InGRID

Supporting expertise in inclusive growth

www.inclusivegrowth.eu

Milestone 114

WEB-BASED METHODOLOGY FOR MONITORING NEW JOBS

Updating the Occupations Observatory

Miroslav Beblavý, Cécile Welter-Médée & Karolien Lenaerts, with contributions from Mehtap Akgüç, Zachary Kilhoffer & Ana Silva

31 October 2018

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998

Abstract

The identification of new and emerging occupations has proven to be a challenging task, in which real-time information on labour market developments is key. At present, the most commonly used data sources do not provide up-to-date information, are narrow in scope or limited in size. In this light, online job portals have been suggested as an interesting data source for real-time labour market analysis. This report aims to contribute to the identification of new and emerging occupations by presenting an updated version of the methodology underpinning the Occupations Observatory developed by Beblavý et al. (2016). We use data extracted from online job boards using web scraping techniques, compare newly identified occupations with existing occupational classifications, and present examples of the tasks and skills required. With this update, we set out to further fine-tune the data collection, processing and analysis steps, but also to make the methodology and outputs more user-friendly, while providing more information at the same time. The proposed revised methodology consists of seven stages, and has been tested for the case of Ireland.

This report constitutes Milestone 114, for Work Package 12 of the InGRID-2 project.

October 2018

General contact: inclusive.growth@kuleuven.be
p.a. InGRID
HIVA - Research Institute for Work and Society
Parkstraat 47 box 5300, 3000 LEUVEN, Belgium

For more information miroslav.beblavy@ceps.eu

Please refer to this publication as follows:
Beblavý, M., Welter-Médée, C., & Lenaerts, K. with contributions from Akgüç, M., Kilhoffer, Z., & Silva, A. (2018). *Web-based methodology for monitoring new jobs, Updating the Occupations Observatory* (Milestone 114). Leuven, InGRID-2 project 730998 – H2020.

This publication is also available via http://www.inclusivegrowth.eu

This publication is part of the InGRID-2 project, this project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998.

InGRID
Supporting expertise in inclusive growth

Contents

1. Introduction

1.1 Identifying new and emerging occupations

As *labour market transformations* are giving rise to *new occupations and skills*, the identification of such new occupations and skills is important to prevent *skill gaps and mismatches*, contribute to a *well-functioning labour market,* and ensure the *competitiveness of Europe's economy,* and *to support policy-making*. While the need to have up-to-date information on new occupations and skills has been widely acknowledged, only little information is available to date.

Traditionally, new occupations - and by extension, new skills - are identified on the basis of trade publications, surveys, data from job advertisements, employer interviews and existing occupational classifications (Beblavý et al., 2016). These methods, however, tend to rely on data that are updated irregularly or are simply outdated, focused on a specific case or sector, or derived from the opinion of an expert or stakeholder and, therefore, hard to validate or generalise.

To address these issues, recent data and a methodology that allows for quick and easy identification are needed. *Real-time labour market information* obtained from *online data sources* would be an excellent candidate for this purpose. Previous research has suggested that the Internet has become a prominent data source for research, notably for labour market analysis (Askitas & Zimmermann, 2009; 2015; and D'Amuri & Marcucci, 2010; Benfield & Szelmko, 2016). The Internet has, indeed, radically changed job search, recruitment and matching (Carnevale et al., 2014; Kuhn, 2014; Kuhn & Mansour, 2014). In this context, a number of authors have used data obtained from *online job boards to explore questions on jobs and skills*. In this analysis, we will follow a similar strategy. Online job boards are a particularly interesting and rich data source, as they bring together *detailed information on jobs, skills and tasks across different sectors,* which is *updated* on a very regular basis and allows to capture recent trends. Most job boards have in fact developed into career boards that also provide other information such as CVs, sector analysis, company reviews, and more.

1.2 The 'Occupations Observatory'

Against this background, Beblavý et al. (2016) developed a methodology that makes use of real-time labour market information to identify new and emerging occupations. This methodology was at the core of their *'Occupations Observatory'*, which set out to provide up-to-date information on labour market developments to policy-makers, academics, job seekers, business owners and other relevant stakeholders in a format that was easily accessible, yet sufficiently detailed (labelled 'occupation card'). The Occupations Observatory not only describes occupations that did not exist before, but also new occupations in terms of awareness on their existence and importance in a specific country or sector. The methodology was piloted for 11 countries: Belgium, the Czech Republic, Denmark, France, Germany, Hungary, Italy, the Netherlands, Poland, Slovakia, Spain and the UK.

Methodologically, the Occupations Observatory relies on data obtained from online job boards - one job board for each of the countries listed above - through web scraping. One of the innovative features of the Occupations Observatory was that the methodology made use of the *underlying occupational classification* that the online job boards themselves use to structure their website and vacancies. This classification can be established on the basis of a *'tag'* system (when tags are attached to each vacancy and used to structure the website) or keywords (directly extracted from the vacancy text, but this is less precise). Research showed, however, that most commonly, a tag system was used. This

method is faster and less data-intensive than extracting full vacancies texts, but consequently generates less information on the potentially new occupation.

In practice, first a *benchmark* list of occupations available on the job board was established, after which a new list was extracted a month later and compared to the benchmark. Any occupation that was not included in the benchmark was considered a candidate for a potentially new occupation, and subject to further examination. Occupations that could already be found on the benchmark list were removed; occupations that were not yet part of the benchmark were added to it; and then the process was repeated.

1.3 Updating the 'Occupations Observatory'

After piloting the Occupations Observatory methodology in the different countries over a six-month period, Beblavý et al. (2016) concluded that the pilot was a successful proof of concept. The feasibility of the approach to identify new and emerging occupations was confirmed. Nevertheless, there were a *number of lessons to be learned*, which the current project aims to address:

- Although the occupational classification of a job portal is easy to obtain and this does not take much time in terms of data collection, processing and analysis, some portals seemed to update the classification much more regularly than others. It was not entirely clear why this was the case. This also meant that for some countries there was a lot of volatility in the data; while for other countries there was hardly any variation. A related point was that monthly periodicity of extracting the occupational classification from the job boards was too frequent in the majority of the cases.
- While some steps of the methodology were automated, the entire process was rather labour intensive and involved a lot of manual checking by researchers with different expertise (e.g. knowledge of a country and its labour market, knowledge of a specific language - also see the next points).
- As the occupational classification itself does not provide information on the occupation's skills and tasks, it was necessary to go back to the vacancy text to get information about the potentially new occupation. This step of the methodology was not automated, and required a researcher to go to the portal and attempt to find the vacancy that triggered the new occupation. This was not always an easy task, especially when some time had passed between extracting the classification and the analysis of results. It is also in this light that Beblavý et al. (2016) recommended to combine different approaches.
- The output of the methodology 'occupation cards' provided a lot of information, but took a lot of time to compile (very labour-intensive), and some of the information is repeated across the card. Both the user-friendliness and the process of preparing such a card should therefore be re-assessed. Other ways to provide the results are worthy to explore.
- Issues related to translations - see below.

