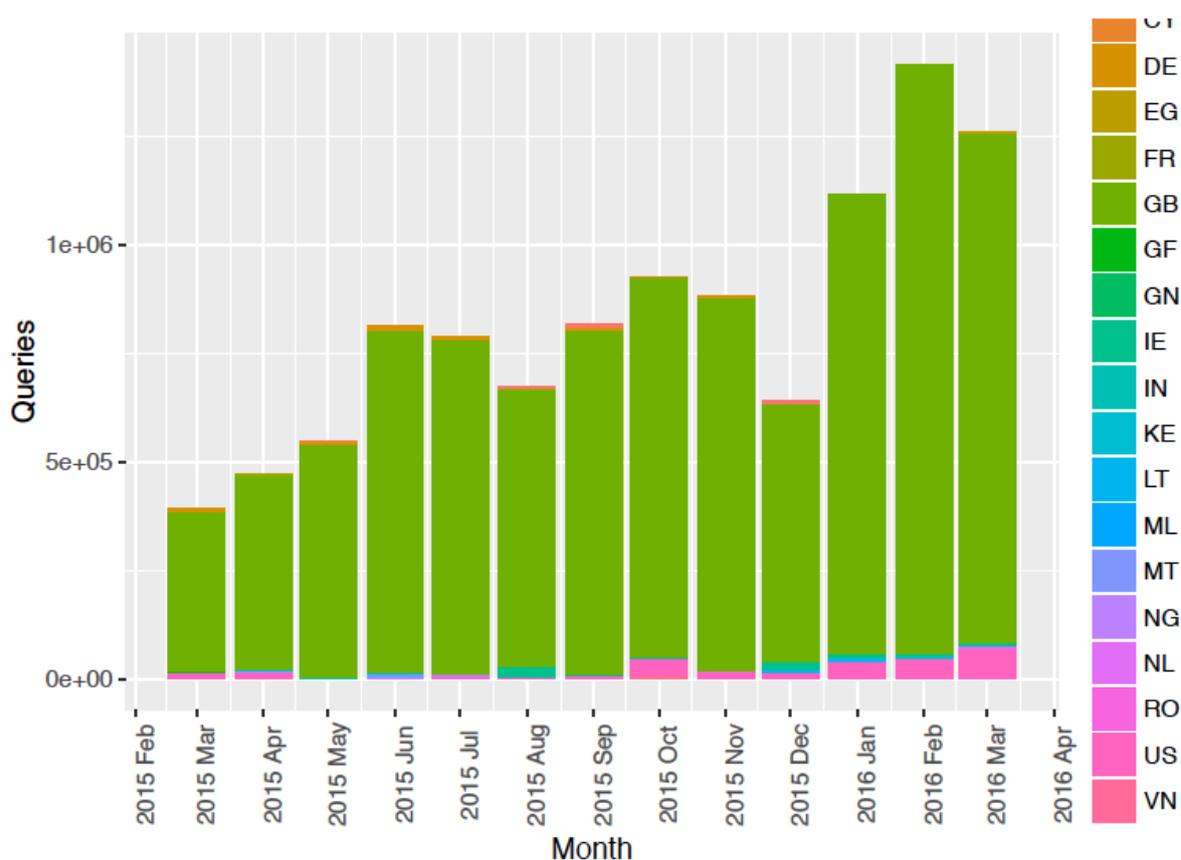
Figure 3.6 LMI for All traffic by endpoint and month, 2015-2016



3.5.4. Traffic by country

LMI for All traffic mainly comes from the UK, with some interest from the US growing over recent months.

Figure 3.7 LMI for All traffic by country, 2015-2016



3.6. Engagement activities

Dissemination and stakeholder engagement activities (such as keynotes, presentations, workshops and exhibition stands) have been undertaken through a series of events. The aim was to: increase awareness of the LMI for All service; gain feedback that can inform the continued, and future, development of the service; and explore the potential for linking to the data tool from existing websites (namely the Careerometer, icould, RCU, plotr and Adviza). During this phase of the project, the LMI for All service has been promoted and disseminated at 25 events across the UK to a range of audiences (approximately 1,500 people) and six international events (reaching over 100 people). The key stakeholder audiences have included:

- ❖ Careers professionals and managers from across the sector;
- ❖ Careers delivery organisations from across the UK;
- ❖ Careers membership associations;
- ❖ Schools and membership networks and associations;
- ❖ Further education colleges;
- ❖ Higher education institutions.
- ❖ Developers and information service providers; and

- ❖ Local government bodies.

General messages that have come through consistently from engagement activities include the following:

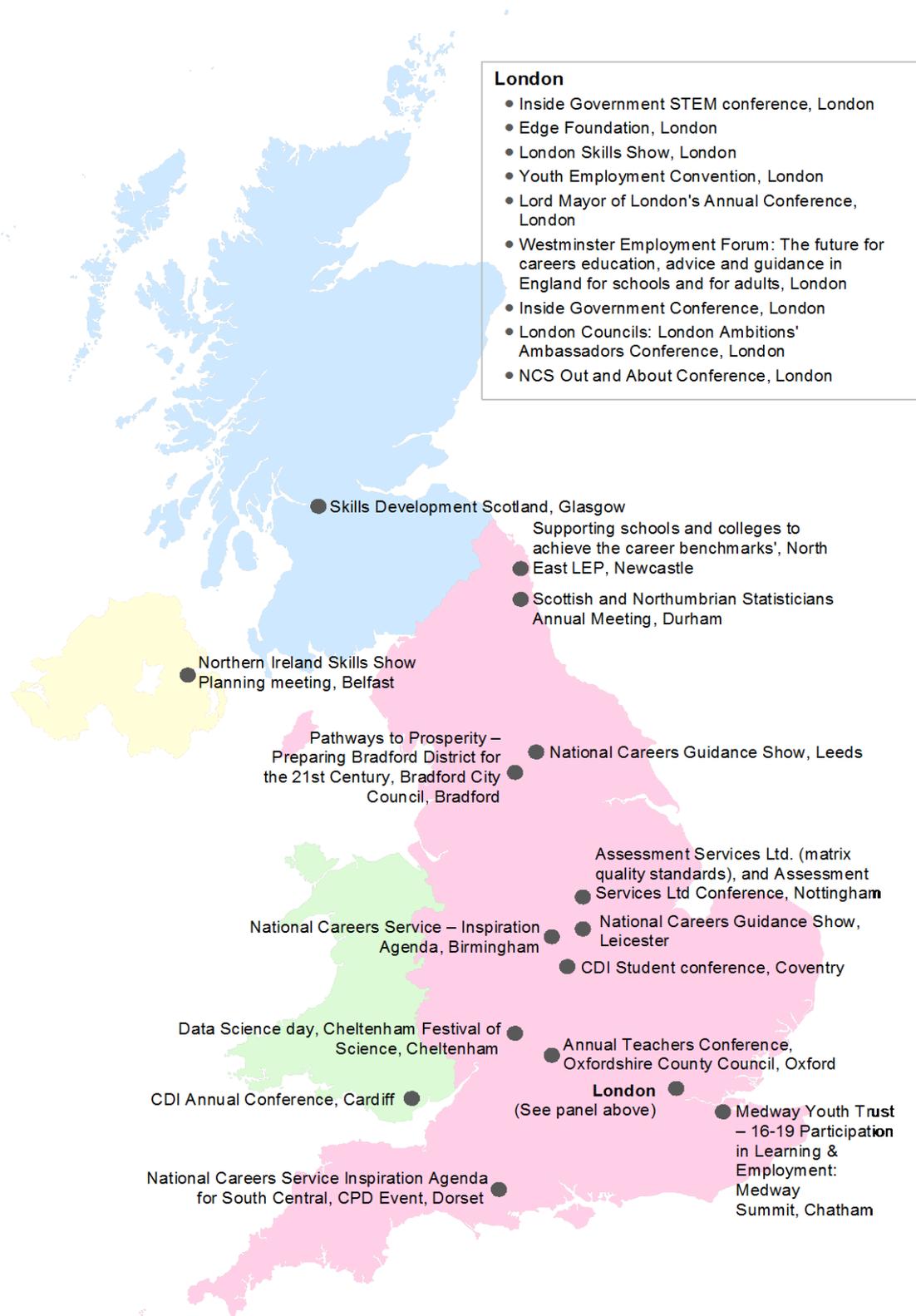
- ❖ A high level of awareness of the LMI for All service exists amongst careers professionals.
- ❖ A low level of awareness of the service at events targeted at schools and colleges. However, awareness amongst this group has improved as evidenced by the take-up of Careerometer predominately by schools (see Section 3.4.2 and Annex A).
- ❖ Workshops and seminars have been the most successful dissemination activities for gauging opinion on the usefulness of the service and potential applications.
- ❖ Positive feedback has consistently been received from a range of stakeholders, potential third party users, developers and others (such as policy makers). The potential for the service is clear for a range of stakeholders, so generating buy-in has not be an concern for those taking ideas back to their organisations and management. Problems are apparent when moving from understanding the potential of the LMI for All service to producing an interface or application.
- ❖ Major issues for potential users of the service have included:
 - Understanding the development process;
 - Identifying a developer and then briefing that developer; and
 - Resourcing the development process.

