InGRID

Supporting expertise in inclusive growth

www.inclusivegrowth.eu

Deliverable 12.2

REPORT ON HOW TO IDENTIFY AND COMPARE NEWLY EMERGING OCCUPATIONS AND THEIR SKILL REQUIREMENTS

Zachary Kilhoffer

March 2020

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

Abstract

Identifying new and emerging occupations is important to ensure well-functioning labour markets, yet most data on the topic is quickly outdated. We therefore build on a methodology from the 'Occupations Observatory' (Beblavy et al., 2016), using web scraped data from an Irish job board to identify new occupations, while automating the text analysis process as much as possible.Data were web scraped from the job board Indeed once per month for 18 months, resulting in nearly 500,000 occupation titles. Using several text analysis techniques, we compared these with established occupations from ISCO, ESCO and O*NET databases. We find potentially new and emerging occupations including network engineer, digital marketing executive, network security engineer and cloud architect. Throughout, we encountered theoretical and practical issues. The main practical issue is that occupations posted to job boards are not intended to be objective and used for analysis, but for recruitment. On the theoretical side, it is very difficult to determine when one occupation ends and another begins.We conclude that data from job boards are a promising resource for labour market analysis, but not very well suited for identifying new occupations. Moreover, the analysis described is difficult to automate, requiring contextual and cultural information both in the programming process (i.e. text cleaning), and to adequately interpret results. Future research might use job board data differently, or explore more sophisticated natural language processing techniques than we employ.

This report constitutes Deliverable 12.2, for Work Package 12 of the InGRID-2 project.
It builds on Milestone 114 - Web-based methodology for monitoring new jobs: an earlier report of the InGRID-2 project authored by: Miroslav Beblavý, Cecile Welter-Medee, Karolien Lenaerts, Mehtap Akgüç, Zachary Kilhoffer & Ana Silva

March 2020

© 2020, Brussels – InGRID-2, Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy – project number 730998

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 zachary.kilhoffer@ceps.eu

Please refer to this publication as follows:
Kilhoffer, Z. (2020), Report on how to identify and compare newly emerging occupations and their skill requirements, Deliverable 12.2, Leuven, InGRID-2 project 730998 – H2020

Information may be quoted provided the source is stated accurately and clearly.
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.

The information and views set out in this paper are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Contents

1. Introduction

1.1 Identifying new and emerging occupations

As labour market transformations give rise to new occupations and skills, identifying new occupations and skills is important to prevent skill gaps and mismatches, contribute to a well-functioning labour market, ensure the competitiveness of Europe's economy, and support evidence-based policymaking.

What are occupations? 'Occupation' refers to a set of *jobs* whose main tasks and duties are characterised by a high degree of similarity. Occupations can be used as *job titles*, which is the sense the present paper uses.

-See ILO Standard Classification of Occupations

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 (Beblavỳ et al., 2016). These methods, however, tend to rely on outdated or irregularly updated data, data focused on a specific case or sector, or derived from the opinion of an expert or stakeholder, and therefore difficult to generalise. To address these issues, identifying new occupations using more recent and representative data would be useful.

Real-time labour market data obtained online are an excellent candidate for this purpose using *web scraping techniques*.[1] Previous research has suggested that the internet has become a prominent source of big data for research, notably for labour market analysis (Askitas & Zimmermann, 2015; D'Amuri & Marcucci, 2010; Benfield & Szlemko, 2006). Just as the internet has changed job search, recruitment and matching (Carnevale et al., 2014; Kuhn, 2014; Kuhn & Mansour, 2014), a number of authors have used data obtained from online job boards to explore questions on jobs and skills. Online job boards are a particularly interesting and rich data source to capture recent trends as they bring together detailed information on jobs, skills and tasks across different sectors, and are updated on a very regular basis.

1.2 The 'Occupations Observatory'

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 policymakers, academics, job seekers, business owners and other relevant stakeholders in a format that was easily accessible, yet sufficiently detailed. The Occupations Observatory describes occupations that did not previously exist in a specific country or sector. The methodology was piloted for eleven countries: Belgium, the Czech Republic, Denmark, France, Germany, Hungary, Italy, the Netherlands, Poland, Slovakia, Spain and the UK.

The Occupations Observatory relied 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

[1] Webscraping is a process whereby data is retrieved from websites. It generally refers to automated processes using a 'bot' or 'webcrawler', which is simply a custom-written computer programme.

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 by the website and used to structure it) or keywords (like tags, but deduced by the researcher from the vacancy text).

In practice, a *benchmark* list of occupations available on the job board was established (month one). Once web scraping was complete for month two, it was compared to the benchmark. Any occupation from month two that was not already in the benchmark list was considered a candidate for a new occupation and subject to further examination. Then, the benchmark list was updated with month two's occupations. By repeating this process, new occupation candidates were found each month.

1.3 Updating the 'Occupations Observatory'

After piloting the Occupations Observatory methodology for eleven countries over a six-month period, Beblavý et al. (2016) concluded that the pilot was a successful proof of concept to identify new and emerging occupations. Nevertheless, there were a number of lessons learned and takeaways for future researchers to build on:

- the occupational classification is easy to obtain from job portals, but some portals seemed to update the classification more regularly than others. Thus, for some countries there was hardly any variation between each month's scrapes. In the majority of cases, one month was too small a time interval to observe meaningful changes;
- while some steps of the methodology were automated, the entire process was rather labour intensive. Researchers with different expertises (e.g. knowledge of a country and its labour market, knowledge of a specific language) needed to spend a great deal of time at different stages;
- the occupational classification itself does not provide information on the occupation's skills and tasks, so it was necessary to go back to the vacancy text when one requires more information about an occupation. This step required a researcher to go to the portal and attempt to find the vacancy listing a new occupation. When some time had passed between extracting the classification and the analysis of results, the data may have already been removed from the website. While the entire vacancy text could have been stored, it would have required an impractically large amount of storage capacity and greatly increased the amount of time required for web scraping;
- for all countries except the UK, automated translations were used. These proved insufficiently precise, and manually correcting them proved difficult and time consuming.

In this revised methodology, we attempt to improve how we identify new occupations. This process began with web scraping occupation titles monthly over a year and a half. We then compare the web scraped data to existing occupational classifications and perform text analysis over a number of stages to identify candidates for new and emerging occupations.

1.4 Selecting a test case

To avoid difficulties with translations and better focus on developing the methodology, an English-speaking country was selected. An additional reason for choosing an English-speaking country is that all occupational classifications for identifying existing occupations are available in English in their most detailed format.[2]

Of the EU Member States where English is an official language, *Ireland* was selected. Ireland has a dynamic labour market, high levels of Internet usage and availability of related infrastructure, and a

[2] 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.

booming tech sector. Moreover, its labour market is smaller than the UK's and has fewer major online job portals, which makes it easier to obtain representative data from a single website.