In order to overcome these issues and further improve the functionality of the Occupations Observatory, the methodology and presentation of the output were completely revised in this update. This was a crucial first step, before expanding the methodology to additional cases. In this revised version, we go back to the vacancies by extracting job titles, and then use job titles to identify occupations. These occupations are then compared to existing occupational classifications to verify whether they are new. The idea is not to generate a comprehensive list of new occupations, but rather to develop a methodology that allows for an easy and fast identification of new occupations using an innovative data source.

1.4 Selecting a test case

To test the updated methodology in practice, the Irish job portal *Indeed* (https://ie.indeed.com/) was selected. This choice was motivated by two main reasons. First, although the pilot study of Beblavý

et al. (2016) considered 11 EU Member States, the *translation process* significantly complicated the analysis. We, therefore, want to avoid such issues when testing the revised methodology. In the pilot, information was extracted from a job portal in the native language, translated via Google Translate and cross-checked by a native speaker. Whereas using Google Translate worked quite well in some cases, in other cases there were many mistakes (e.g. incorrect translation due to different nuance of a word, or words that were skipped for various reasons). It would, therefore, be necessary to manually check the quality of translations, but this is very time-consuming and complex. Another reason for choosing Ireland is that all occupational classifications are available in English in their most detailed format. ISCO-08, for example, is available in English, French and Spanish, but the most detailed version - which holds over 7,000 occupations - only exists in English. Similarly, O*NET is only available in English. Finally, the choice for Ireland, rather than another English-speaking country, is motivated by the size of the country and its labour force, the dynamism of its labour market, the use of the Internet and availability of IT infrastructure, the relevance of online portals in job search and recruitment, and other variables. *Indeed* is one of the largest and most popular portals in the country.

2. Methodology

In what follows, a detailed description is provided of the revised methodology of the 'Occupations Observatory'. This methodology comprises seven stages and encompasses data collection, processing and analysis. In terms of the methodology's main contributions, besides an easy and fast identification of potentially new and emerging occupations, it can be that the methodology generates a longlist of thousands of current occupations and that potentially new occupations are assessed against the most detailed occupational classifications, as well as against the background lists of occupations used to construct these classifications.

2.1 Stage I – Scraping jobs titles from an online portal identified as (quasi) exhaustive

Stage I consists of *data collection*. Data are collected by scraping job titles from the Irish *Indeed* job portal[1] using web scraping techniques. *Indeed* is one of the most widely used job portals in Ireland that gathers vacancies published on job sites, newspapers, associations, company career pages and other sources for positions of various skill levels, job types, sectors, and locations. *Indeed*'s combined vacancies can, therefore, be considered quasi-exhaustive.

To extract the job titles from the *Indeed* job portal, a Python programme is developed (programme 0, on which more information is provided in the box below). This programme, as well as the other two Python programmes which are described in more detail below, are written in Python Notebooks (corresponds to Jupyter compiler, from the Anaconda suite). Each programme starts by calling the useful packages. All the Excel files (inputs and output files) need to be saved in the same folders as the Python programmes. Note that only the job titles were extracted and not the full text of the job vacancy, as this would take up a huge amount of time, but not necessarily provide more insights for the purpose of our analysis.

Table 1. Python programme for web scraping

Python Programme 0 - *Scraping of Indeed website.ipynb*	
Aim	This programme is dedicated to the selected website scraping
Input	Detailed url of the website; in the case of *Indeed* Ireland, this url is: https://ie.indeed.com/jobs?q=&l=Ireland&start=0
Output	Extraction_EI_Indeed_MM_DD.xlsx: an MS Excel file containing all scraped job titles, associated to their specific url
Duration	Almost 4 hours to scrap Indeed Ireland, extracting 25,930 vacancies or 81% of total available vacancies (August 2018)
Modifications	- Marginal modifications needed to adapt the programme to other *Indeed* websites or rename the output file - Potentially major modification needed to adapt the programme to other job board websites

When scraping the *Indeed* website, some technical issues emerged that had to be resolved. The *Indeed* website stops loading job offers at 1,000 vacancies. The website does not allow its users to manipulate

1 https://ie.indeed.com/jobs?l=Ireland&start=0

the selection criteria so that more than 1,000 vacancies are shown. To overcome this issue, the Python programme relies on the list of available job types as well as on a list of keywords that are entered in the website's search box, and extracts the corresponding list of job titles that the website returns. The resulting lists of job titles are compared and duplicates are removed. For the moment, extractions are not exhaustive; the last extraction gathered job titles from 25,930 vacancies (81% of the database of vacancies for Ireland).

Figure 1. Example of output of Stage I

	ident	Job
2	/company/PHMR/jobs/Medical-Statistician-c57724f7576f5ad8?fccid=fa4f72764ed4e85a&vjs=3	Medical Statistician
3	/rc/clk?jk=f5952122edb00f6a&fccid=5e964c4afc56b180&vjs=3	Corporate Responsibility and Wellness Lead
4	/company/Two-Ten-Health/jobs/Business-Analyst-ee4ec9e251816ec9?fccid=5c430a89c7a93ff6&vjs=3	Business Analyst
5	/company/OurTeam/jobs/Facility-Coordinator-b13cb69f87e4f5aa?fccid=a7b96f6f5780c8d5&vjs=3	Facilities Coordinator
6	/company/AXA-Insurance-Ireland/jobs/Business-Performance-Analyst-c578ed1fd9168cf9?fccid=dbf795	Business Performance Analyst
7	/rc/clk?jk=41bcbe2915da3bfb&fccid=729c9ea43d23df82&vjs=3	Medical Workforce Research Officer GVII
8	/rc/clk?jk=13f7d2230240baba&fccid=351e5044b73f6efb&vjs=3	Medical Administrator
9	/company/Chris-Mee-Group/jobs/Hseq-Advisor-6794b5e156e88dd2?fccid=c2a6f9afd466b966&vjs=3	HSEQ Advisor
10	/rc/clk?jk=f0c36a5eacb3adb3&fccid=cd289537629e66e3&vjs=3	Project Manager
11	/rc/clk?cmp=MEIC%2FGlan-Agua-LTD&ti=Hsqe+Advisor&jk=798ef0bb77452e94&fccid=86e9e4962d9dba7	HSQE Advisor
12	/company/Go--Ahead-Dublin/jobs/Office-Support-42ca3a011cdc9e6f?fccid=c85e67783ea7e850&vjs=3	Office Support
13	/company/PHMR/jobs/Senior-Epidemiologist-Real-World-Data-Analyst-f37c7e5b5eb5cf2e?fccid=fa4f72	Senior Epidemiologist/Real World Data Analyst

All the pre-processing tasks described below are implemented on the job titles from the first output described above (i.e. the data collected through the web scrapping, the output of programme 0). Note that these pre-processing steps are done each time the full database of job titles is called (also in later programmes).