However, it should be noted that many organisations have been able to talk through and resolve these issues. It is evident that there is a need to further support stakeholders in creating an interface. This could be achieved by creating a database of developers that have worked on LMI for All.

- ❖ Identifying a target audience, articulating their particular needs, and selecting what data to use in an interface have not been problematic.
- ❖ A number of stakeholders are still progressing with the creation of their own interface to LMI for All, so continued use of the service and the likelihood for future developments are high.

All stakeholder engagement activities are listed in Annex B, but the following map provides an overview of geographic coverage of the stakeholder engagement services.

Figure 3.8 Map of stakeholder engagement activity, 2015-2016



An article on the LMI for All service was published in the SSAT Journal (see Barnes, Bimrose, and Hughes, 2015), which reaches members of schools at all phases of education.

3.7. Recommendations for future stakeholder engagement services

- ❖ Continuation of dissemination activities to maintain the momentum of the 2015-2016 activities. This is particularly important, as there is a need to address recent uncertainties about the future of LMI for All. These dissemination activities could be a combination of presentations and workshops at key stakeholder conferences and events, as well as articles in professional journals (such as CDI, ASCL, NAHT and NUT).
- ❖ A high profile dissemination event to showcase the LMI for All service and demonstrate the longevity of the service. The event could also showcase good and interesting policies and practices from across the UK. The key aim would be to cascade learning in the use of exploiting LMI for All for end users. There may also be a need for a smaller event in which developers and potential users of the LMI for All service could be introduced, and the potential of data demonstrated.
- ❖ A series of regional workshops aimed at:
 - Schools and colleges – to explore ways of using LMI for All within and across the curriculum. The potential of the different datasets in LMI for All could also be explored to encourage their use.
 - Local enterprise partnerships and local/combined authorities – to explore ways of using LMI for All within local development plans.
- ❖ The Careerometer widget has been successful in getting people to engage with the LMI for All service. The development of a set of widgets each focused on making specific datasets within LMI for All accessible could be developed. These widgets would demonstrate the potential of different datasets.
- ❖ There is a need to continue working with those stakeholders who have started the development process. It is evident that the development process can take a number of months from design to implementation.
- ❖ The LMI for All website could also be developed to include an animation on LMI for All for non-technical users. This would help promote the service and illustrate how third parties can access and use the service.
- ❖ As the number of LMI for All interfaces has increased, there is now potential to undertake a small-scale end-user evaluation. This could be undertaken alongside third parties and their users to gain feedback on the service. Broader audiences could also be engaged to determine the potential for LMI for All services to be used in other contexts such as in local enterprise partnerships and further and higher education course planning.
- ❖ To date, end user engagement has been a key principle in stakeholder engagement activities. So potential end users have identified the data sets they require as part of

their development. However, there is potential in the future to promote particular datasets (e.g. take up of the HESA destinations data is currently low, yet represents a potentially valuable resource for schools), by developing worked practical examples of how these data could be used creatively in applications, alongside other datasets.

4. Data development services

4.1. 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);⁷
- ❖ Sector (up to the 2-digit level of SIC2007, about 80 categories);
- ❖ Geographical area (12 English regions and constituent countries of the UK);⁸
- ❖ 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.⁹

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, i.e., RDs);
- ❖ Pay (for a recent year (currently 2015, for employees only));
- ❖ Hours (for a recent year (currently 2015, for employees only)).

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

- ❖ Occupational descriptions for 4-digit occupations (based on ONS information);
- ❖ Skills data mapped to 4-digit occupations based on US O*NET information;
- ❖ Changes in pay by detailed 4-digit occupation (between 2014 and 2015);
- ❖ 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 and real-time vacancy data from Burning Glass;
- ❖ Census (limited information for 2011 on occupational employment at a detailed geographical level and on travel to work distances); and

⁷ 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.

⁸ Plus in some cases additional information on: age; gender; status; and qualification (highest held).

⁹ In particular historical time series are not available at the 4 digit level.

- ❖ Occupational destinations of graduates by detailed 4-digit occupation (based on HESA data);
- ❖ Apprenticeship Frameworks and Standards including vacancy information.

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. This was in line with Sir John Holman's original vision that the service should be dynamic. 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* employment database (See, for example Wilson *et al.* (2016a, 2016b). 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;¹⁰ 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 would 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 as well as a method to derive implications for the distribution of pay (estimates of medians and deciles).

4.2. Summary of data and indicators included in the portal

4.2.1. 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 new Regulated Qualifications Framework (which has now replaced 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, available at the full level of detail for 2014 only, are constrained to match the main *Working Futures* database using an extended version of the algorithm

¹⁰ 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.

developed to produce the main *Working Futures* database (For details see Wilson *et al.* (2016a)).

Although estimates can be generated for the full level of detail shown at the start of this section, not all of these are reliable and robust. In order to rule out such information, the API censors results that fall below a certain threshold and flags up cases where the estimates may be less reliable. These criteria are based on rules 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 historical estimates and projections by occupation at the 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.¹¹ 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

¹¹ 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*).