After considering a number of options and lessons learned from the pilot study, the Irish job portal *Indeed*[3] was selected. This portal is very active and frequently updated and contains job postings for all sectors in Ireland. *Indeed* works by compiling job openings web scraped from other job portals, newspapers, association websites, companies' career pages, and so on. Additionally, information is still retrievable even after a vacancy is removed. Finally, because *Indeed* is popular in other countries, it should be relatively easy to adapt this methodology for replication elsewhere.

Previous research found that online job boards are imperfect representations of labour markets, tending towards more technical and higher-skilled occupations (Carnevale et al., 2014). For example, Carnevale et al. (2014) found that between 60% and 70% of job openings are posted online, whereas more than 80% of jobs requiring a Bachelor's degree or higher are posted online. Nevertheless, the internet has become an even more important advertising tool since 2014, and this probably means that job boards host increasingly thorough data on job openings. Furthermore, *Indeed* seems to be the most widely used job board in Ireland, and it hosts ads for occupations of all skills and requirements levels.

[3] https://ie.indeed.com/

2. Methodology

This section details the revised methodology of the 'Occupations Observatory'. This methodology comprises seven stages and encompasses data collection, processing and analysis. The methodology aims to contribute to future research by outlining a relatively fast and easy way to identify candidates for new and emerging occupations. It does so by web scraping occupations on *Indeed* over a period of time, then performing text analysis.

2.1 Stage I – Web scraping occupation titles

Stage I consists of data collection. To collect the data of interest, we web scraped *Indeed* to collect two data points – occupation titles and associated URL. To do so we developed Python Programme 0[4] (see Table 1) to extract the job titles and URLs.[5]

Table 1. Python programme for web scraping

Python Programme 0	
Aim	*Web scraping* – extract job titles and associated URLs from Indeed
Input	*None*
Output	*Extraction_EI_Indeed_MM_DD.xlsx*: an MS Excel file containing all scraped job titles and associated URLs
Run time	Around 4 hours
Modifications	- Marginal modifications needed to adapt the programme to other Indeed websites - Moderate modifications needed to extract more data from each job posting (e.g. employer's name, detailed job description) - Major modifications (essentially a complete rewrite) needed to adapt the programme to other websites

In order to web scrape *Indeed*, a few technical issues had to be resolved. Foremost is that the website is not structured in a way that easily lends itself to extracting all data.[6]

To retrieve all job listings, in principle the programme would work by querying *Indeed's* detailed base URL[7] for all job openings in Ireland, retrieving all job postings on the page, then repeating this process for all subsequent pages until all job listings are exhausted. However, *Indeed* stops loading new job offers after 100 pages.

To overcome this issue, Python Programme 0 does not request all job listings at once, but in batches using keyword pairs. The first set of keywords are occupations or sectors (e.g. 'health', 'marketing', 'manager', and 85 others), while the second set of keywords are contract types ('permanent', 'fulltime', 'contract', 'temporary', 'part-time').

[4] 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).

[5] Only the job title and URL are extracted because our analysis does not use the full text of the job vacancy. While it would have been possible to extract all other text, it would require substantially more time and data storage. Should it be necessary, the researcher can access each listing with the URL at any time. This is a significant advantage of *Indeed* distinguishing it from other job boards.

[6] At the time of writing, ie.indeed.com offers a public API. Therefore, it should now be possible to gather data of interest with API calls rather than web scraping.

[7] https://ie.indeed.com/jobs?q=&l=Ireland&start=0

To illustrate this, we search for 'health permanent', 'health fulltime', 'health contract', 'health temporary', and 'health part-time'. For each of these keyword pairs, we extract occupation titles and hyperlink texts for 1-100 individual pages. We then repeat this procedure for each of the 88 sectors, for a total of 440 base URLs, each of which returns 1-100 pages of occupation titles.

This method results in many duplicates, so we simply drop duplicates before writing the data frame to an Excel sheet output. However, one important weakness of note is that certain keywords combinations still exceed 100 pages of results. Because only the first 100 pages can be retrieved, any listings beyond page 100 are not successfully scraped. Additionally, not every job listing is well-catalogued with keywords. Those that are not, or which do not correspond to our selection of keywords, are also not extracted. Still, by searching for every combination of keywords, our methodology extracts over 80% of vacancies on the website.[8]

We ran Python Programme 0 monthly from September 2018 to January 2020. We were very fortunate that the website structure remained virtually the same over this time period. Otherwise, substantial modifications to the code may have been necessary.

To ease subsequent analysis, all files containing scraped data were merged into a single data frame of 480,044 occupations. An additional column 'date' for the date of retrieval was added to the data as shown in Table 2. Lastly, we removed rows for which the 'Job' field was blank.[9] We move to subsequent analysis with our web scraped data of 479,924 occupations, URLs, and dates of retrieval.

Table 2. Head of merged web scraped data

	Url_suffix	Occupation_title_raw	Date
1	/company/Zevo-H…	Wellness Co-Ordinator	20180913
2	/company/patien…	Patient Engagement Coordinator and Health Research Analyst	20180913
3	/company/PHMR/j…	Medical Statistician	20180913
4	/rc/clk?jk=89e4…	Health Care Assistant	20180913
5	/company/MedLab…	Medical Scientist	20180913
6	/company/PHMR/j…	Health Economics & Real-World Evidence (HE-RWE) Scientist	20180913
7	/company/Shaw-A…	Head of Health & Wellness Education	20180913
8	/company/Meaghe…	Health and Wellbeing Advisor	20180913
9	/rc/clk?jk=5bf4…	General Assistant - Blanchardstown (Part Time)	20180913

Note: '...' indicates a longer text has been cut short for presentation.

2.2 Stage II – Pre-processing

The data requires a significant deal of pre-processing to remove uninformative information surrounding the occupation title. This necessitates automatically processing large amounts of text, ideally with human interaction kept to a minimum. For this purpose we developed Python Programme 1, which makes use of several Python text analysis libraries[10] and custom regular expression replacements.

[8] The programme extracts around 81% of the total job postings. For August 2018, for example, the programme retrieved 25,930 job titles.

[9] Evidently these were posted online by mistake, but in any event they are not useful for our analysis.

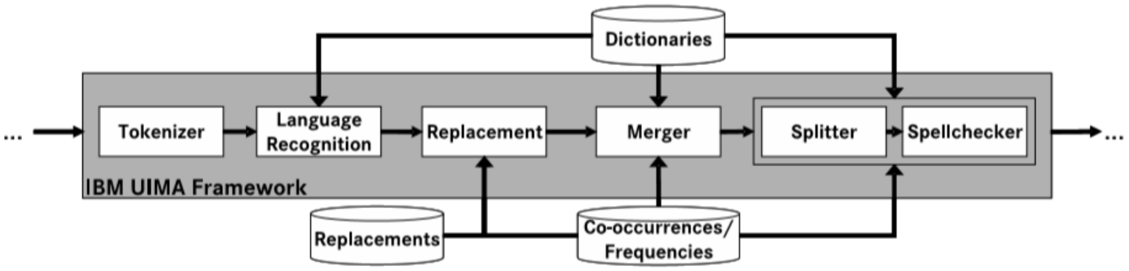
[10] NLTK, stopwords, etc.