The *output* of this stage requires some further pre-processing steps to prepare the data for analysis. More specifically, the following steps are taken as part of the data cleaning process:

- *Convert all titles into lower-case*: jobs titles are transformed into lower case, as this helps remove duplicates later on. For example, when calculating the word count, 'Manager' and 'manager' would be counted as different words if they are not corrected.
- *Removal of punctuation*: in the context of job titles, punctuation is non-informative and can be removed.
- *Removal of stop-words and non-informative words*: stop-words (or commonly occurring words) should be removed from the data because they do not carry specific information. For this purpose, one can either create a list of stop-words ourselves or use predefined libraries. Here, a combination of a predefined library and an ad-hoc list (called 'negative list') of non-informative words is used. The latter list includes city or company names and languages, for example (e.g. Dublin, Cork, Galway, English, Irish, etc.). Such words are tracked in a separate document.
- *Replacement of the plural words by singular words if possible*: for this purpose, we use lemmatisation to convert a word into its root word. This process is based on a vocabulary and does a morphological analysis to obtain the root word. Thanks to the condensed version of the occupations titles, lemmatisation extracts all the singular words from the titles. This step is also done for the O*NET data later on, because all its occupations are plural.

2.2 Stage II – Words frequency analysis

Once the vacancies' titles are extracted from the job portal, they are split into words, after which all words are sorted according to their frequency through the whole database.

Table 2. Python programme for words frequency analysis

Python programme 1 - *Words Frequency Analysis.ipynb*	
Aim	This programme pre-processes the job titles and conducts a words *frequency analysis* on the cleaned data
Inputs	- Output file programme 0: *Extraction_EI_Indeed_MM_DD.xlsx* - Negative list containing words that are to be removed from the analysis: *Negative_list.xlsx* (as an input only for pre-processing job titles)
Output	*Analysis_EI_Indeed.xlsx*: An MS Excel file containing the list of words with their frequency and distribution. Based on the close to 25,000 vacancies, a list of 6,950 words (after removing stop words and words from the negative list) can be compiled
Duration	A few seconds
Modifications	- Marginal modifications needed to rename the input and/or output files - Marginal modifications needed to change the language of the stop-words collected from the adapted Python package

The words in this list with the higher frequencies can be considered as non-informative, as these words are common and therefore not likely to signal new or emerging occupations. Furthermore, the words with the highest frequencies are generic and do not appear to be linked to any specific sector or occupation. Examples include 'manager', 'assistant', 'sales', 'engineer', 'agent', or 'commercial'.

Moreover, in order to support users in choosing what keywords to analyse, all words that appear five times or less, are removed from the output. There are two reasons for doing so: first, these words are insufficiently informative to keep them in the sample. For example, among the words with the lowest frequencies, there are many that are misspelled. Second, although these words on themselves are rare (e.g. 3,699 of these words are only mentioned once, and 904 are only mentioned twice in the 25,930 job titles examined), there are quite of few of them which makes it more difficult for users to get a clear overview of the output and correctly interpret it.

After removing all the words appearing less than five times, 1,418 words from the 6,950 list of words extracted from the job titles remain. Examples of words that are removed include 'scooter', 'английских' (which is obviously a mistake), 'gastrointestinal', 'simulator', 'translation', 'physics' and 'college'. To further guide the user's choice, words are broken down into five sheets according to their frequency of appearance, as illustrated below.

Figure 2. Breakdown of the MS Excel sheets according to frequency of appearance

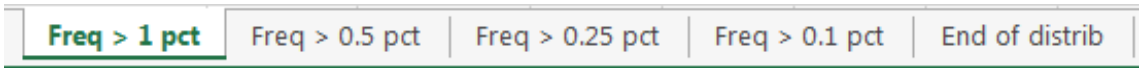


Figure 3. Example of output of Stage II: Words with the highest frequencies (extracted from vacancy titles)

	Words	Frequence	Percentage
2	manager	3325	47%
3	engineer	2473	35%
4	senior	1901	27%
5	assistant	1562	22%
6	sales	1239	18%
7	analyst	1122	16%
8	administrator	857	12%
9	specialist	852	12%
10	executive	739	11%
11	chef	715	10%
12	project	705	10%
13	support	687	10%
14	business	664	9%
15	contract	635	9%

2.3 Stage III – Identification of keywords of interest

Building on Stage II, the idea is not to further investigate the most frequent words but rather look at the *words in the tail of the distribution*. Note, however, that not all segments of the tail are equally informative. The words with the lowest frequencies, for example, are not informative at all, as these words are very rare (e.g. 3,699 of these words are only mentioned once, and 904 are only mentioned twice in the 25,930 job titles examined) and many of them are spelling mistakes rather than words with an actual meaning. Such words are clearly not helpful to capture new and emerging occupations. The challenge is to identify that part of the distribution in which the keywords are sufficiently frequent so they could identify new or emerging occupations but rare enough to correspond to such occupations.

As a starting point, words with a frequency between *0.5% - 0.9%* are kept. In total, this selection contains 168 words, of which 12 words are retained for further analysis. The selection of the keywords that are most likely to indicate new or emerging occupations is not automated, but should be done by a researcher or expert with knowledge on the national labour market and its recent developments. In doing so, the expert should rely on literature and other materials for the country of interest as well as other countries. While this is discretionary step, it simply cannot be automated as occupations have clear *time and space dimensions*; in other words, an occupation that would be regarded as new in a specific sector or country may have existed for many years elsewhere.

Each of the 12 keywords is linked back to the *job titles* that had been obtained through the web scraping in the first stage in Stage IV. For that reason, it is recommended that no more than 30 words are selected for additional analysis. Considering that each of keyword is associated to almost 40 to 60 jobs titles, users will have to look at (40 to 60) x 30 = 1,200 to 1,800 job titles to check as a starting point for proposals of new occupations. By linking the 12 keywords back to their corresponding job titles and vacancies, potentially new occupations can be derived.