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.

Figure 4.1 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 aspects 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	Number of vacancies
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)	Burning Glass
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	Number of vacancies
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
Period	2002-2014 #	2014-2024	2014-2024	2013, 2014 and 2015	2013, 2014 and 2015	2011-2015	2011, 2013 and 2015	2015
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	annually

Dataset	Number of vacancies	Skills, Knowledge, Abilities, Interests	Occupational descriptions	Current vacancies	UK Census Population	Higher education destinations	Apprenticeships
Source	Burning Glass	O*NET database	ONS Standard Occupational Classifications	Universal JobMatch	UK Census Population (England and Wales)	HESA	Apprenticeship Service
Indicator	Number of vacancies	Skills, Knowledge, 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	Apprenticeship data and frameworks
Dimensions*	Occupation	Occupation	Occupation	Occupation	Occupation (1,2,3 digit), Geography**	Occupation, Qualification, Qualification required for job, Subject of study	Occupation
Period	2015	2015	2010	current	2011	2011/12-2013/14	current
Available updates (if known)	annually	every 2-3 years	only required when SOC is updated	constant	every 10 years	annually	constant

Notes:

* Occupation (SOC2010 4-digit), Industry (SIC2007, 75 industries), Qualification (RQF 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 2000-2013 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 was 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 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.¹² 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.

4.2.2. 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.

¹² In practice, these employment data are provide to the Technical Team alongside the Pay and Hours data so it is not necessary to delve into the separate part of the database containing the full *Working Futures* employment estimates.

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 subset of set of skills and related attributes selected for their utility in the present context. 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	Description¹³
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
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 associated with the Work Context content area – physical

¹³ For more information on these indicators see: <https://www.onetonline.org/find/descriptor/>

Categories	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 SOC2010.¹⁴ This is the LFS. This adopts the standard ILO definition of unemployment (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 SOC2000 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.

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 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.

¹⁴ The official claimant series uses SOC2000 and hence cannot be used.

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 becomes increasingly out of date 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 Burning Glass)

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 vacancies generated by Burning Glass. In contrast to the DWP/Monster data considered in previous cycles of LMI for All, these data are coded using SOC occupational data coded to the SOC2010. The Burning Glass data feed includes steps to de-duplicate multiple postings of vacancies. The data provided are one-month old, and were updated on a daily basis during the present pilot test of the vacancy data which ended in March. A “frozen” version of the dataset (i.e. no longer updated on a daily basis) is currently available until of May to developers who signed up to the dataset prior to 31 March. At present only the most recent value is available for a given occupation and it is recommended that a solution be explored to allow a time series of the data to be made available to applications.

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. In the absence of the latter development, an alternative solution from Burning Glass has been sought and incorporated that is matched to SOC.

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 cover the entire UK and use the SOC2010 classification, as did the 2011 survey.

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 2015. 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 and Burning Glass)

In principle, the data on vacancies collected by Monster on behalf of DWP provided 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 included 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 included 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 used 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 implemented a “fuzzy matching” based on reported job titles, which provides a feed of vacancy information from the DWP/Monster website. This included details of actual vacancies rather than any attempt to quantify the overall number of vacancies or estimate a vacancy rate. The information however had limited value, as it was 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 CASCOT¹⁵ software package as a possible solution to the mapping problem.

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 have allowed Monster to make data available recoded to SOC2010. At present such data are not in the public domain nor

¹⁵ 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>

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 seemed 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 was 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. Since jobcentre vacancies only represent a fraction of the market, it is necessary to complement these with other sources of vacancy data.

By way of exploring an alternative source of real-time vacancy information that is matched to SOC2010 the UK Commission has licenced an API feed from Burning Glass. This provides the most recent vacancy data after a one-month time lag, updated on a daily basis, but currently excluding the preceding time series of values. Cross-classifications by SOC,

industry, country, region and travel to work area are provided in the LMI for All API. Given the commercial nature of the Burning Glass data, limitations on crawling the complete dataset have been introduced by means of API keys.

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 for job seekers is the provision of data for neighbourhoods 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 (though there is no more recent alternative source of this information). 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	2013/2014
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 (2013/14)	Principal subject of study.
F_XJACS201OLD	Subject of study	This is now redundant and only contains 'na' - Not Applicable,

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.

Apprenticeship vacancies

The Skills Funding Agency (SFA) make vacancy data for apprenticeships in England available through an API. The aim was to offer this information as an integral part of LMI for All, even though apprenticeship Frameworks and Standards are not classified to SOC 2010. The task involved the classification of the job role descriptors that are included in each Framework or standard to occupation (SOC 2010) using the CASCOT software, as well as expert judgement to resolve any ambiguities. Linking the vacancy data via SOC 2010 then enabled end-users of LMI for All to identify apprenticeship opportunities that are relevant to the occupation they are interested in. The LMI for All API links to the SFA apprenticeship vacancy API feed and allows an application to provide information on apprenticeship programmes by either SOC2010 code, programme code, or a text search.

Example use-cases include the following:

- ❖ An application that presents occupational data on pay, employment etc., but alongside it renders information on current apprenticeship vacancies relevant to that occupation.
- ❖ Selection of an apprenticeship title to obtain the related LMI information derived from the associated SOC codes, including the set of job titles that are applicable.
- ❖ Conversely, knowing a job title (and hence the API being able to derive the SOC code) should allow a user to be presented with a list of potentially applicable apprenticeship frameworks.
- ❖ Finally, given the latter information, the system should be able to pull out available apprenticeships in the local geographical area.

The common link across LMI for All datasets of SOC 2010 means that labour market opportunities, skills and returns can be linked to apprenticeship opportunities that may be relevant to a given occupation.

4.3. Data development summary

The data development strand of Phase 2B 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, including the piloting of the Burning Glass vacancy feed, but could not be completely resolved within 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 of Population, *Working Futures* and ESS, Apprenticeship frameworks and standards) 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 Regulated Qualifications Framework [RQF]);
- ❖ 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.

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 of these sources.

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 provided 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 was a priority to make these data available (or an equivalent dataset) for the LMI for All database. As a result an API feed of vacancy data generated by Burning Glass has been incorporated. Initially incorporated on a trial basis, the efficacy of the Burning Glass vacancy data needs to be evaluated, along with the restrictions imposed by the licence in relation to open-data principles which underlie LMI for All.

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. Trying to locate and include course information and data into LMI for All continues to be problematic, as identified in earlier phases of the LMI for All development (see Bimrose *et al.*, 2012, 2015). There are issues with mapping to occupation (defined by SOC) and subject classification and collating course information. For example, there is no occupational flag in the BIS data on FE outcomes and destination data from HMRC and so no easy way to integrate into LMI for All. Consistency and quality also continue to be a concern with available sources. Skills Match has developed a limited mapping between SOC and Skills Funding Agency subject areas for London. However, overall, it seems unlikely that the disparate national sources of course data can readily be pulled together to create a complete dataset.

LMI for All includes the highest qualification held (9 RQF categories) related to employment, replacement demand, unemployment rates, as well pay and earnings. However, for the reasons indicated, more detailed breakdowns by course would be difficult to map to occupation by SOC2010.