Table 3. Python programme for pre-processing and word frequency analysis

Python programme 1	
Aim	Pre-process the occupation titles and conduct a **word frequency analysis**.
Inputs	- Output of Python Programme 0 - Negative list containing words that are to be removed from the analysis
Output	An MS Excel file containing the list of words with their frequency and distribution.
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. - Moderate work needed to create appropriate 'Negative list' of non-informative words.

Certain practices for text cleaning are generally followed prior to text analysis. Schierle et al. (2008) describe a process with many common themes, as shown in Figure 1.

Figure 1. Framework for text preparation



Source Schierle et al. (2008: p. 4)

While we generally followed established guidelines, our particular text of occupation titles required a somewhat different cleaning approach to become useful for our purposes. To briefly highlight a few reasons why, job ads have their own vocabulary (Beblavỳ et al., 2016), which complicates spellchecks with pre-made dictionaries. Job ads are filled with words appearing in no dictionaries – the names of small companies, sector-specific acronyms like 'QCer' (quality checker), 'OTE' (on-target earnings), and many more problematic terms. These and other complications added a great deal of manual work to a process that would ideally be almost entirely automated.

Before beginning text cleaning, we create a duplicate of the column 'occupation_title_raw' column titled 'occupation_title_preprocessed'. This preserves the original data while allowing us to observe the results of processing.

Thereafter, the first step in our text cleaning is to convert all letters to *lower case*. Otherwise 'Manager' and 'manager' would be considered two different words.

Next, we *remove punctuation*, which is generally uninformative in the context of job titles.[11] To ensure we handle hyphenated words in the data (e.g. 'daily-rate'), and to avoid creating mistakes with a few other characters (e.g. '&', '/'), we first replace these characters with a single space. Then, we simply remove all additional punctuation. This intermediate step means a compound word like 'daily-rate' becomes 'daily rate' instead of 'dailyrate', 'male/female' becomes 'male female' instead of 'male-female', etc.

The third step is to *tokenise*. This means that all occupation titles are split by empty space characters into single 'tokens' or words. For example, 'senior manager' becomes a list of two words: ['senior', 'manager'].

[11] Although hyphenated words and some other examples can make this step more complex. See discussion in Schierle et al (2008).

Fourth, we *remove stop-words*: commonly occurring words like 'an', 'and', 'the', and so on. These are removed because they do not carry specific information, while they clutter the data. We rely on a commonly-used, predefined library of English stop-words. The only change made to this predefined list was to remove the word 'it'. For example, if the original title was 'IT Consultant', 'IT' means 'information technology' and should not be removed.[12]

Fifth, we *lemmatise* the tokens. Lemmatisation[13] converts a word into its singular form. This process is based on an existing lexicon and morphological analysis to obtain the root word, converting all the plural words found in occupation titles.[14] Otherwise 'manager' and 'managers' would not be recognised as the same word.

Sixth, we *remove additional non-informative information* appearing in the occupation titles. This step proved to be necessary in our attempt to identify new and emerging occupations, but very complex and arduous to achieve satisfactory results. To illustrate this, consider a few unprocessed occupation titles retrieved on 15 October 2018:

- 25092018 – CX Design and Improvement Manger (11 Month Maternity Cover);
- 3 Full licence Driving Senior Home Care Workers Urgently Needed for North Dublin Areas;
- 30 hr Sale associate (Christmas Temp);
- Construction Operative with SLG 3 Day/SLG 1 Day ticket in New Ross, Co. Wexford - IMMEDIATE START!!

From these it becomes clear that we are dealing with big data. Taking the first example, 'CX Design and Improvement Manager' (presumably not 'manger')[15] is the occupation title we are after. The second example contains a great deal of irrelevant information. We know that three positions are available, a full driver's licence is required, the urgency with which the positions are looking to be filled, and the location. However, just 'Senior Care Worker' appears to be the occupation title.

Common junk information removed includes location names (e.g. Dublin, East Dublin, Cork),[16] company names,[17] timing (e.g. immediate, full-time, weekends, 1 year, 10:00am), level (e.g. entry-level, senior, basic grade), desired language proficiencies, compensation, numbers,[18] and words relating to the search for workers (searching, seeking).

Even beyond these, a great deal more changes were needed to clean the text. For example, many truck driver positions included the weight and length of the truck. We used regular expressions to remove this information, because simply removing words would not suffice.[19] We also changed the data to take certain terms common for an occupation in Ireland; 'nan' and 'nanny' become 'childminder'.

When every appearance of a single word could be safely removed (e.g. Dublin, French, excellent, searching), these words were stored in a separate document, then matching tokens were removed from job titles.[20] When removing an entire word risked changing the meaning of the occupations, we wrote regular expressions. For example, if a job title reads 'full time manager', we cannot remove

[12] Compounds with the word 'it' were removed, however.

[13] Note that 'stemming', or removing the ends of words but retaining the stems, is an alternative approach.

[14] This step is also performed for the O*NET data later on, because all its occupations are plural.

[15] In all 113 times it appears in the data, it seems to have been a misspelling of 'manager'. This is another example of why simple spelling checks referencing a dictionary are insufficient for these data.

[16] The top 100 Irish cities by population were added to the 'Negative List', as well as a few large cities (e.g. Paris, Madrid, London), many countries (e.g. UK, US, Spain), other words that are only indicative of location (e.g. the cardinal directions, 'republic', 'island').

[17] Initially select company names were added to the 'Negative List' as they are non-informative. However, because it is not possible to remove all companies, and removing a few of the largest companies had minimal impact on subsequent steps of the analysis, this strategy was abandoned.

[18] Numbers and letter-number combinations are often the job opening's index used on another website.

[19] For example, one listing read 'Driver – 7.5T'. We used the regular expression '\b\d\.\dt' to identify and remove '7.5T' (after converting to lowercase), but avoid removing other instances where the data maybe more indicative.

[20] This step requires discretion. For example, 'full' might be non-informative in most job titles (e.g. 'full-time developer'), and prove informative in others (e.g. 'full stack developer'). It is better to leave some non-informative words than to remove too many.

‘full’ and ‘time’ from all occupation titles. Otherwise, ‘full stack developer’ would become ‘stack developer’, and useful information is lost. Instead, we needed to remove ‘full time’, ‘fulltime’, etc.