Users have to create manually and save this list of keywords in an Excel file *Word_list.xlsx*, which is called as an input in programme 2 (see Figure 4).

Figure 4. Example of output of Stage III: keywords selected for further analysis

	Top
2	Research
3	Network
4	Chain
5	Cloud
6	Technology
7	Media
8	Online
9	SAP
10	EMEA
11	technologist
12	Web

2.4 Stage IV – Selection of job titles containing these keywords

Once the list of keywords for further examination is prepared in Stage III, all jobs titles containing these words are automatically selected from the first sample of jobs titles (collected in Stage I of the methodology). From these job titles, potentially new occupations can be derived (shortlist).

Table 3. Python programme to identify new occupations

Python Programme 2 - *Identification of new occupations.ipynb*	
Aim	All jobs titles from the sample of jobs titles collected in Stage I are compared with the keywords selected in Stage III. Jobs titles are selected if they contain at least one of these keywords and associated to the corresponding keywords
Inputs	- A file containing the keywords, selected in Stage III, from the previous programme's output: *Word_list.xlsx* - The output from Python programme 0 that holds all scraped vacancies' job titles: *Extraction_EI_Indeed_MM_DD.xls*
Output	None, output file is generated after the following stages
Duration	A few minutes
Modifications	Marginal modification needed to rename the input and/or output files

2.5 Stage V – Check jobs titles against occupations listed in ISCO, ESCO and ONET

Then, all the selected job titles are compared to the occupational classifications from ISCO, ESCO and O*NET, on which more details are provided in Table 4. The output of Stage V of the methodology is, as illustrated below, a list with potentially new and emerging occupations.

More specifically, each entry in the shortlist generated in Stage IV is matched to the full list of occupations in the ISCO, ESCO and O*NET classifications. After the matching, three groups can be separated: (i) *perfect match,* i.e. occupations for which a perfect match is found in ISCO, ESCO or O*NET (e.g. occupation 'research director' is a perfect match to ISCO's 'director, research'), (ii) *partial match*, i.e. occupations that contain an entire occupation title from ISCO, ESCO or O*NET but do not match perfectly (e.g. occupation 'health research director' is a partial match with ISCO's 'director, research'), and (iii) *no match*; i.e. occupations that do not contain a full occupational title form ISCO, ESCO or O*NET.

For example, if 'Research Director' is in the shortlist, then it is associated to 'Director, research' which is part of the ISCO classification. If 'Health Information Research Officer' is in the list, it will also be associated to 'Officer, information', which is also part of ISCO. The output of Stage V of the

methodology is, as illustrated below, a list with jobs titles associated to our keywords and their correspondence with existing classifications if they exist.

Occupations that correspond to an entire occupational title from ISCO, ESCO or O*NET ('perfect match') should be eliminated from further analysis, as these occupations, by definition, are not new. The other two types of occupations, 'partial match' and 'no match', are potentially new occupations that should be further examined by the researcher. It is important to keep track of occupations that are partial matches to the classification, because research shows that occupations tend to develop out of existing occupations or combinations thereof. For example, while 'information officer' might not be new, 'health information research officer' can very well be a new occupation.

There are some caveats that users need to be aware of, though. For efficiency reasons, the *process* stops as soon as a first perfect or partial match is found within each classification - this should not be problematic, given that the main result from the comparison is that the occupation seems to exist and in any case partial matches are cross-checked by a researcher in a following step. For example, the *Indeed* occupation 'Research and Development Programme Manager, includes ISCO occupation 'Manager, programme: research'. After this association is found, the matching process is stopped for this specific occupation, even if further in the ISCO classification it appears another ISCO occupation included in it: 'Manager, research and development'.

Table 4. The ISCO, ESCO and O*NET classifications

Classification	Description
ISCO	The International Standard Classification of Occupations (ISCO), prepared by the ILO, is one of the main international occupational classifications. The first version of ISCO (ISCO-58) was adopted in 1957 by the Ninth International Conference of Labour Statisticians (ICLS). It was then replaced by ISCO-68 adopted in 1966, and ISCO-88 in 1987. ISCO has last been updated in 2008, to take into account developments in the labour market since 1988 and to make improvements in light of experiences gained with ISCO-88. The 2008 update did not change the basic principles and top structure of ISCO-88 but significant structural changes were made in some areas. Many countries are now updating their national classification better match it with ISCO-08. ISCO recognises 10 major occupational groups: 1 Managers, 2 Professionals, 3 Technicians and Associate Professionals, 4 Clerical Support Workers, 5 Services and Sales Workers, 6 Skilled Agricultural, Forestry and Fishery Workers, 7 Craft and Related Trades Workers, 8 Plant and Machine Operators and Assemblers, 9 Elementary Occupations, and 0 Armed Forces Occupations. In 2018, 10 years after the ISCO classification was last updated, it is already clear that the ISCO classification is not fully accurate anymore. Simple checks show that some sectors of activity that are very popular currently are not included in this list (no title containing 'online' for example).
ESCO	ESCO stands for European Skills, Competences, Qualifications and Occupations. It is a classification developed by European Commission, under the direction of DG Employment, Social Affairs and Inclusion (tasked to manage the continued development and updating of ESCO). The ESCO classification is composed of modules with elements such as occupations, knowledge, skills and competences, qualifications (following the ISCO hierarchy). When combined and interrelated, these modules make up the whole classification. Moreover, ESCO is available in 27 languages, making it an excellent point of comparison for a possible transcript of the proposed methodology in other languages.
O*NET	The Occupational Information Network (O*NET) is specific to the US economy, developed under the sponsorship of the US Department of Labor/Employment and Training Administration (USDOL/ETA). The O*NET database contains hundreds of standardised and occupation-specific descriptors on almost 1,000 occupations covering the entire US economy. The database, which is available to the public on a dedicated website, is continually updated from input by a broad range of workers in each occupation. The O*NET database was initially populated by data collected from occupation analysts; this information is updated by ongoing surveys of each occupation's worker population and occupation experts. This data is incorporated into new versions of the database on an annual schedule, to provide up-to-date information on occupations as they evolve over time. The O*NET taxonomy is based on the Standard Occupational Classification, the O*NET-SOC taxonomy currently includes 974 occupations which currently have, or are scheduled to have, data collected from job incumbents or occupation experts. To keep up with the changing occupational landscape, the taxonomy is periodically revised; the last revision was in 2010.