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:

- ❖ 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, EU Skills Panorama, ESCO and EURES

Cedefop projections, Skills Panorama 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 were initially classified using ISCO 88, which is not directly comparable with SOC2010 – however, a broad brush mapping can be derived (see below).
- ❖ The new data published in 2014/15 used 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 and the next phase will move from the current policy orientation to careers guidance counselling, but retaining the limited resolution of 2-digit occupational classification for most labour market information. The current version can be found at: <http://euskillspace.cedefop.europa.eu>

EURES portal

The EURES portal evolved from a vacancy database in 1994 to its present form in 2006. It constitutes the European network for worker mobility with European Commission and the public employment services (PES) of the EU countries and Norway, Iceland and Liechtenstein (EEA Member States) as well as the social partners' organisations. The EURES network exchanges information and enables cooperation among its stakeholders to facilitate freedom of movement of workers across the EU. The services offered by EURES include advice, information and assistance for placement and recruitment. The EURES portal collects together job vacancies processed by national PES, which constitutes around 40 per cent of the overall European job vacancy market.

Further information on EURES can be found at: <https://ec.europa.eu/eures/public/homepage>

Though EURES does not constitute open data, it is possible for a third-party service to obtain access to the EURES API after an approval process has been completed. Thus, it is worth exploring whether the API to these European wide vacancy and mobility services can be included in future cycles of LMI for All.

Future developments of ESCO

The European Skills, Competences, Qualifications and Occupations (ESCO) is the multilingual classification (currently in 25 languages) which identifies and categorises skills, competences, qualifications and occupations for the EU labour market, education and training (European Commission, 2013). ESCO offers occupational profiles of the relationships between occupations, skills, competences and qualifications. It has been developed using open IT standards and may be accessed through an online portal. ESCO is the only pan-European standard terminology which maps skills and competences and qualifications to occupations. It will be available free of charge to all labour market, education institutions and other stakeholders. After some years in development and beta testing an API for what shall be the first full release of ESCO will go live later in 2016. As the classification has been developed with international experts in a European context and resolves down to the 5-digit level of the International Standard Classification of Occupations, future cycles of LMI for All may benefit from exploring mapping that occupational classification to UK SOC2010 using a similar approach to the skills information from O*NET, i.e., via the

CASCOT classification software. By this means a plug-in to the ESCO API would become possible.

Additional information on ESCO can be located at: <https://ec.europa.eu/esco/portal/home>

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.

5.2. Stakeholder engagement activities: maintaining momentum

5.2.1. General considerations

Considerable progress has been made in working with stakeholder groups and organisations to help them understand the potential of LMI for All and begin the process of integration into practice. Practical examples have been provided, through the case studies (see section 3, above) to illustrate the processes that potential users need to use to integrate the LMI for All resource into their practice and maximise its potential benefit. The momentum established in relation to this activity needs to be maintained, through additional stakeholder engagement along similar lines.

5.2.2. What works?

Clear insights have been gained into what types of engagement activities work for particular outcomes. A future stakeholder strategy would benefit from building on the learning derived from this phase of the project, so that metrics can be built up. For example, it has been found that workshops are the most effective dissemination mechanism to collect data on perceived usefulness for potential users.

5.2.3. Overcoming challenges to implementation

Stakeholder activity has indicated that schools currently have the lowest level of awareness (though this has been improved by the availability of the careerometer). A particular challenge for schools is the technical development process. As different organisations have different types of resources available, a differentiated support system could be developed, with particular account taken of resources available.

A further issue to consider is the promotion of potentially useful datasets, like the HES data, which are currently underused. A strategy could be developed to identify underused datasets and design activities that would encourage/support take up.

References

- Barnes, S-A., Bimrose, J. and Hughes, D. (2015). 'LMI for All': supporting access to careers labour market information for schools, *SSAT Journal* (Winter).
- 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., Barnes, S-A., Owen, D., Li, Y., Green, A., Bosworth, L., Millar, P., Holden, A., Attwell, G., Rustemeier, P., Elferink, R. and Higginbottom, J. (2015). *LMI for All: Developing a Careers Database (Final report)*. London: UK Commission for Employment and Skills.
- 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>
- European Commission (2013). ESCO European classification of skills, competences, qualifications and occupations (the first public release), A Europe 2020 initiative. Luxembourg: Publications Office of the European Union. Available from: <https://ec.europa.eu/esco/portal/document/en/8e9cf30d-9799-4f95-ae29-e05c725b24c7>
- 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
- Li, Y and R A Wilson (2016) *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-output 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>
- Wilson, R.A., Sofroniou, N., Beaven, R., May-Gillings, M., Perkins, S. and Lee, M. (2016a). *Working Futures 2014-2024: Main Report*. UK Commission for Employment and Skills: Wath on Dearne.
- Wilson, R.A., May-Gillings, M., Pirie, J. and Beaven, R. (2016b). *Working Futures 2014-2024: Technical Report*. UK Commission for Employment and Skills: Wath on Dearne.

Annex A: Users of Careerometer

East

Moving On Careers and Qualifications (careers magazine), Saffron Walden
<http://movingonmagazine.co.uk/careerometer/>

East Midlands

Vision West Nottinghamshire College, Mansfield
<http://iag.wnc.ac.uk/4098-2/>

Ducere

<http://ducere.org.uk>

London

Connexions Ealing at Northolt High School, West Ealing, Greater London
<http://northolthighconnexions.weebly.com/lmi-and-other-resources.html>

The College of Haringey, Enfield and North East London
<http://www.conel.ac.uk/courses.html>

Spark! Kick Starting Careers, Hounslow

<http://sparklondon.org/students-and-parents/advice/careerometer/>

Central London Careers Hub (CLCH)

<http://www.centrallondoncareershub.co.uk/stuff/lmi-for-all>

North East

The King Edward VI School, Morpeth
<http://www.kevi.info/careerometer/>

Kenton Multi Academy Trust, Newcastle upon Tyne

<http://www.kenton.newcastle.sch.uk/careerometer.php>

Hartlepool College of Further Education, Hartlepool

<http://www.knowyourcollege.co.uk/contact/4575415599>

Greenfields School (on their moodle), Newton Aycliffe

<http://www.greenfieldschool.net/>

Teeside University, Middlesbrough

http://www.tees.ac.uk/depts/careers/careers_help/occupational_information.cfm

East Durham College, Durham, North East

<http://www.eastdurham.ac.uk/careerometer>

Connexions-TW, Tyne and Wear, North East

<http://www.connexions-tw.co.uk/compare-careers>

North West

Burnage Academy for Boys, Manchester

<http://www.burnage.manchester.sch.uk/index.php/school-life/ceiag/careerometer>

Fazakerley High School, Liverpool

<http://www.fazakerleyhigh.org/careers.php>

University of Salford, Salford

<http://www.careers.salford.ac.uk/page/Labour-Market-Information>

South East

Hillview Sixth Form, Tonbridge, Kent

<http://www.hillview.kent.sch.uk/current-students-2/careerometer/>

Career Guidance Charts, Tunbridge Wells, Kent

<http://www.careerguidancecharts.com/career-comparison.html>

New Line Learning Academy, Maidstone

<http://www.futureschoolstrust.com/New-Line-Learning/Skills-For-Life/Career-Comparison>

