Statistics Canada
Employment and Social Development Canada
Statistique Canada
Emploi et Développement social Canada
Canada
LMI CIMT

# LMI INSIGHTS

ISSUE № 32
June 2020

# Through the Looking Glass: Assessing Skills Measures Using 21st Century Technologies

## Key Findings

- Advances in computing power, data management and machine learning enable efficient, automated large-scale collection and analysis of data from online job postings.
- Online job posting data provide two important benefits that make it a valuable complement to official statistics on vacancies: (1) They provide insights on the work requirements of jobs, including skills, that employers are seeking in near real-time at both local (e.g. a municipality) and granular (e.g. detailed occupations) levels. (2) They can be analyzed at a fraction of the cost of traditional survey methods.
- When using online job posting data, there are some important limitations to be aware of. The first relates to its representativeness, as online job posting data may be skewed towards occupations in certain industries and regions and by firm size and educational requirements. Second, some firms (e.g. SMEs or those operating in sectors like agriculture) are more likely to hire by word of mouth than to post online. Third, differences in data processing methods among data providers can yield different results in terms of work and skill requirements. Finally, there is no way to tell which work requirements are critical for the position in question. In fact, the way a job posting is written does not necessarily convey the actual work required. The data allow us only to observe that certain requirements are more frequently stated by employers across online job postings.
- Data obtained from online job postings offer an immense opportunity to complement existing data sources on the skill requirements of jobs. Like all data sources, such as O*NET, the caveats and limitations should be transparent to inform decision making.

## Introduction

During the past year, LMIC has been working on a joint project with Employment and Social Development Canada (ESDC) and Statistics Canada (STC) to describe jobs in terms of their skill requirements, as well as other job/worker characteristics. One possible approach previously explored is leveraging the breadth of information available through the US Occupational Information Network (O*NET). This *LMI Insight Report* focuses on another approach: collecting, analyzing and structuring job requirements, including skills, from online job postings using

21st century approaches such as machine learning, application program interfaces (APIs) and web scraping (see Box 1).

## Explain, Please

The popularization of open source programming languages, advances in computing power and improvements to information management have allowed for the efficient, automated large-scale acquisition of data from online sources. At the same time, advances in the field of machine learning have enabled us to analyze these large sets of data to generate new insights in near real time.

Increasingly, raw data from online job postings are being leveraged via machine learning to complement official statistics for labour market analysis. Since these data — available in real-time by detailed location (e.g., city or town) — are collected from publicly available websites, they help improve our access to more local, granular, timely labour market information (LMI). Moreover, they offer the opportunity to explore skills information from the perspective of employer demand for talent.

## From Job Requirements to Skills Information

When unstructured data is collected from an online job posting, the text needs to be parsed and categorized. A common parsing technique is Document Object Model (DOM) parsing, which allows a program to extract information from a web page by referencing its position within the document. This method, however, returns all the data at the specified location. In other words, the result is simply raw text that still needs to be formatted for analysis. To accomplish this, data analytics firms typically apply advanced — often proprietary — natural language processing (NLP) algorithms that use machine learning to classify the unstructured text according to a pre-existing taxonomy of skills and work requirements (see Box 2).

### Box 1: What are 21st Century Technologies?

**Machine Learning**

Machine Learning (ML) is a set of modern approaches to statistical analysis that rely on algorithms to process large sets of data. Distinguished from other statistical approaches, ML algorithms are designed to become more accurate in predicting outcomes (e.g. classifying text) without being explicitly programmed to do so. In this sense, ML is a subset of Artificial Intelligence (AI). In the case of online job postings, algorithms are designed to associate job postings with an occupation and to categorize the raw text into a set of work requirements (see Box 2).

**Web Scraping**

According to Statistics Canada, web scraping is a process through which information is gathered from public websites for retrieval and analysis. One application of web scraping is collecting data from online job boards and corporate websites and then "cleaning" the text. This process can either be done in-house or outsourced. Currently, several data analytics firms (e.g., Vicinity Jobs, Burning Glass Technologies and TalentNeuron) collect and analyze job posting data from multiple Canadian corporate websites and aggregators (e.g. indeed.ca).

**Application Programming Interface (API)**

APIs are a software intermediary that allow an external computer program to obtain information from an internal source through a set of protocols defining the type of data to be accessed. Crucially, APIs allow independent systems (or computer languages) to speak to one another. For example, whether you use a Windows computer or an iPhone to access flight information, the data can be delivered to you via the same API access point. Several job posting aggregators make APIs available so that third parties can more easily access and download structured information.