Table 5. Python programme to identify new occupations

Python Programme 2 - *Identification of new occupations.ipynb*	
Aim	To compare the shortlist of potential occupations created in Stage IV with the occupational classifications from ISCO, ESCO and O*NET
Inputs	- The output from Python programme 0 that holds all scraped vacancies' job titles: *Extraction_EI_Indeed_MM_DD.xls* - Negative list containing words that are to be removed from the analysis: *Negative_list.xlsx* (as an input only for pre-processing job titles) - The most recent occupations from ISCO: *index08-draft.xlsx* - The most recent occupations from ESCO: *occupations_en.xlsx* - The most recent occupations from O*NET: *Occupation Data.xlsx* *Three last inputs must be downloaded from the corresponding websites.*
Output	*output.xlsx*: an MS Excel file containing a sheet per keyword, with all the corresponding job titles, and if it exists, the corresponding occupation(s) from ISCO, ESCO and O*NET
Duration	A few minutes
Modifications	Marginal modification needed to rename the input and/or output files

Figure 5. Example of output of Stage V: job titles and their associated ISCO, ESSCO and ONET classifications

Jobs titles	ISCO Correspondance	ESCO Correspondance	ONET Correspondance
2nd level network support			
bluecoat network engineer			engineer
cctv network design engineer 60k		design engineer	
cisco network engineer			engineer
custody network manager	manager network		manager
developer c network infrastructure			
director global network operations			
director nursing st lukes radiation oncology network	director nursing		
dutch level network support associate			
dutch network support level			
ict senior network analyst	analyst network		

2.6 Stage VI – Identification of potential new occupations

From the last available list of jobs titles, the programme is able to suggest potential new occupations, based on the most frequent bigrams from the selected job titles.

Bigrams are the combination of two words used together. The basic principle behind n-grams is that they capture the language structure, like what letter or word is likely to follow the given one. The longer the n-gram (the higher the n), the more context you have to work with. The optimal length really depends on the application: if n-grams are too short, one may fail to capture important differences. On the other hand, if they are too long, one may fail to capture the 'general knowledge' and only stick to particular cases.

In our particular case, we are looking for new occupations, in that sense that an occupation should describe something general enough to certainly match many different job offers. Occupations in the existing classification are quite short, it is why we decided to focus on bigrams (in the programme 2, monograms, called tokens, and trigrams are also identified, but not treated, they are left in the programme in case a user would like to analyse their frequencies).

Bigrams frequency is analysed for both the entire list of job titles associated to a keyword, and for the partial list of job titles remaining after removing those titles which have any correspondence in the ISCO, ESCO or O*NET classifications (perfect or partial match). As an illustration, for each of the 12 keywords, the five most common bigrams before comparison to the existing occupational

classifications and the five most common bigrams after comparison to existing classifications are put in the output MS Excel file (see Figure 6).

Both are interesting, for example looking at keyword 'chain', the bigram 'supply chain' appears a lot in all job titles (before and also after comparison to existing classifications). It is a collocation, a bigram used as a composed word. Bigrams 'chain manager' and 'manager supply' are dropped from the most frequent bigrams after removing occupations with existing correspondences in ISCO, ESCO or O*NET. After this step, the bigram 'chain planner' appears to be relatively more frequent, whereas 'chain manager' disappears. Note that should there be a bigram that is of particular interest to the researcher but that has a partial match, the researcher can always go back to a previous output of the Python programme; no information is lost.

On this basis, researchers should *have an idea of potential new occupations*, based on similarities and frequencies of the suggested bigrams.

Figure 6. Example of output of Stage VI: most common bigrams

Before comparison

Key word	Bigrams	Frequence
research	('research', 'development')	9
research	('research', 'fellow')	7
research	('senior', 'research')	7
research	('clinical', 'research')	6
research	('research', 'engineer')	6
network	('network', 'engineer')	15
network	('network', 'support')	12
network	('senior', 'network')	10
network	('level', 'network')	8
network	('engineer', 'network')	8
chain	('supply', 'chain')	56
chain	('chain', 'manager')	8
chain	('manager', 'supply')	8
chain	('chain', 'planner')	7
chain	('chain', 'analyst')	5
cloud	('engineer', 'cloud')	13
cloud	('cloud', 'platform')	8
cloud	('watson', 'cloud')	6
cloud	('software', 'engineer')	6
cloud	('marketing', 'cloud')	5

After comparison

Key word	Bigrams	Frequence
research	('research', 'fellow')	7
research	('clinical', 'research')	6
research	('senior', 'research')	6
research	('research', 'engineer')	6
research	('research', 'officer')	5
network	('network', 'engineer')	13
network	('network', 'support')	11
network	('senior', 'network')	9
network	('engineer', 'network')	9
network	('level', 'network')	8
chain	('supply', 'chain')	40
chain	('chain', 'planner')	7
chain	('chain', 'analyst')	5
chain	('analyst', 'supply')	5
chain	('chain', 'supply')	4
cloud	('engineer', 'cloud')	7
cloud	('cloud', 'platform')	7
cloud	('watson', 'cloud')	6
cloud	('senior', 'cloud')	5
cloud	('marketing', 'cloud')	5

Table 6. Python programme to identify new occupations

Python Programme 2 - *Identification of new occupations.ipynb*	
Aim	To find the most frequent bigrams from the selected job titles associated to the keywords
Inputs	- The output from Python programme 0 that holds all scraped vacancies' job titles: *Extraction_EI_Indeed_MM_DD.xls* - Current data frame of the programme
Output	*output.xlsx*: an MS Excel file containing - Two additional sheets, containing the most frequent bigrams for each keywords, before and after comparison to the three classifications
Duration	A few minutes
Modifications	Marginal modification needed to rename the input and/or output files

2.7 Stage VII – Linking a sample of vacancies to each new occupation

As a final step in the methodology, vacancies corresponding to the bigrams that have been identified are retrieved, as well as their link on the *Indeed* website. It is important to do this part of analysis in the continuity of previous steps (i.e. immediately after the identification of the new occupations), to be able to retrieve the corresponding vacancies.[2] These vacancies are used as illustrations of the types of tasks the vacancy entails, and the associated education and skills requirements (see Figure 7).

Then, the researcher should be able to identify potential new occupations, from both bigrams and associated job titles. For example, 'supply chain' and 'chain planner' are both frequent bigrams, and 'supply chain planner' appears in many associated job titles (Figure 8). As a consequence, a potential new occupation could be *supply chain planner*. It is also informative for the researcher to click on the link and see the vacancy details on the website, only for interesting job titles, regarding the potential new occupation he should have found looking at job titles and bigrams.