FHS Careers

<http://fhscareers.weebly.com/>

Frances Cushway (career coach), Godalming

<http://www.francescushway.com/careerometer>

What's your next move?, Surrey County Council

<http://www.careersadviceinsurrey.co.uk/GeneralPage.aspx?sectionID=4>

South West

Careerpilot, Western Vocational Progression Consortium (comprises seven higher education institutions, collaborative colleges and training provider partners in the northern sub-region of the South West), Bath

<http://www.careerpilot.org.uk/info/a-job-or-career/search-job-sectors/>

Next steps Cornwall, Falmouth

<http://www.nextstepscornwall.co.uk/info-and-resources/careers-advice-and-employment-op/careerometer-widget>

Yorkshire and the Humber

Higher Education Progression Partnership (for the Sheffield City Region), Sheffield

<http://apprentice-forward.org.uk/careers-advice/careerometer>

C&K Careers online, Huddersfield

<http://ckcareersonline.org.uk/uncategorised/93-job-trends>

Redcar and Cleveland Borough Council, Redcar

<http://www.redcar-cleveland.gov.uk/jobs/Careerometer.html>

West Midlands

Aim Higher West Midlands (partnership of Higher Education Institutions, schools, academies and colleges), Birmingham

<http://www.aimhigherwm.ac.uk/for-learners-2/aged-11-to-14/career-planning>

Northern House School, Solihull

<http://northernhousesolihull.schooljotter2.com/news/specific-subject-news/careers>

Ormiston Sandwell, Oldbury

<http://www.ormistonsandwell.org.uk/Students/Careers/>

Northern Ireland

KareerHub

<http://kareerhub.co.uk>

Annex B: Engagement activities

The following is a list of engagement activities undertaken as part of this phase of the LMI for All project:

Scottish and Northumbrian Statisticians Annual Meeting, Durham University

Date: 22/05/15

Lead: Nicholas Sofroniou

Audience: academics working with open data (mathematicians, economists and statisticians)

Activity: presentation of LMI for All including RCU and icould, demonstration of LMI for All api and data modelling

Numbers: 50

Data Science day, Cheltenham Festival of Science, Cheltenham

Date: 22/05/15

Lead: Nicholas Sofroniou

Audience: families and young people

Activity: demonstration of icould website and buzz test as a way to introduce LMI for All

Numbers: 25

Inside Government STEM conference, London

Date: 29/09/15

Lead: Deirdre Hughes

Audience: Higher education lecturers, admissions and senior managers

Activity: conference keynote, leaflets distributed

Numbers: approximately 100

Edge Foundation, London

Date: 13/10/15

Lead: Deirdre Hughes

Audience: UTCs, training providers and school leaders/managers

Activity: expert Panel input

Numbers: audience approximately 100 delegates. Short talking heads video produced

Pathways to Prosperity – Preparing Bradford District for the 21st Century, Bradford City Council, Bradford

Date: 07/10/15

Lead: Deirdre Hughes

Audience: senior leaders from schools and colleges, councils and LEAs

Activity: keynote presentation

Numbers: approximately 100

CDI Annual Conference, Cardiff

Date: 09/11/15 - 10/11/15

Lead: Deirdre Hughes and Sally-Anne Barnes

Audience: higher education, careers professionals, local authorities, RAF, sole traders

Activity: workshop and exhibition

Numbers: 15 participants attended the workshop, 30 plus delegates attended the exhibition stand

London Skills Show, London

Date: 13/11/15

Lead: Deirdre Hughes

Audience: schools, colleges, students, training providers and National Careers Service

Activity: seminar session linking London Ambitions and LMI for All

Numbers: 25 (1,000 plus attendees at main event)

Medway Youth Trust – 16-19 Participation in Learning & Employment: Medway Summit, Chatham

Date: 17/11/15

Lead: Sally-Anne Barnes

Audience: key policymakers, stakeholders including school and college leaders, careers advisers, schools careers co-ordinators, head teachers, counsellors, cabinet ministers, youth training providers

Activity: keynote

Numbers: 60

Skills Development Scotland, Glasgow

Date: 20/11/15

Lead: Deirdre Hughes

Audience: Senior Skills Development Scotland Leadership Team

Activity: review of careers system in Scotland, including use of LMI for All

Numbers: 8

Youth Employment Convention, London

Date: 24/11/15

Lead: Deirdre Hughes

Audience: youth agencies, careers professionals, NEET experts and young people

Activity: expert panel input

Numbers: approximately 100

Lord Mayor of London's Annual Conference, London

Date: 29/11/15

Lead: Deirdre Hughes

Audience: schools, colleges, Councils and LEP members

Activity: expert panel input within workshop

Numbers: 200

Assessment Services Ltd. (matrix quality standard), Nottingham

Date: 30/11/15

Lead: Deirdre Hughes

Audience: Board and Independent Quality and Governance Board members

Activity: update on LMI for All and use of portal by matrix Assessors

Numbers: 15

Northern Ireland Skills Show Planning meeting, Belfast

Date: 02/12/2015

Lead: Deirdre Hughes

Audience: industry and education experts

Activity: planning meeting

Numbers: 12

Westminster Employment Forum: The future for careers education, advice and guidance in England for schools and for adults, London

Date: 03/12/15

Lead: Sally-Anne Barnes

Audience: key policymakers, stakeholders including school and college leaders, careers advisers, universities, representatives of employers and trade unions, academics and others

Activity: seminar presentation 'Career development for adults, improving labour market awareness and new approaches, available online

<http://www.westminsterforumprojects.co.uk/forums/event.php?eid=1038>

Numbers: 100

'Supporting schools and colleges to achieve the career benchmarks', North East Local Enterprise Partnership, Newcastle

Date: 15/01/16

Lead: Sally-Anne Barnes

Audience: School and college head teachers and leads, LMI leads, LEP representatives, career professionals

Activity: presentation, network event

Numbers: 30

Annual Teachers Conference, Oxfordshire County Council, Oxford

Date: 25/02/16

Lead Deirdre Hughes

Audience: mixed, mostly careers co-ordinators, teachers and careers service providers

Activity: keynote

Numbers: approximately 100

Inside Government Conference, London

Date: 03/02/16

Lead Deirdre Hughes

Audience: careers teachers, careers advisers and local authority representatives

Activity: keynote

Numbers: approximately 70

London Councils: London Ambitions' Ambassadors Conference, London

Date: 11/02/16

Lead: Deirdre Hughes

Audience: local authority representatives, careers teachers and careers professionals with representatives from Battersea Power Station

Activity: keynote

Numbers: approximately 60

National Careers Guidance Show, Leicester

Date: 25/02/16

Lead: Jenny Bimrose

Audience: careers professionals

Activity: keynote on using LMI for All as an example in practice

Numbers: 120

National Careers Guidance Show, Leeds

Date: 09/03/16

Lead: Deirdre Hughes and Graham Attwell

Audience: mainly teachers, careers providers and careers professionals

Activity: two keynote presentations focused on LMI developments, including LMI for All update