Correcting spellings is another method frequently used in text cleaning. However, this relies (at least initially) on an existing dictionary. Existing dictionaries would not have many of the terms used in job titles (e.g. small villages, small company names or acronyms), so correcting misspellings with our data would be more complicated and labour intensive than usual. Additionally, contextual information is very important to correctly identify and handle misspellings programmatically (Schierle et al., 2008). Occupation titles are short, stand-alone texts which often disregard grammar conventions, and thus contain less context than other corpora. For this reason we do not attempt to identify and correct spellings.

Instead, we remove the least frequent words in the data. This is a rather quick and messy way to address many problems in the data, including misspellings. Following text cleaning as described, the data contain 1,411,653 total words and 14,831 unique. Of the unique words, 88.60% appear fewer than 50 times. These are saved into a separate Excel sheet for later reference. We then duplicate the ‘occupation_title_preprocessed’ column,[21] and in the new column we remove any appearances of the infrequent words. Removed words include noise such as misspellings, words in foreign languages,[22] small village names, obscure acronyms, and reference codes from other websites, but also informative words like ‘statistician’. While some useful information is lost, it is necessary to remove less useful words in bulk. Without more complex text analysis methods,[23] it is not feasible to proceed otherwise.

Lastly, the remaining tokens are reassembled back into occupation titles. The head of results after pre-processing are shown in Tabel 4. In the final row, we see that the misspelling ‘architectural’ is present after the initial processing (in the column occupation_title_processed), but removed in the final column, which contains no infrequent words.

Ultimately, the text cleaning required a great deal of checking and manual work to achieve usable data. The difficulty resulted from a number of factors: occupation titles have their own vocabulary; Ireland has its own vocabulary; and most importantly, *we have not web scraped occupation titles per se, but rather the titles of advertisements on a job board.*

[21] The new column is imaginatively titled ‘occupation_title_preprocessed_no_infrequent’.

[22] Some job openings seem to have been deliberately posted in a non-English language because they require it. Others were probably mistakenly picked up by Indeed’s own web scraping.

[23] Discussed in 5.4 below.

Table 4. Head of pre-processed occupation titles

Url_suffix	Date	Occupation_title_raw	Occupation_title_processed	Occupation_title_processed _no_infrequent
/compa…	20190724	3D Artist	3D artist	3D artist
/rc/cl…	20190319	Senior 3D Artist	3D artist	3D artist
/rc/cl…	20190219	Senior 3D Artist	3D artist	3D artist
/rc/cl…	20190919	3D Artist/Animato…	3D artist animator imaging	3D artist
/rc/cl…	20190821	3D Artist/Animato…	3D artist animator imaging	3D artist
/rc/cl…	20190529	3D Artist/Animato…	3D artist animator imaging	3D artist
/rc/cl…	20190621	3D Artist/Animato…	3D artist animator imaging	3D artist
/rc/cl…	20190724	3D Artist/Animato…	3D artist animator imaging	3D artist
/rc/cl…	20190724	3D Artist/Animato…	3D artist animator imaging	3D artist
/compa…	20200115	Mid 3D Rigging Ar…	3D rigging artist	3D artist
/compa…	20181113	Architetcural Vis…	architectural visualisatio…	3D artist

Note: the index column at far left is not shown.
Note: '...' indicates a longer text has been cut short for presentation.

After preprocessing, some additional occupation titles contain no data. For example, one occupation title simply read '-Ireland',[24] which was removed due to being in the negative list. Considering only non-blank occupation titles, 479,875 remain for analysis.

2.2.1 Word frequency analysis

With pre-processing complete, we proceed with the word frequency analysis. Python Programme 1 breaks each occupation title into individual words, then adds all individual words to a single list. Every unique word is then assigned a frequency for how many times it appears, e.g. (manager: 67,479, engineer: 51,584).

We then calculate each word's frequency relative to the entire corpus by dividing the number of times the word appears by the total word count. In the output file, words are displayed in five sheets according to their frequency of appearance:

- >1%,
- 0.5%-1%
- 0.25%-0.5%
- 0.1%-0.25%
- 0%-0.25%

Python Programme 1's output, sheet one (>1%), is shown in Table 5.

[24] By the author's reckoning, '-Ireland' is not an occupation title. Still, the internet is a place where nearly anybody can post nearly anything. Such 'junk' in big data sets needs to be handled appropriately.

Table 5. Word frequency in job ads - above 1%

	Word	Frequency	Percentage
0	Manager	67,479	4.78
1	Engineer	51,584	3.65
2	Assistant	38,105	2.70
3	Sale	26,053	1.85
4	Analyst	19,170	1.36
5	Service	17,141	1.21
6	Administrator	15,716	1.11
7	Specialist	15,674	1.11
8	Account	15,455	1.09
9	Project	14,610	1.03
10	Support	14,464	1.02

As we would expect, the head of the data contains very general words used in many types of occupations and sectors. At the other end, the data have quite a long tail. If we had not already removed the least frequent words, 98.80% of words (or more precisely, tokens) would appear with less than 0.1% frequency.[25] In the lower frequency, where words appear more than 50 times but account for less than 0.1% of the corpus, many words are quite difficult to interpret alone (e.g. 'keeping', 'submission', and 'optimiser'). Others are quite straightforward, like 'beautician', 'hairstylist' and 'hostesses.

The latter are quite interesting because we know these occupations are still common, so we would expect them to appear more. We can infer that other terms are used more often. Checking this assumption with the data confirms that hundreds of ads exist for 'beauty advisor', 'beauty ambassador', 'beauty consultant' 'beauty therapist', 'beauty specialist', and so on. However, 'beautician' is not currently the preferred nomenclature on Irish job boards.

The words become more interesting between 0.1% and 0.25% in frequency, where words associated with tech occupations appear more often. Here we find terms like 'java', 'web', 'devops', 'cloud', and others likely to be associated with new, emerging, and evolving occupations.

2.3 Stage III – Identifying keywords

The next step is to use the word frequency analysis and desk research to identify keywords that are likely to appear in new or emerging occupations. The selection of the keywords requires some knowledge of the national labour market and its recent developments. The expert should rely on literature and other materials for the country of interest.

As discussed, the most frequent words (e.g. 'manager', 'assistant' and 'sales') are not linked to any specific sector or occupation, less illustrative of the jobs' content, and not likely to signal new or emerging occupations. The words with the lowest frequencies are difficult to interpret, and with very low frequency, less likely to indicate any identifiable trend. Therefore, the challenge is to identify that part of the distribution in which the keywords are sufficiently frequent that they indicate a trend, but rare enough to correspond to something new. As a starting point, the researcher should closely consider words with a frequency between 0.1% - 0.25%.[26]

[25] Although this is in large part because of the text cleaning, which removed hundreds of common words.