2 Although an advantage of the *Indeed* website is that information related to a removed vacancy is still available on the website, using the link associated to each vacancy in the output file.

Figure 7. Example of vacancy

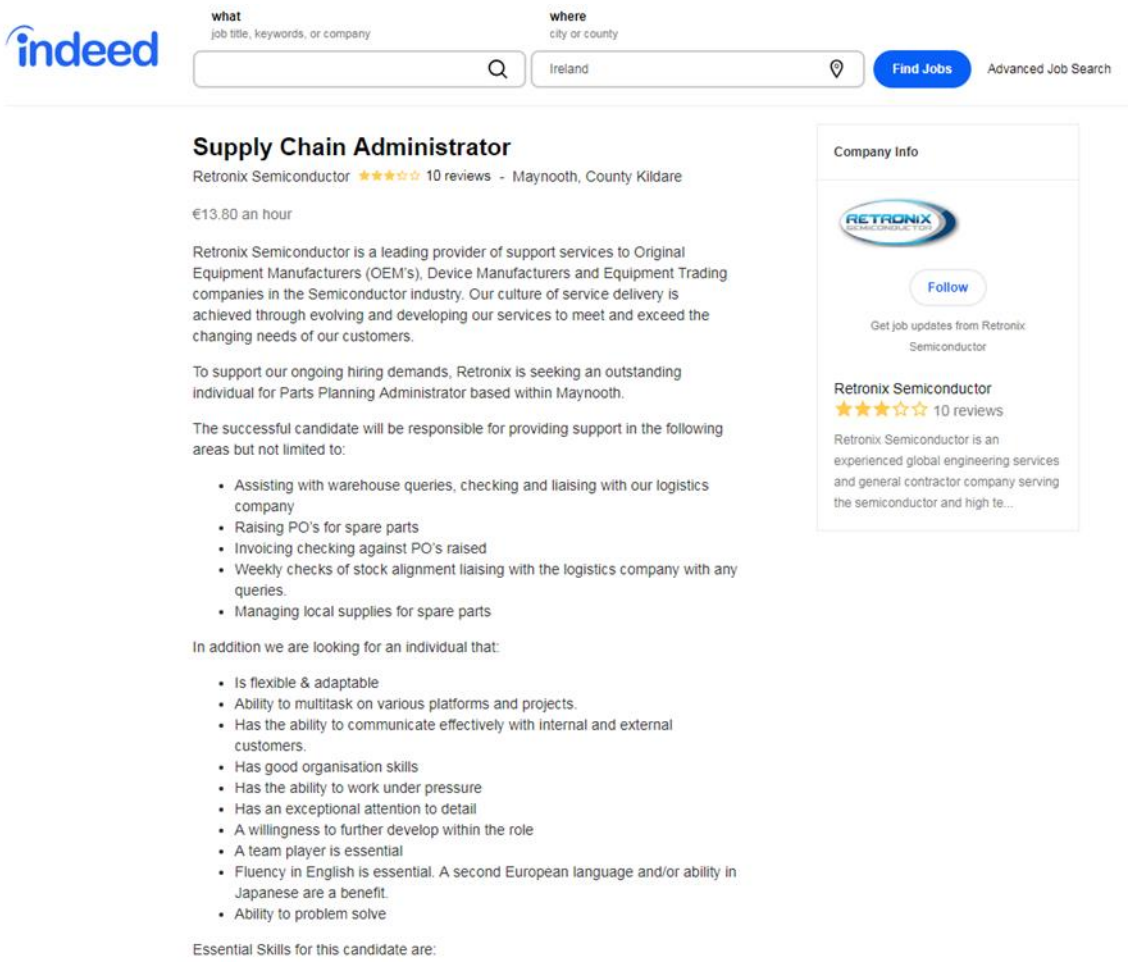


Figure 8. Example of output of Stage VII: list of job titles linked to bigrams

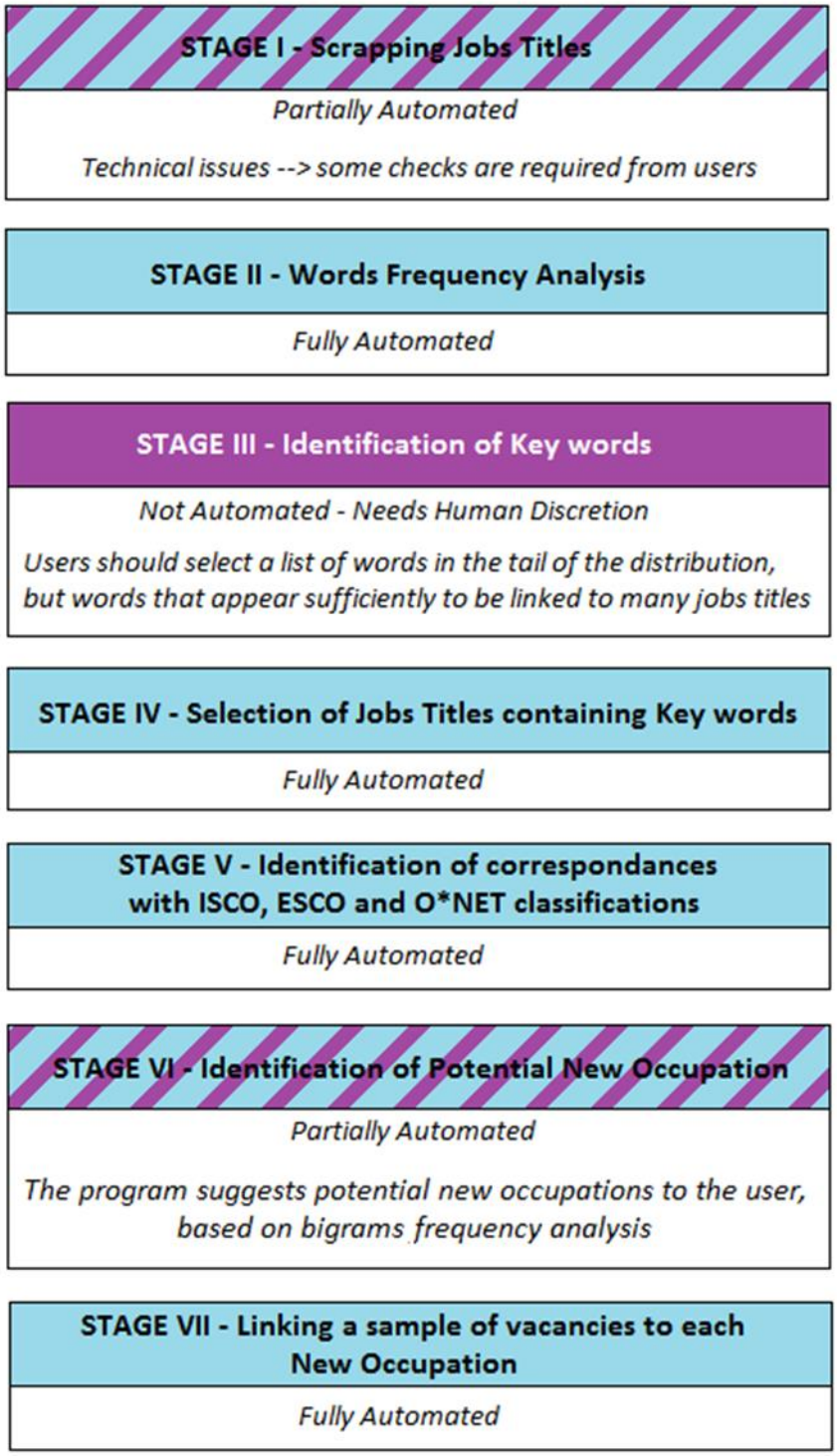
Keyword	Bigrams	Job	Website prefixe	ident
chain	('supply', 'chain')	supply chain specialist iv	https://ie.indeed.com	/rc/clk?jk=0a28d91e2d401878&fccid=e4b075354d7c2865&vjs=3
chain	('chain', 'planner')	senior supply chain planner	https://ie.indeed.com	/rc/clk?jk=c3fd2627dfe0ce8c&fccid=f017d6ee002c4760&vjs=3
chain	('chain', 'planner')	supply chain planner	https://ie.indeed.com	/rc/clk?jk=76183ed1c06d3833&fccid=7942658690eba67c&vjs=3
chain	('chain', 'planner')	supply chain planner	https://ie.indeed.com	/rc/clk?jk=8f078cb5c8d45f1e&fccid=40406c0ac98fe22e&vjs=3
chain	('chain', 'analyst')	supply chain analyst mullingar	https://ie.indeed.com	/rc/clk?jk=636ea264d2413d48&fccid=14b7c83f2c02772e&vjs=3
chain	('chain', 'analyst')	senior supply chain analyst	https://ie.indeed.com	/rc/clk?jk=13a2f8eb850a3bf4&fccid=11f7c022afb890b5&vjs=3