Numbers: approximately 100

National Careers Service Inspiration Agenda for South Central, CPD Event, Dorset

Date: 22/03/16

Lead: Deirdre Hughes

Audience: school careers co-ordinators and local authority representatives

Activity: LMI for All keynote presentation

Numbers: 50

CDI Student conference, Coventry

Date: 27/03/16

Lead: Sally-Anne Barnes

Audience: student careers professionals and a number of leaders in the careers field

Activity: keynote, LMI for All as offering a possible future for careers LMI

Numbers: 50

West Midlands, National Careers Service – Inspiration Agenda, Birmingham

Date: 08/04/16

Lead: Deirdre Hughes

Audience: NCS, Careers and Enterprise co-ordinator(s) and Job Centre Plus

Activity: update on LMI for All

Numbers: 15

Assessment Services Ltd Conference, Nottingham

Date: 21/04/16

Lead: Deirdre Hughes

Audience: matrix assessors

Activity: keynote, LMI for All update and how the use of LMI contributes to quality standards

Numbers: 70

National Careers Service 'Out and About' Conference, London

Date: 29/04/16

Lead: Jenny Bimrose

Audience: social workers, DWP, SFA, Home Office

Activity: keynote

Numbers: 50

International events

Slovenian National Statistics Agency, Ljubljana, Slovenia

Date: 06/15

Lead: Graham Attwell

Audience: technologists, developers and statisticians

Activity: presentation of LMI for All including demonstration of LMI for All api as part of discussion on medium to long term skill demands

Numbers: approximately 60

Educational Technologists Conference, Italy

Date: 09/07/15

Lead: Graham Attwell

Audience: technologists

Activity: workshop

Numbers: approximately 25

IAEVG – Restructuring Careers Over Unexpected Powerful Forces, Japan

Date: 18/09/15-21/09/15

Lead: Jenny Bimrose

Audience: international audience, careers professionals

Activity: conference presentation, <http://www.iaevgconf2015.jp/program.html>

Numbers: 55

How will ICT developments influence the provision of mobility-related services?, Brussels, Belgium

Date: 19/04/16

Lead: Nick Sofroniou and Sally-Anne Barnes

Audience: EU commissioners, academics and data experts from across the EU

Activity: seminar presentation

Numbers: 19

Cedefop workshop – Labour market information (LMI) in Lifelong guidance (LLG), Thessaloniki, Greece

Date: 16/05/16

Lead: Nick Sofroniou and Jenny Bimrose

Audience: Careers guidance and LMI data experts

Activity: workshop presentation

Numbers: 20

**Department of Employment and Social Development, Canadian federal government
Department symposium on “Learning and Labour Market Information (LLMI)”, Ottawa,
Canada**

Date: 16/05/16

Lead: Deirdre Hughes

Audience: domestic and foreign experts involved in the production and dissemination of
LLMI, senior government officials

Activity: presentation

Numbers: unknown

Annex C: Core data sources included in LMI for All

C.1 Introduction

LMI for All aims to provide detailed data on a range of key labour market indicators for those interested in careers prospects and progression (Bimrose, 2012). These data 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 (2016).

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 C.2 summarises the case for making detailed data on Employment, Pay and Hours available as part of the LMI for All database.
- ❖ Sections C.3 and C.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 C.3 deals with Pay and weekly Hours worked and Section C.4 with employment.
- ❖ Section C.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 C.6 describes the Checking Algorithm used to avoid publishing unreliable estimates of Pay and Hours.
- ❖ Section C.7 provides technical details of the regression analysis undertaken for pay predictions.
- ❖ Section C.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 C.9 concludes by providing a summary of the main data on employment, pay and hours provided in the LMI for All database.

C.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);¹⁶
- ❖ Sector (up to the 2-digit level of SIC2007, about 80 industry categories);
- ❖ Geographical area (12 English regions and constituent countries of the UK);^{17, 18}
- ❖ Gender and employment status (full-time, part-time employees and self-employed).

The main official data sources for such data are:

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

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

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

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

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

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

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

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

A typical earnings function takes the form:

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

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

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

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

$$\ln(E) = a + b \cdot A + c \cdot A^2 + D \cdot X + u$$

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 C.2;
- ❖ D is a vector of parameters associated with the vector X ;
- ❖ a , b and c are also parameters to be estimated;
- ❖ u is the standard regression error term.

X includes:

- ❖ Gender (default is Male (0), a 1 indicates Female);
- ❖ Region (default is London, 11 other 0/1 dummies one for each other region);
- ❖ Sector (default is currently Agriculture, plus 78 other 2-digit SIC2007 categories as used in *Working Futures*)²⁰;
- ❖ Occupation (default is Chief executives and senior officials, plus 368 other 4-digit SOC2010 categories);
- ❖ Qualification (default is a degree or equivalent and 5 other qualification categories²¹ (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 and the present update 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 Data Service (SDS)). These parameters are then used to generate the estimates for the careers database. Table C.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 C.4 below).

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

²¹ Including 'none' and 'don't know'.

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 C.4. This provides, for example, information on the percentage of jobs that are part time.

Table C.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.

C.4 Providing detail without being disclosive – Employment

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

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

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

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

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

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

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

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

C.4.2 The *Working Futures* database

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

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

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.

C.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).

Analyses 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.

²³ Effectively the generation of the *Working Futures* database can be regarded as equivalent to estimating the probability of employment in a certain category defined by: industry (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.

²⁴ For full details of how the *Working Futures* database is constructed see Wilson *et al.* (2016b)

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

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

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.

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

Originally, many of the more detailed data used to construct the *Working Futures* database were only available via NOMIS.²⁷ It was agreed with ONS 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.

The *Working Futures* database has now been redesigned to be dependent only upon data in the public domain which removes 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.

C.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

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

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.

C.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).

C.7 Details of the regression analysis for pay predictions

C.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 (C.7.1) provides a brief introduction. Section C.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 C.7.3 discusses the specification of the earnings functions and how the estimated results are used for predicting pay. Section C.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 C.7.5 compares ASHE and LFS, outlining the advantages and limitations of both datasets. Section C.7.6 concludes this discussion.

C.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 2014 and 2015 are pooled together to form an aggregated sample of 266,974 different individuals covering two years. It is worth noting that the 2014-2015 pooled LFS sample covers four quarters in 2014 and three quarters in 2015 due to the fact that only data of the first three quarters in 2015 were available at the time of this updating work. Since LFS are representative of the UK working population, it is expected that the missing of the

last quarter in 2015 will not affect the pay analysis. For the purpose of this exercise, the pooled sample is further constrained to employees aged 16 and over, leaving 63,640 full-time employees and 21,260 part-time employees for the 2015 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). For the 2015 pay estimation, a pooled sample from the 2013 and 2014 waves of ASHE has to be constructed because the most recent ASHE available to access in the Secure Data Lab was ASHE 2014 at the time of this updating work. The assumption is that the estimated pay based on ASHE 2013 and 2014 is not hugely different from the estimated pay from ASHE 2014 and 2015 if ASHE 2015 was available. Constraining to employees aged 16+, the pooled sample has 243,095 full-time workers and 115,331 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 C.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).²⁸ This process is described in more detail in Section C.7.6.