[26] The frequency depends in great part on the quality of text cleaning.

For desk research, we made use of national data on growing professions and in-demand skills from the OECD[27] and CEDEFOP.[28] Ultimately the research must make a selection of 10-30 keywords that they expect to correspond with new professions. This list of keywords[29] is then saved to use as an input in Python Programme 2.

The keywords selected for further analysis are as follows: ai,[30] chain, cloud, cyber, digital, mobile, model, network, online, platform, research, tech, and web.[31]

2.4 Stage IV – Create longlist of occupation titles containing keywords

Next, these keywords are used to filter occupation titles obtained through web scraping. All titles containing one or more keywords are retrieved to form a longlist of potentially new and emerging occupations. Of the 479,875 occupation titles, 16,200 contain at least one of the keywords. After dropping duplicates by keyword and occupation title, 4,147 rows remain. If we drop duplicates by only the occupation title, we have 3,962 unique. Thus, some occupation titles match more than one keyword. The two most frequently appearing occupation titles for each keyword are shown in Table 6.

Table 6. Most frequent occupations matching keywords

Keyword	Example occupation 1	n	Example occupation 2	n
Ai	Head data scientist lead innovation ai lab	12	Enterprise business development executiv ...	8
Chain	Supply chain manager	236	Supply chain planner	96
Cloud	Cloud architect	65	Smb sale executive google cloud platform	60
Cyber	Cyber security consultant	29	Cyber security manager	22
Digital	Digital marketing executive	269	Digital marketing specialist	140
Mobile	Mobile phone repair technician	41	Mobile phone technician	40
Model	Snr manufacturing engineer model farm rd	21	Model validation analyst	14
Network	Network engineer	376	Network security engineer	68
Online	Online account manager	51	Strategic partner manager online partner ...	30
Platform	Technical web analyst digital marketing ...	34	Smb sale executive google cloud platform	30
Research	Research assistant	100	Research development engineer	82
Tech	Tech support	42	Tech service engineer	29
Web	Web developer	274	Full stack web developer	57

Note: the occupation titles are pre-processed with the most infrequent words removed.
Note: '...' indicates a longer text has been cut short for presentation.

Some of these occupation titles are immediately recognisable and unlikely to represent anything new, like 'research assistant' (n=100). Others seem like good candidates for new and emerging occupations, like 'cloud architect' (n=65).

Other examples may or may not be new, but are more difficult to interpret, such as 'snr manufacturing engineer model farm rd' (n=21). This occupation title contains a good deal of jargon that was

[27] See OECD's Skills for Jobs Database: https://www.oecdskillsforjobsdatabase.org/imbalances.php#IE/_/_/_/[%22abilities%22]/co. It contains the most in-demand skills, knowledge, and abilities.

[28] See CEDEFOP's Skills Panorama: https://skillspanorama.cedefop.europa.eu/en. For example, it contains national data on sectors that will experience the most employment growth.

[29] We imaginatively name the Excel file *Word_list.xlsx*.

[30] In this case of 'ai', the loop was programmed to search for a stand-alone word rather than testing if the string was in the occupation. Otherwise, 'p*ai*nter' would appear as a positive result. For other words, it is desirable to test for the string rather than a stand-alone word. Thus, '*tech*nologist', '*tech*nology', and 'tech' are all captured by searching for the keyword 'tech'.

[31] Unsurprisingly, we expect new and emerging occupations to be largely found in the tech sector.

not removed in the text cleaning. While this specific occupation title is the most frequent one containing the keyword 'model', it is not very illustrative. This is because a single firm posted the same job opening with very slight variations over many months, increasing its n. However, the second example occupation matching 'model' is 'model validation analyst', which is quite a tidy occupation title and much more generalisable.

2.5 Stage V – Filter longlist using existing occupations in ISCO, ESCO and ONET

Next, we compare each occupation in the longlist generated in Stage IV with the established occupations of ISCO, ESCO and O*NET.

Table 7. Programme to identify candidates for new and emerging occupations

Python Programme 2	
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 - Pre-processing materials (e.g. negative words list, regular expression replacements) used in Programme 1 - The most recent occupations from ISCO - The most recent occupations from ESCO - The most recent occupations from O*NET
Output	An MS Exel file containing a sheet per keyword, with all the corresponding job titles, and if it exists, the partially-matching occupation(s) from ISCO, ESCO and O*NET.
Duration	A few minutes
Modifications	Marginal modification needed to rename the input and/or output files.

For this we wrote a function that compares two lists of words and returns the number of matching elements. If the number of matching elements is equal to the length of the reference list (regardless of elements' order), they are considered a perfect match. In this case we remove the occupation title from consideration because it cannot be new if it also perfectly matches an established occupation.

To create these two lists, each occupation title in the longlist is broken down into a list of tokens. The occupation titles from the ISCO, ESCO, and O*NET data sets are pre-processed in the same way as the occupation titles in Stage II, then each is also broken up into a list of tokens. Then we compare each occupation title against ISCO, ESCO, and O*NET in turn. Thus, if 'Research Director' is in the longlist, then it matches the ISCO occupation 'Director, research', and is removed from consideration as a new or emerging occupation.[32]

From the longlist of 4,147 unique occupation titles containing a keyword, 3,668 (88.45%) exactly match a known occupation from the ISCO classification. 3,516 (84.78%) of the job titles have a correspondence with the ESCO classification, and 2,456 (59.22%) with the O*NET classification. This leaves us with a shortlist of 321 potentially new and emerging occupations.

However, upon looking at the matches more closely, we found a minor issue. Because we applied the same text cleaning to the ISCO, ESCO and O*NET data as we did to the web scraped occupation titles, many reference occupations are only one word. For example, ISCO and ESCO contain reference occupations simply reading 'manager', 'engineer', and 'developer', which are general terms that apply to many occupations. If we had found a bleeding-edge occupation titled 'space-time recalibration manager', it would match 'manager' and be disregarded as a candidate for new or emerging occupations. Therefore this matching strategy results in many false positives.

[32] Note that once the Pyton Programme 2 finds a match in ISCO, ESCO and O*Net occupations references, it does not continue to look for more.

To mitigate this issue, we recoded our function comparing two lists such that matches only count if ISCO, ESCO or O*NET references are more than one word.[33] With this approach, we find matches for 2,541 of our 4,147 web scraped occupation titles (61.27%). Just under 30% of our candidate occupation titles match with ISCO and ESCO: 29.18% and 27.22%, respectively. Once again we find fewer matches with O*NET, where 483 occupation titles or 11.65%. Thus, after narrowing our longlist to only job titles with no correspondence in ISCO, ESCO and O*NET, the list is still quite long at 2,541 potentially new and emerging occupations. These are further discussed in Results.