Table 7. Python programme to identify new occupations

Python Programme 2 - *Identification of new occupations.ipynb*	
Aim	To link the vacancies related to the most frequent bigrams found in the previous stage
Inputs	- The output from Python programme 0 that holds all scraped vacancies' job titles: *Extraction_EI_Indeed_MM_DD.xls* - Current data frame of the programme
Output	*output.xlsx*: an MS Excel file containing - A sheet listing the vacancies linked to the bigrams, and their url address
Duration	A few minutes
Modifications	Marginal modification needed to rename the input and/or output files, and to adapt the Website prefix to the selected country

2.8 Summary of the methodological process

Figure 9. Methodological process used to identify new occupations



3. Next steps

3.1 Repetition over time

Because the methodology described above is nearly fully automated, it would be very informative and feasible to repeat it over time for a better understanding of how new occupations emerge. The initial exercise presented here only provides a snapshot of the new and emerging occupations that can be identified at this point in time. Repeating this process every month[3] or quarter for at least a year, however, would make the analysis and monitoring of potential new occupations easier and enable researchers to fully exploit that web data are available in real time. It would also allow researchers to see repetition of 'new' occupations and facilitate the identification of seasonal effects on vacancies. One of the most important next steps, therefore, consists of repeating the analysis over time. This process can be automated, or the three programmes can simply be re-run. Our initial proposal is to repeat the analysis on a quarterly basis, so that the next data collection rounds would take place in the last quarter of 2018 (end of November/beginning of December) and then in the first and second quarter of 2019.

Our initial proposal is to repeat the analysis on a quarterly basis, so that the next data collection rounds would take place in the last quarter of 2018 (end of October/beginning of November) and in the first and second quarter of 2019 (end of January/beginning of February, end of April/beginning of May). At that point, a full year of data collection has been completed (i.e. pilot + three repetitions) and the frequency of the repetitions can be re-assessed.

3.2 Validation of results

Another step that will be elaborated on in the implementation of the methodology over time is that the findings, i.e. the identified new and emerging occupations, will be validated. Among the methods that will be used to do so are: stakeholder feedback (e.g. consulting experts from trade unions, sector representatives, public employment office, academia, job board managers, and others on whether these occupations are in fact new or emerging), cross-check with other data sources (e.g. the country's own occupational classification should this differ from ISCO or ESCO) and publications (e.g. trade publications, government reports), and other validation methods.

To further fine-tune the methodology, a major validation exercise that combines different methods is scheduled to be conducted in the second quarter of 2019, after three repetitions of data collection and analysis. Building on the insights gathered at that stage, the most effective validations methods will be added into the methodology and used in future repetitions. The findings of the methodology in terms of new and emerging occupations will then be contrasted with the results from this exercise.

3.3 Scaling up the methodology

When it comes to scaling up the methodology presented in this document, there are a few options to explore. A first option is to extend the methodology to other countries, including those where English is not among the national languages. Another option is to continue using the Irish case as a baseline, either expanding the work of other job portals or by zooming in on specific sectors. By limiting the

3 Although the pilot study of Beblavý et al. (2016) showed that the frequency should not be too high, this new approach allows researchers to collect much more information, using data sources that are updated more frequently.

scope of the analysis, it is likely easier to subsequently broaden the analysis to other countries, which would complement the analysis of new and emerging occupations. A final decision on this issue will be made in the next stages of the research.

All the difficulty of the transposition of the methodology to other cases, whether they are other job portals in Ireland or other countries, lies in updating the scraping programme and translating key inputs and outputs of the programmes. Looking at the **translation issue** first, transposing the methodology to other countries will require to specify the language in the Python functions which are language-dependant as the function which load the list of stop words (at the beginning of programme 1 and programme 2):

Figure 10. Part of the Python code related to stop-words that needs updating

```
from stop_words import get_stop_words #Upload of stop-words
stop = list(get_stop_words('en'))  #About 900 stop-words
```

It will also be necessary to create an ad-hoc negative list for each country, containing the names of the cities or of big companies in the country. Moreover, except if job titles are available in English and not in the language of the country, it would be necessary to add in the pre-processing of job titles a patch able to translate the job titles into English, in order to compare the job titles to existing classifications, which do not exist in all languages.