## What Does the Data Look Like?

To provide a general overview of the data available, we present a summary of the information from 2,467,709 online job postings from across Canada in 2019. The specific results shown here are provided by the Big Data analytics firm Vicinity Jobs, but the type and frequency of information available is common across all data providers.[1] This data has already been cleaned and structured by the scraping firm. The key variables of interest and the frequency with which each appears are displayed in Table 1. Job posting information from Vicinity Jobs will be available on LMIC's forthcoming Job Posting Dashboard.

### Box 2: What Are Taxonomies and Why Are They Useful?

In its simplest form, a taxonomy is a classification system. In terms of skills, taxonomies are used to organize the wide-ranging information into usable categories. Taxonomies also provide a common language for discussing the skills required for jobs and occupations.

Of the several skills taxonomies in existence, one of the most popular is embedded in the US O*NET system. Other taxonomies include the European Skills/Competences, qualifications and Occupations and Canada's new Skills and Competencies Taxonomy.

**Table 1. Availability of information contained in Vicinity Jobs (2019) data by variable**

| Category | Variable | Description | Share of Job Postings |
|---|---|---|---|
| **Employer** | Employer's name | The name of the employer responsible for the posting | 39% |
| **Location** | City or town | City or town where job is located | 91% |
| | Census Division | Census Division where job is located | 93% |
| | Economic Region | Economic Region where job is located | 93% |
| | Province or Territory | Province where the job is located | 100% |
| **Industry (NAICS)** | 6-digit NAICS | Detailed industry classification of posting | 43% |
| **Occupation (NOC)** | 1-digit NOC | Broad occupational classification of posting | 87% |
| | 4-digit NOC | Detailed occupational classification of posting | 71% |
| **Experience** | Experience | Whether experience is explicitly stated as a requirement | 16% |
| **Education** | Education | Type of education required for the position (e.g., high school completion, graduate degree) | 42% |
| **Certification** | Certification | License, certification, professional development required | 23% |
| **Offered Salary/ Wage** | Salary | Remuneration indicated in posting | 16% |
| **Duration** | Permanent or temporary job | | 38% |
| **Type** | Full-time or part-time job | | 95% |
| **Work Requirements** | Tools, skills, knowledge, technology, and other descriptors identified by the employer as required for the job | The full set of work requirements* categorized into ESDC's Skills and Competencies Taxonomy:<br>1) Knowledge<br>2) Skills<br>3) Tools and Technology<br>4) Other[2] | 90% |

Source: Vicinity Jobs (2019).
*Organized by LMIC in partnership with Vicinity Jobs and ESDC.

## What Insights Can This Data Provide?

As discussed in **LMI Insight Report no. 14**, one of the main advantages of data collected and analyzed from online job postings is the ability to identify the work requirements that employers explicitly ask for. Using the data provided by Vicinity Jobs, we can calculate the shares of specific work requirements (e.g., Knowledge, Skills, Tools and Technology) listed by employers for different locations and occupations. In **Figure 1** we show the top four work requirements by group (i.e., Skills, Knowledge, Tools and Technology, and Other) for nurses (NOC 3012) in Winnipeg. We can also look at one category, such as skills. In **Figure 2** we present the top five skills across all job postings in Winnipeg in 2019. Similarly, in **Figure 3** we see the top five tools and technologies for computer programmers and interactive media developers (NOC 2174).

**Figure 1. Communication, CPR, flexibility, and Microsoft Word top the work requirements for registered nurses and psychiatric nurses (NOC 3012) in Winnipeg (2019).**