2.6 Stage VI – Identify bigrams corresponding to potential new occupations

From the longlist of job titles created in Stage IV, Python Programme 2 also finds the most frequent bigrams from the selected occupation titles. From these, we have an idea of what words new and emerging occupation may contain.

N-gram models are widely used tool in statistical natural language processing. They have many uses because they capture the context in which language is used, reflecting which words are likely to precede or follow another. The longer the n-gram (the higher the n), the more context is captured, though the optimum length depends on the application. In our case, occupations tend to contain few words,[34] so we decided to use bigrams,[35] which are simply two adjacent words.

What is the value of bigrams in this analysis? In essence, the present report is based on applying word frequency analysis techniques to a 'box of words'. This is very useful to understand salience - i.e. how frequently a word appears - but less so to understand context. Bigrams help to shore this up, providing a helpful tool to better understand the context in which a text appears.

Take for example one of the occupation titles identified in Stage V: 'digital marketing executive'. In our web scraped data, we have many other occupations with titles like 'digital marketing lead', 'digital marketing associate', 'digital marketing analyst', 'digital marketing consultant'. In these, the most descriptive words are 'digital marketing', while the other words are variable. Digital marketing seems to be a *field*, whereas the other words in the occupation title indicate something of less interest - mostly related to seniority and contract. Bigrams help us sort this out, further isolating the signal from the noise in our data.

We analysed bigram frequency for the longlist created in Stage V (which only contains occupations containing a keyword, and without a match in the ISCO, ESCO or O*NET classifications). The top five most frequent bigrams are shown in Table 8.

[33] If we had a one-word occupation title, this strategy means we would not be able to find a match. However, only eight of our 4,147 occupation titles in the web scraped data are a single word. Upon reviewing them, they do not contain enough information to form a meaningful occupation title, so they were discarded.

[34] Median length of an unprocessed job title is five words while mean length is just over eight words.

[35] We also tested trigrams. They ultimately dropped from analysis, as they did not seem to generate better insights than bigrams.

Table 8. Top five bigrams by keyword

Ai	Chain	Cloud	Cyber
Ai engineer	Supply chain	Cloud support	Cyber security
machine learning	Chain planner	Support engineer	Security engineer
Engineer ai	Director supply	Google cloud	Director cyber
Developer ai	Global supply	Engineer cloud	Cyber risk
Ai cyber	Chain specialist	Cloud platform	Cyber threat
Digital	**Mobile**	**Model**	**Network**
Digital marketing	Mobile developer	Model validation	Network engineer
Marketing executive	Mobile app	Model farm	Network support
Digital medium	Mobile team	Farm rd	Network security
Digital development	App developer	Model risk	Security engineer
Social medium	Mobile patrol	Risk model	Network operation
Online	**Platform**	**Research**	**Tech**
Online partnership	Cloud platform	Research analyst	Tech company
Online content	Platform engineer	Research assistant	Tech support
Partnership group	System platform	Research fellow	Global tech
Online sale	Platform digital	Post doctoral	Accountant tech
Online marketing	Saas platform	Doctoral research	Analyst tech
Web			
Web analyst			
Web service			
Web designer			
Web content			
Amazon web			

2.7 Stage VII – Linking bigrams and occupations

As a final step, we search all pre-processed occupation titles for any containing the most frequent bigrams identified in Stage VI. We opted to use the most frequent five bigrams per keyword, though more or fewer could have been used as well.

The matching occupation titles are then stored in a data frame alongside their associated bigram, original unprocessed occupation title, and an example URL. This allows us to manually check the occupation on *Indeed* if more detail is desired. We then count the number of times each pre-processed occupation title appears, append the count to a new column, then drop duplicates by pre-processed occupation title. This creates our second list of candidates for new and emerging occupations.

3. Results

3.1 Longlist using keywords, ISCO, ESCO and O*NET

Ultimately we identify 2,541 unique occupation titles that (1) match a keyword, and (2) do not match any occupations currently recognised by ISCO, ESCO, and O*NET. In principle all of these are candidates for new and emerging occupations. With this many candidates, it is very difficult to consider all results in depth.

However, we are looking for *trends* and therefore can narrow this down to occupations appearing more often. It is difficult to know how many times an occupation title must appear before we consider it a good candidate for a new and emerging occupation. Taking only the ten most frequently appearing occupations from this shortlist yields the following:

Table 9. Shortlist of potentially new and emerging occupations

Keyword	Occupation title	n
Network	Network engineer	376
Digital	Digital marketing executive	269
Research	Research assistant	100
Chain	Supply chain planner	96
Network	Network support	85
Chain	Supply chain analyst	82
Chain	Supply chain specialist	77
Network	Network security engineer	68
Research	Research analyst	67
Cloud	Cloud architect	65

Some of these occupations intuitively seem like they might well correspond to new and emerging fields. 'Network engineer', 'digital marketing executive', and 'cloud architect' appear to be specialised, differentiated, and quite promising.

However, we also see several false positives. 'Research assistant' and 'research analyst' appear in the shortlist, having no matches in the ESCO, ISCO and O*NET data. Upon inspecting these data, the reason no match turned up is that the reference data use 'research associate' instead of 'research assistant'. Similarly, ESCO contains 'market research analyst', but not only 'research analyst'. While 'supply chain manager' is in our reference data, 'supply chain analyst' and 'supply chain specialist' are not. This again demonstrates the limitations of matching our web scraped data to existing occupations.

Another difficulty is whether we would consider examples like 'supply chain analyst' and 'supply chain specialist' to be two separate occupations, or two ways of describing the same occupation. The present analysis is not able to help us in this regard.

3.2 Bigrams

All in all, we identify 1,561 candidates for new and emerging occupations using the bigram method. Each of these candidates (1) does not match any occupations currently recognised by ISCO, ESCO, and O*NET, and (2) contains one of the bigrams shown in Stage VI.

Again, we must consider how many times an occupation must appear before we assume it to be new or emerging. The candidate occupations appear between 1 and 613 times with a very long tail; the mean is 5.8 and the median is 2. As above, we only show and discuss the top ten most frequent occupation titles here.

Table 10. Top candidates for new and emerging occupations - bigram method

Occupation	Bigram match	n
Marketing executive	Marketing executive	613
Network engineer	Network engineer	376
Technical support engineer	Support engineer	367
Security engineer	Security engineer	282
Digital marketing executive	Digital marketing	269
Support engineer	Support engineer	183
Sale marketing executive	Marketing executive	168
It security engineer	Security engineer	138
Network security engineer	Security engineer	136
Post doctoral researcher	Post doctoral	111

Looking to this table, the results are rather similar to the results found without using bigrams. The top occupation, ‘marketing executive’, does not intuitively seem like anything new. Several others, such as ‘post doctoral researcher’, are quite general. Rather than being new occupations, it seems more likely that they have existed for some time, but did not match occupations in ISCO, ESCO and O*NET closely enough to have been removed from consideration. They should therefore be regarded as false positives.