Other issues, which have already been mentioned in earlier sections, include the difficulties in using Google Translate, both in practical terms and in terms of the quality of the translations, and the lack of a detailed ISCO classification in languages other than English. Especially the latter is an important issue that would need to be addressed: the ISCO classification is used to check whether a potentially new occupation identified in the programme is in fact new. This is a crucial step in the analysis, and therefore national language ISCO lists would be necessary.

As regards *the Python programmes*, the scraping entirely depends on the website structure. It necessitates some basics knowledge in Python programming and the ability to adapt the existing programme to the potential new website if the structure is different.

By selecting a job portal that operates in multiple countries, such as *Indeed*, part of these difficulties could be avoided if the websites have a highly similar structure. It appears that Indeed websites for many countries have the same structure, but a different url. EU countries with corresponding *Indeed* dedicated websites are listed in Table 8.

Table 8. Exploration of the potential to expand methodology to other *Indeed* websites

Country	url to access all the vacancies for specific-countries
Austria	https://at.indeed.com/jobs?l=Austria&start=0
Belgium	https://be.indeed.com/jobs?l=Belgium&start=0
Czech Republic	https://cz.indeed.com/jobs?l=Czech+Republic&start=0
Denmark	https://dk.indeed.com/jobs?l=Danemark&start=0
Finland	https://www.indeed.fi/jobs?l=Finland&start=0
France	https://www.indeed.fr/emplois?l=France&start=0
Germany	https://de.indeed.com/jobs?l=Germany&start=0
Greece	https://gr.indeed.com/jobs?l=Greece&start=0
Hungary	https://hu.indeed.com/jobs?l=Hungary&start=0
Ireland	https://ie.indeed.com/jobs?q=&l=Ireland&start=0
Italy	https://it.indeed.com/offerte-lavoro?l=Italy&start=0
Luxembourg	https://www.indeed.lu/jobs?l=Luxemburg&start=0
Netherlands	https://www.indeed.nl/vacatures?l=Netherlands&start=0
Poland	https://pl.indeed.com/praca?q=&l=Poland&start=0
Portugal	https://www.indeed.pt/ofertas?l=Portugal&start=0
Romania	https://ro.indeed.com/jobs?l=Romania&start=0
Spain	https://www.indeed.es/ofertas?l=Spain&start=0
Sweden	https://se.indeed.com/jobb?l=Sweden&start=0
United Kingdom	https://www.indeed.co.uk/jobs?l=United+Kingdom&start=0

4. References

Askitas, N., & Zimmermann, K.F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107-120.

Askitas, N., & Zimmermann, K. F. (2015). The Internet as a Data Source for Advancement in Social Sciences. *International Journal of Manpower*, 36(1), 2-12.

Beblavý, M., Akgüç, M., Fabo, B., & Lenaerts, K. (2016). Occupations Observatory – Methodological Note. *CEPS Special Report No. 144*, August 2016.

Benfield, J. A., & W. J. Szlemko (2006). Internet-Based Data Collection: Promises and Realities. *Journal of Research Practice*, 2(2), Article D1.

Carnevale, A.P., Jayasundera, T., & Repnikov, D. (2014). *Understanding Online Job Ads Data: A Technical Report*. Georgetown University, McCourt School on Public Policy, Center on Education and the Workforce, April, 28 pp.

D'Amuri, F., & Marcucci, J. (2010*). 'Google It!' Forecasting the US Unemployment Rate with a Google Job Search Index* (FEEM Working Paper, No. 31).

Kuhn, P. (2014). The Internet as a Labor Matchmaker. IZA World of Labor No. 18.

Kuhn, P., & Mansour, H. (2014). Is Internet Job Search Still Ineffective? *The Economic Journal*, 124(158), 1213-1233.

InGRID-2
Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy

Referring to the increasingly challenging EU2020-ambitions of Inclusive Growth, the objectives of the InGRID-2 project are to advance the integration and innovation of distributed social sciences research infrastructures (RI) on 'poverty, living conditions and social policies' as well as on 'working conditions, vulnerability and labour policies'. InGRID-2 will extend transnational on-site and virtual access, organise mutual learning and discussions of innovations, and improve data services and facilities of comparative research. The focus areas are (a) integrated and harmonised data, (b) links between policy and practice, and (c) indicator-building tools.
Lead users are social scientist involved in comparative research to provide new evidence for European policy innovations. Key science actors and their stakeholders are coupled in the consortium to provide expert services to users of comparative research infrastructures by investing in collaborative efforts to better integrate microdata, identify new ways of collecting data, establish and improve harmonised classification tools, extend available policy databases, optimise statistical quality, and set-up micro-simulation environments and indicator-building tools as important means of valorisation. Helping scientists to enhance their expertise from data to policy is the advanced mission of InGRID-2. A new research portal will be the gateway to this European science infrastructure.

This project is supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998.

More detailed information is available on the website: www.inclusivegrowth.eu

Co-ordinator
Monique Ramioul

KU LEUVEN HIVA
RESEARCH INSTITUTE FOR WORK AND SOCIETY

Partners

TÁRKI Social Research Institute Inc. (HU)
Amsterdam Institute for Advanced Labour Studies – AIAS, University of Amsterdam (NL)
Swedish Institute for Social Research - SOFI, Stockholm University (SE)
Economic and Social Statistics Department, Trier University (DE)
Centre for Demographic Studies – CED, University Autonoma of Barcelona (ES)
Luxembourg Institute of Socio-Economic Research – LISER (LU)
Herman Deleeck Centre for Social Policy – CSB, University of Antwerp (BE)
Institute for Social and Economic Research - ISER, University of Essex (UK)
German Institute for Economic Research – DIW (DE)
Centre for Employment and Work Studies – CEET, National Conservatory of Arts and Crafts (FR)
Centre for European Policy Studies – CEPS (BE)
Department of Economics and Management, University of Pisa (IT)
Department of Social Statistics and Demography – SOTON, University of Southampton (UK)
Luxembourg Income Study – LIS, asbl (LU)
School of Social Sciences, University of Manchester (UK)
Central European Labour Studies Institute – CELSI (SK)
Panteion University of Social and Political Sciences (GR)
Central Institute for Labour Protection – CIOP, National Research Institute (PL)

InGRID-2

Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy Contract N° 730998

For further information about the InGRID-2 project, please contact
inclusive.growth@kuleuven.be
www.inclusivegrowth.eu
p/a HIVA – Research Institute for Work and Society
Parkstraat 47 box 5300
3000 Leuven
Belgium