In order to generate predictions of pay by age in the database supplementary age equations are estimated. These results are then used 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.²⁹

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.

C.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 RQF levels of qualifications, RQF1-8, and no qualification. and 8 dummy variables in the regression (base category: RQF8). 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);

²⁸ RAS is an iterative process used to reconcile row and column totals of a two-dimensional data array with some target figures. See McMenemy et al. (1974), Toh, (1998), Miller and Blair (2009) and Lahr and Mesnard (2004) for a general discussion of RAS methods.

²⁹ API stands for Application programming interface.

- ❖ 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).

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 2014 and 2015 in the LFS estimation and 2013 and 2014 in the ASHE estimation, respectively. It is interacted with industries to control time trend differences within each industry. The base groups are industries in 2014 in the LFS estimation and 2013 in the ASHE estimation.
- ❖ 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 2014 in the LFS estimation and 2013 in the ASHE estimation.

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 RQF8 working in the Agriculture sector and are Chief executives or senior officials in 2014 in the LFS estimation and in 2013 in the ASHE estimation. 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: $EXP(\log \text{ expected wage})$. 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.

C.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'.³⁰ 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 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: RQF Levels 4-8;
- ❖ Medium: RQF 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

³⁰ 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.

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 2014-2015 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.

C.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 the same ratios apply across all other dimensions of the database. The median log pay values are calculated to enable 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).

C.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 2015 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, but bias correction methods are used to reconcile the smoothed model estimates with 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.

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

C.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 C.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.³¹ RAS procedures have been developed to generate detailed data on Pay, Employment and Hours consistent with published data from official sources.

C.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 et al. (2016a and b)). This dataset has been expanded from the 25 2-digit occupations in the WF dataset to 369 4-digit categories for LMI for All database. 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.³² For

³¹ 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).

³² 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.

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).

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 C.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 (2015).³³

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).

³³ Other problems associated with the level of detail being attempted also mean that variation over time is difficult to capture precisely. See the discussion of the "change in pay" indicator elsewhere in this report for further details.

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 C.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 ³⁴) 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 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

³⁴ 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.

numbers (these are used as weights rather than the original sample weights).³⁵

- 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 C.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.
- 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).

³⁵ 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.

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 C.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 C.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 C.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 then 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.³⁶ 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.

C.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’.

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).

³⁶ In practice a simpler method was adopted as described above.

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 the years 2011 onwards 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 2015) 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/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:

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

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 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.

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

C.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 (2015). 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 2014 and 2015 is also provided.
3. **Hours:** the hours database again covers all the main dimensions but just for a single year (2015). 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).

C.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 – 2000-2024;
- ❖ 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 2024. 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 C.2 and C.3).

For **occupations** aggregation could be made to the 3 or 2-digit level of SOC2010.

Table C.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 RQF 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 C.5).

Table C.9.1 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		
6. Non-market services	S	94-96	Other service activities	18-20	68-72
	O	84	Public administration and defence etc		
	P	85	Education		
	Q	86-88	Human health and social work		

Table C.9.2 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 C.9.3 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 C.9.4 Qualifications classified using Regulated Qualifications Framework (RQF)

id	RQF	QCF	NQF (old)	qualification	Categories used in supplementary regression analysis
1	RQF 8	QCF8 Doctorate	NQF 5	Higher degree or equivalent	High
2	RQF 7	QCF7 Other higher degree			
3	RQF 6	QCF6 First degree	NQF 4	Higher education	
4	RQF 5	QCF5 Foundation degree; Nursing; Teaching			
5	RQF 4	QCF4 HE below degree level			
6	RQF 3	QCF3 A level & equivalent	NQF 3	GCE, A-level or equivalent	Medium
7	RQF 2	QCF2 GCSE(A-C) & equivalent	NQF 2	GCSE grades A*-C or equivalent	
8	RQF 1	QCF1 GCSE(below grade C) & equivalent	NQF 1	Other qualifications	
9	No Qualification	No Qualification	NQF 0	No qualification	None

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 C.3;
 - b. Occupations - the 2-digit level as in Table C.4;

Any query relating to Replacement Demands is dealt with analogously to employment³⁷:

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)

C.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 (2015). 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.

³⁷ 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 suppresses 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 C.7.5 for more details.

Annex D: O*NET

D.1 Introduction

Skills data from the US O*NET database was one of the sets of indicators used in the pilot phase (focusing 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 and utilised in the present update.

D.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 service 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.³⁸ 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.³⁹ 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:

³⁸ 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.

³⁹ 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.

D.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 (1,110) 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 (1,110 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 D.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

↓

SOC Code	SOC Title	ONET Code 1	ONET Code 2	ONET Code 3
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	11-3051.03
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	11-9199.09
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	11-3049.00

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 D.2 Alternative steps to improving the matching

<p>The matching of O*NET occupational titles to SOC was refined as follows:</p>
<ol style="list-style-type: none">1. Each US O*NET sub-group was matched to a preferred UK sub-group. Using CASCOT, each O*NET sub-group entry (1,110) was chosen using Google and a search on the O*NET site as help, to the best UK SOC match.<ul style="list-style-type: none">• 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 1,110 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 1,110 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. While the possibility of providing a weighted set of occupations is worthy of further exploration, the future availability of detailed information on skills for occupations via ESCO using the International Standard Classification of Occupations (which will also require mapping to the UK SOC) suggests that the complexity of the API interface to non UK SOC occupational standards be developed judiciously.

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 D.3 to D.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 D.3.

Table D.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 D.4 show details from the Abilities.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 D.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 D.4 Abilities.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 D.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 D.3 and D.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** these 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 come 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 E: Other data considered for inclusion but rejected

E.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:

- E.2 ONS Vacancy Survey
- E.3 Annual Population Survey (APS)
- E.4 NOMIS:
 - Employment (at local level)
 - Claimant unemployment rate
 - Job Centre Plus vacancies (historical series)
- E.5 Census of Population (other indicators)
- E.6 Cedefop – pan-European employment projections
- E.7 Other European datasets
- E.8 Course information

E.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.

E.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.

E.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 unemployment⁴⁰). 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 interface⁴¹.

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.

⁴⁰ 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.

⁴¹ This is a web API implemented using HTTP and REST principles, which is often JSON. The API is hypertext driven.

The indicators that could be derived include:

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

However, access to the data is problematical, because these surveys are collected under the Statistics of Trade Act 1947 which promises to maintain the confidentiality of data provided by survey respondents. Hence all users have to apply for and purchase a 'Notice' from the Department for Business Innovation and Skills in order to use the data. There are proposals (March 2016) for easier access to a less detailed and less disclosive version of 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 fairly up-to-date, it is available for different levels of aggregation and the statistical flags attach identify which data is reliable and indicate the limits of data disaggregation.

Unemployment claimant count

NOMIS provides access to monthly ONS claimant count statistics from June 1971 onwards. The official definition of the unemployment count changes occasionally and is currently the number of people claiming Job Seekers Allowance and National Insurance Credits. The method of collection changed from manual to computerised processing in 1982. Since 1982 monthly or quarterly data on stocks and flows of people claiming unemployment benefit have been produced, disaggregated by age, gender and duration of claim. Since 2005, these statistics have been disaggregated by previous occupation (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.