Several others including ‘network engineer’, ‘technical support engineer’, and ‘IT security engineer’ are more promising candidates for new and emerging occupations.

4. Reflections and limitations

This analysis made use of a rich source of data - nearly 500,000 job titles scraped over 18 months - with the ultimate aim to identify potentially new and emerging occupations. We attempt this by comparing web scraped occupation titles with occupation titles from ISCO, ESCO and O*NET data sets. While the web scraped data is representative and timely, it is not without problems. Here we discuss two overarching issues that the present methodology cannot overcome.

The first core issue is that *online job ads are designed for recruitment purposes rather than analysis* (Carnevale et al., 2014), which makes it very difficult to sort out the signal from the noise. Online job ads are not designed to be as objective as possible, but to maximise the chance that candidates of a certain profile notice the ad and click it. On the other hand, we compare online job ads with data sets worded objectively with a view to analysis.

The terminology used in occupation titles might be generously described as *inventive*. A franker assessment would be that a great portion of online occupation titles are full of marketing nonsense. A few examples in the scraped data are illustrative of this argument. Take the title 'Tech-savvy Google Cloud Help Desk 33K'. Looking through the details contained in this listing, a more accurate title, which corresponds to established occupation lists, might be 'Computer administrator, entry-level'. Another example found in the data is the occupation 'sandwich artist', which fast-food restaurant Subway calls its entry-level labourers.[36] Simply stated, companies often rely on euphemisms and advertising language to make occupations seem more novel and desirable.

Unfortunately, *the unobjective occupation titles significantly complicate analysis*. While the text cleaning of Stage II did a decent job of removing much of the noise in the data, more complex text analysis is necessary. Beyond simply matching occupations word-for-word in Stage V, we have not attempted to identify which established occupations a given job title might correspond to.

Taking the example of 'sandwich artist', a human researcher might realise that the occupation corresponds to ESCO's 'quick service restaurant crew member', ISCO's 'Handler, food: fast food', and O*NET's 'Combined Food Preparation and Serving Workers, Including Fast Food'. However, it is much more difficult to *programmatically* match 'sandwich artist' to these known occupations, given that the words 'sandwich' and 'artist' do not. While we could have coded our programmes to replace 'sandwich artist' with something else, we could not have done this for the thousands of other occupation titles with similar problems. Natural language processing - a subset of machine learning - could be used for such a purpose, but even this requires a great deal of training data produced by humans.[37] A further requirement would be web scraping the descriptions associated with each occupation, which would allow us to better understand the content of occupations, and not simply their titles.

A second core issue is that the names of occupations, and what the occupation entails, evolve over time. One common example is that 'data scientist' seems to have largely replaced 'statistician' in previous years.[38] To some extent this reflects different contents of occupations. Today's data scientist is more likely to use machine learning than yesterday's statistician, and has access to exponentially more data, but one may question if these are two separate occupations or the evolution of one.

[36] For a description, see http://www.subway.ky/layouts/page_careers_sandwich.html.

[37] Platforms such as Amazon Mechanical Turk, Clickworker, and many more outsource such data training tasks to human workers in the cloud. Humans would, for example, be asked if x is the same as *y*.

[38] The word 'statistician' appears in the scraped data fewer than 50 times, so it was not even considered for subsequent analysis. As for what data scientist means, one writer on Twitter quipped that 'A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician'. See https://twitter.com/josh_wills/status/198093512149958656.

Ultimately, identifying new occupations can easily become more of a philosophical or linguistic discussion rather than an economic one. *Using the data and methods described, it is very difficult to determine if we are measuring a change in occupations, or the change in the way occupations are discussed.*

Other important questions remain, and to the best of our knowledge, are difficult to satisfactorily answer. For example, how often must some variation of an occupation appear before we accept that it is an emerging occupation? Does an emerging occupation need to be meaningfully different than established occupations, and how can this be determined?

In summary, the present analysis is limited in a few important ways.

- many occupation titles in online ads are not objective;
- the data set only contains the occupation titles without descriptions;
- while few could contest that occupation change all the time, agreeing on which distinct occupations exist at a given point in time is very complex.

5. Next steps

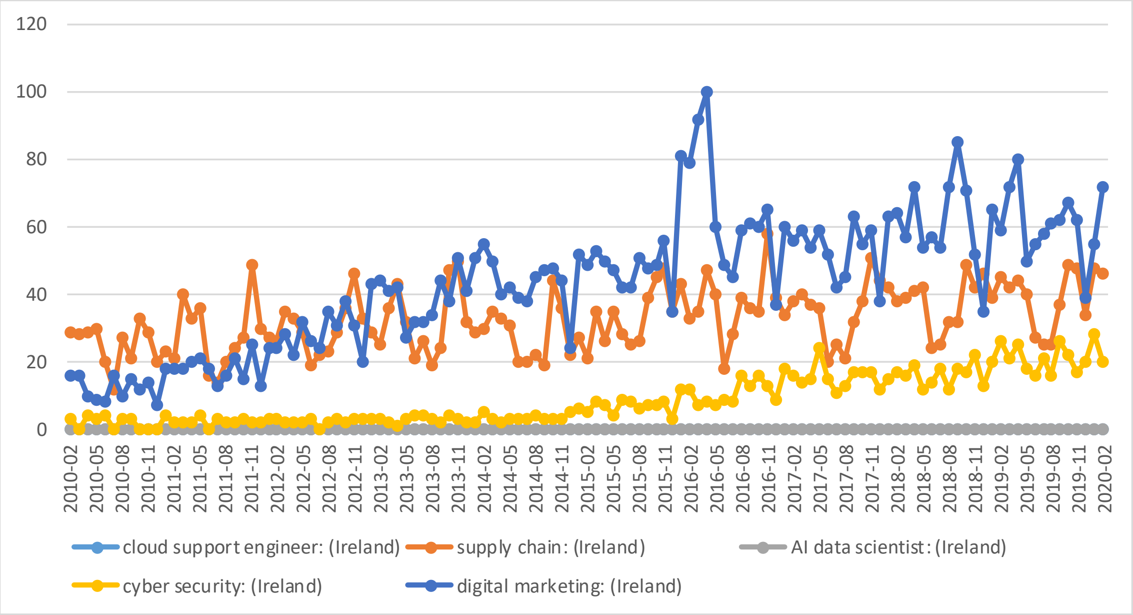
5.1 Repetition over time

Because the web scraping is automated, it could be very useful to continue extracting data over time. Different analysis methods can be applied at any time, but most useful to future research is ensuring that the source data is extracted regularly, ideally at the same one month intervals. Future scraping should ensure that the website can be scraped exhaustively, not merely 80% or so as described in Stage 1.

It would also be very interesting to see how ESCO, ISCO and O*NET are updated over time, and whether the new versions reflect occupations identified as new and emerging. It is not clear what timeframe would be required for such an endeavour, as ESCO, ISCO and O*NET are updated at different intervals, and it is not clear how long of a period is required to observe meaningful changes in occupations.

5.2 Validating results

For any occupations proposed as new and emerging, a logical next step is validating results. Among the methods that might be used are: stakeholder feedback (e.g. consulting experts from trade unions, sector representatives, public employment offices, academics, or others on whether these occupations are in fact new or emerging), cross-checking 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).



An additional, easy check to see whether an occupation seems to be growing is using Google Trends. We demonstrate this below using a few bigrams and occupation titles. The previous ten years do seem to indicate overall growth in interest for most of these terms. Of course, increased search prevalence corresponds to a term's salience rather than labour market realities, but this still could be a valuable clue as to whether an occupation is emerging.

5.3 Scaling up the methodology

In principle, this methodology can be applied to other countries. However, several challenges appear insufficiently resolved.

Part of the difficulty of transposing this methodology to other cases, whether they are other job portals in Ireland or other countries, lies in updating the scraping programme. Not all websites are scrapable, but most job boards are, at least in principle. However, each scraping programme must be custom written for a particular website. This requires time and some basic programming skills. As we selected a job portal that operates in multiple countries (see Table a1), this difficulty could be partly avoided. Other *Indeed* websites would be natural candidates for applying the present methodology or some variation thereof.

In the case of job boards using any language other than English, handling translations will be key. The methodology must rely on a mostly-automated but sufficiently accurate method, as previous results found that Google Translate API was not accurate enough. We also encountered significant semantic issues when using just English data. For this reason, we consider a multi-lingual expansion to be quite an ambitious proposition.

5.4 Natural language processing

This report attempts to match occupation titles found online with occupation titles from databases like ISCO, ESCO, or O*NET. Even after substantial data cleaning, this proved a difficult task using rather simple methods of text analysis.

Natural language processing offers more powerful techniques that would be useful in handling data such as ours. These could help to address some of the contextual and semantic problems described. A good start would be scraping the job descriptions in addition to the titles. While this requires a huge amount of storage space for a list of occupations as long as ours, machine learning is only as effective as the training data it uses.

With a sufficiently large corpus, machine learning can be used to further refine the data. For example, it might correct misspellings and other errors in the data, correctly identify which tokens are specific acronyms or company names, etc.

5.5 Other uses for job board data

Job board data are very rich, but we encountered many difficulties using *occupation titles* to identify new occupations. However, different avenues may still be very promising for future research using job board data. For example, researchers may gather occupation titles *and accompanying text data* (e.g. job descriptions) to analyse new and emerging occupations, as well as other topics of interest like skills demand. The job descriptions often contain quite detailed descriptions of what the job entails, which provides critical contextual information and enables more detailed analysis.

At least one other study has already done something very similar. In an ambitious project led by Cedefop, 'Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis', researchers used web scraped online vacancy data (alongside expert interviews and other steps) to assess changing skills demand, job requirements, as well as new and emerging occupations and skills for the entire EU-28. The project resulted in an excellent EU overview, as well as background report for individual countries.[39] By web scraping the entirety of job vacancy ads, and spending more effort understanding and validating findings with national labour market experts, it avoided some of the pitfalls we encountered.

While an excellent research project, Cedefop collected data over a much shorter timeframe than we did – six or fewer months compared to our 18. To understand how the labour market develops over time, and account for seasonality of job demand, new research conducted over a longer timeframe may still yield valuable insights.

[39] See all publications here.

appendix 1

In addition to the methodology described above, which treats all job titles gathered over 18 months as a single group, we experimented with time series analysis in an attempt to understand which occupations emerged from the first month to the last.

The time series has a few advantages in that we do not limit occupations to a set of keywords determined by the researcher. The other advantage is that we are comparing the web scraped data to itself, rather than to a benchmark of ISCO, ESCO and O*NET data. This avoids comparing objective data with messier data gathered online.

In the time series, we used a moving benchmark similar to that in Beblavý et al. (2016). Essentially, all occupation titles gathered in month one become the first benchmark. Next, we test the next month against this benchmark. Any new occupations are added to a list of potentially new occupations, then they are also added to the benchmark. This continues for all subsequent months.

However, this method achieved rather poor results. The occupations appearing as new were almost always slight variations of the same thing – only distinguished from one another by junk information that was not removed in the text cleaning. Unlike Beblavý et al. (2016), which used keywords associated with a job opening, we simply used the occupation title. Our strategy did not work as well because the occupation title contains much more noise than the tags a website uses.

This finding is another demonstration of two lessons from the present analysis. First, when web scraping data for analysis, one ought to err on the side of gathering too much data, as it may come in handy. The tags and keywords in each occupation could have been quite useful in our analysis. Second, the ‘occupation titles’ are better thought of as the titles of advertisements, which are not intended, or easily suited, for text analysis.

Table a1. Additional *Indeed* websites in EU

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

Table a2. 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.

References

Askitas, N. & K.F. Zimmermann (2015), "The internet as a data source for advancement in social sciences", *International Journal of Manpower*.

Beblavỳ, M. et al. (2016), "Occupations observatory-methodological note", *CEPS Special Report*, No. 144.

Beblavý, M. et al. (2016), "Skills Requirements for the 30 Most-Frequently Advertised Occupations in the United States: An analysis based on online vacancy data", CEPS Special Report 132, CEPS, Brussels, March (https://www.ceps.eu/publications/skills-requirements-30-most-frequently-advertised-occupations-united-states-analysis).

Benfield, J.A. & W.J. Szlemko (2006), "Internet-based data collection: Promises and realities", *Journal of Research Practice*, Vol. 2, No. 2, pp. D1–D1.

Carnevale, A.P. et al. (2014), "Understanding online job ads data: a technical report", Georgetown University, McCourt School on Public Policy, Center on Education and the Workforce, April.

D'Amuri, F. & J. Marcucci (2010), "'Google it!'Forecasting the US unemployment rate with a Google job search index".

Kuhn, P. & H. Mansour (2014), "Is internet job search still ineffective?", *The Economic Journal*, Vol. 124, No. 581, pp. 1213–1233.

Kuhn, P.J. (2014), "The internet as a labor market matchmaker", *IZA World of Labor*.

Schierle, M. et al. (2008), "From spelling correction to text cleaning–using context information", *Data Analysis, Machine Learning and Applications*, Springer.

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

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

All rights reserved. No part of the report may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, and may not be quoted or cited, without prior permission in writing from the Project Co-ordinator.

The views expressed during the execution of the InGRID-2 project, in whatever form and or by whatever medium, are the sole responsibility of the authors. The European Union is not liable for any use that may be made of the information contained therein.