The claimant unemployment count datasets in NOMIS also classify the stock of unemployed and flows into and out of unemployment by occupation. However, the SOC2000

classification is used and are therefore these data sets are 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.

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.

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. It is a source that 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

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).

This material in this section will need to be re-evaluated given that there is a current ONS proposal (subject to consultation) to relax the access constraints for less detailed and less disclosive data.

E.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 households. 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 has been 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" built up from unit postcodes 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 was only 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 was only completed in mid-2015. The custom datasets required were in actuality only received in late 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 and commissioned tables are subjected to disclosure control, in which any information which might potentially disclose details of identifiable individuals or businesses is suppressed.

Though it is becoming ever more outdated, the Census still provides the most detailed (and up-to-date) 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 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 to 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 and published figures are thus for the entire population. 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 are derived from the Workplace Population and journey-to-work tables, which are always produced toward the end of the publication process (in this case, 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 'Workplace Zones' (a new 'workplace geography' created for the 2011 Census) 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 the larger geographical areas they nest into. 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 (RQF) 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 (RQF) 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 (RQF) 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 journey-to-work 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 that 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 due to 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 E.5.1 Labour market questions in 2011 Census of Population

<p>26 Last week, were you:</p> <ul style="list-style-type: none"> ➤ Tick all that apply ➤ Include any paid work, including casual or temporary work, even if only for one hour <input type="checkbox"/> working as an employee? ➔ Go to 32 <input type="checkbox"/> on a government sponsored training scheme? ➔ Go to 32 <input type="checkbox"/> self-employed or freelance? ➔ Go to 32 <input type="checkbox"/> working paid or unpaid for your own or your family's business? ➔ Go to 32 <input type="checkbox"/> away from work ill, on maternity leave, on holiday or temporarily laid off? ➔ Go to 32 <input type="checkbox"/> doing any other kind of paid work? ➔ Go to 32 <input type="checkbox"/> none of the above 	<p>33 In your main job, are (were) you:</p> <ul style="list-style-type: none"> <input type="checkbox"/> an employee? <input type="checkbox"/> self-employed or freelance without employees? <input type="checkbox"/> self-employed with employees?
<p>27 Were you actively looking for any kind of paid work during the last four weeks?</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>	<p>34 What is (was) your full and specific job title?</p> <ul style="list-style-type: none"> ➤ For example, PRIMARY SCHOOL TEACHER, CAR MECHANIC, DISTRICT NURSE, STRUCTURAL ENGINEER ➤ Do not state your grade or pay band <p><input type="text"/></p> <p><input type="text"/></p>
<p>28 If a job had been available last week, could you have started it within two weeks?</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>	<p>35 Briefly describe what you do (did) in your main job.</p> <p><input type="text"/></p> <p><input type="text"/></p>
<p>29 Last week, were you waiting to start a job already obtained?</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>	<p>36 Do (did) you supervise any employees?</p> <ul style="list-style-type: none"> ➤ Supervision involves overseeing the work of other employees on a day-to-day basis <p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>
<p>30 Last week, were you:</p> <ul style="list-style-type: none"> ➤ Tick all that apply <input type="checkbox"/> retired (whether receiving a pension or not)? <input type="checkbox"/> a student? <input type="checkbox"/> looking after home or family? <input type="checkbox"/> long-term sick or disabled? <input type="checkbox"/> other 	<p>37 At your workplace, what is (was) the main activity of your employer or business?</p> <ul style="list-style-type: none"> ➤ 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 <p><input type="text"/></p> <p><input type="text"/></p> <p><input type="text"/></p>
<p>31 Have you ever worked?</p> <ul style="list-style-type: none"> <input type="checkbox"/> Yes, write in the year that you last worked <input type="text"/> ➔ Go to 32 <input type="checkbox"/> No, have never worked ➔ Go to 43 	<p>38 In your main job, what is (was) the name of the organisation you work (worked) for?</p> <ul style="list-style-type: none"> ➤ If you are (were) self-employed in your own organisation, write in the business name <p><input type="text"/></p> <p><input type="text"/></p> <p><input type="checkbox"/> No organisation, for example, self-employed, freelance, or work (worked) for a private individual</p>
	<p>42 In your main job, how many hours a week (including paid and unpaid overtime) do you usually work?</p> <ul style="list-style-type: none"> <input type="checkbox"/> 15 or less <input type="checkbox"/> 16 - 30 <input type="checkbox"/> 31 - 48 <input type="checkbox"/> 49 or more

Figure E.5.2 Journey-to-work questions in 2011 Census of Population

<p>40 In your main job, what is the address of your workplace?</p> <p>➤ If you work at or from home, on an offshore installation, or have no fixed workplace, tick one of the boxes below</p> <p>➤ If you report to a depot, write in the depot address</p> <p><input type="text"/></p> <p><input type="text"/></p> <p><input type="text"/></p> <p><input type="text"/> Postcode <input type="text"/></p> <p>OR</p> <p><input type="checkbox"/> Mainly work at or from home</p> <p><input type="checkbox"/> Offshore installation</p> <p><input type="checkbox"/> No fixed place</p>	<p>41 How do you usually travel to work?</p> <p>➤ Tick one box only</p> <p>➤ Tick the box for the longest part, by distance, of your usual journey to work</p> <p><input type="checkbox"/> Work mainly at or from home</p> <p><input type="checkbox"/> Underground, metro, light rail, tram</p> <p><input type="checkbox"/> Train</p> <p><input type="checkbox"/> Bus, minibus or coach</p> <p><input type="checkbox"/> Taxi</p> <p><input type="checkbox"/> Motorcycle, scooter or moped</p> <p><input type="checkbox"/> Driving a car or van</p> <p><input type="checkbox"/> Passenger in a car or van</p> <p><input type="checkbox"/> Bicycle</p> <p><input type="checkbox"/> On foot</p> <p><input type="checkbox"/> Other</p>
---	---

E.6 Cedefop database

Cedefop publish a range of skills demand and supply projections that are available in the public domain.⁴² 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.

⁴² See <http://www.cedefop.europa.eu/EN/about-cedefop/projects/forecasting-skill-demand-and-supply/index.aspx>

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;
- ❖ 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. Cedefop now provides country-level forecasts using a methodology similar to Working Futures, but they are constrained to utilise only pan-European data sources. In the absence of updates to Working Futures, these 2-digit occupation projections might be drawn upon, but with the corresponding limitations in resolution and modelling constraints.

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);
- ❖ 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 E.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 E.6.1 Mapping from ISCO08 to SOC2010

1	2	3	4	5

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 E.6.2 Map from ISCO 88 to SOC2010 at 2-digit level

ISCO88 Categories as used in Cedefop Projections	2010		SOC2010 categories as used in <i>Working Futures</i>	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

E.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 anonymised 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 of 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. Data for six rounds of the survey are now available.

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 the 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.

E.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.

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 them 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 and 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 from the qualification requirements for new entrants.