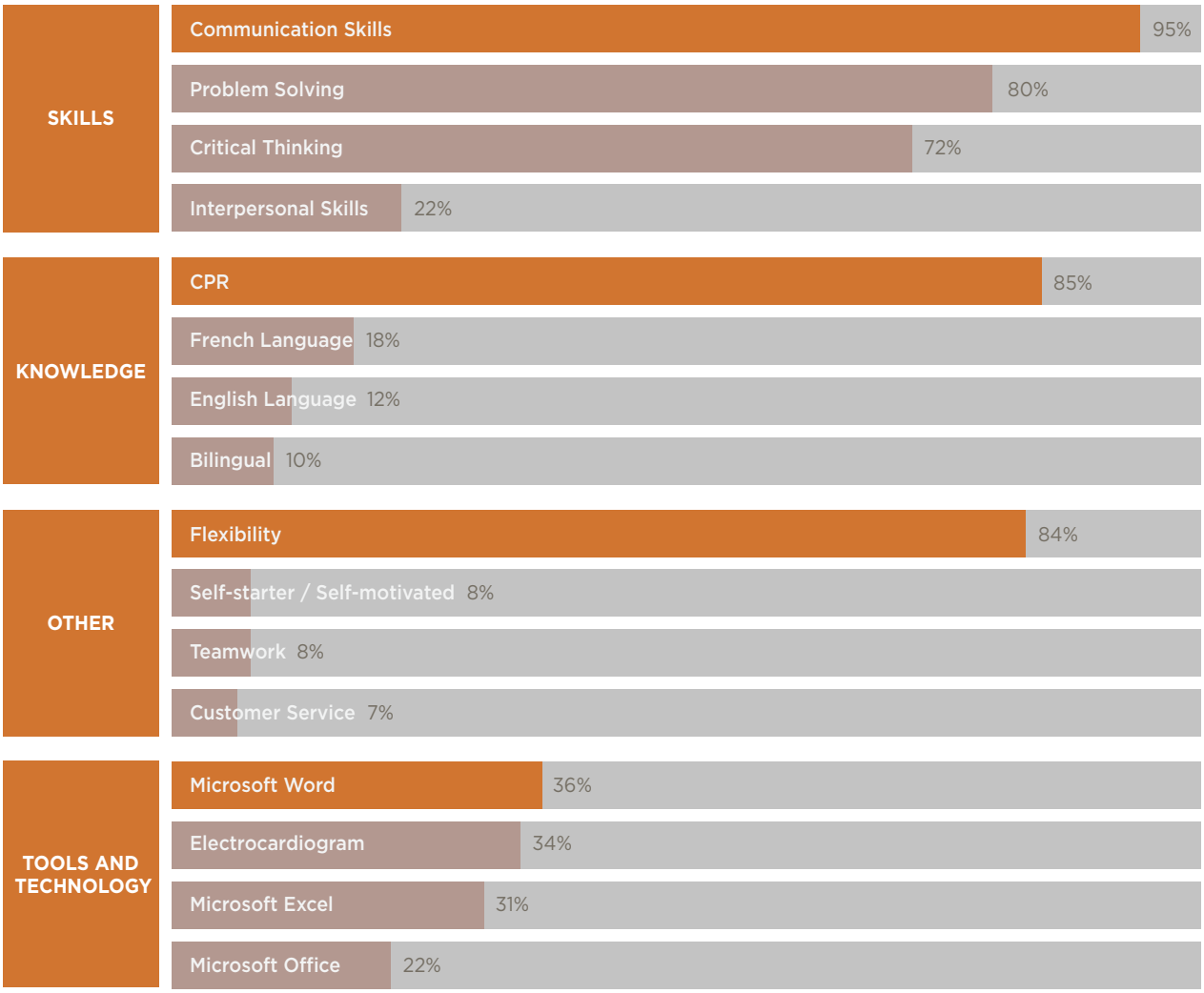


**Figure 2. Communication skills are the most frequently demanded skill in Winnipeg across all job postings (2019).**

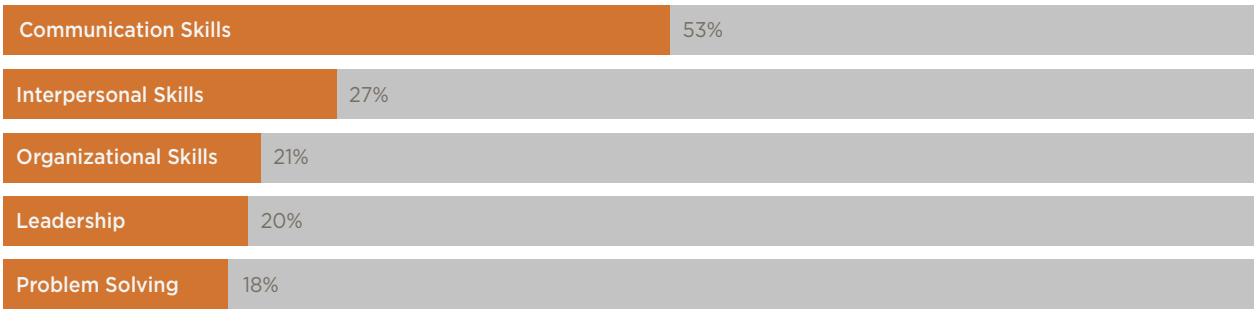
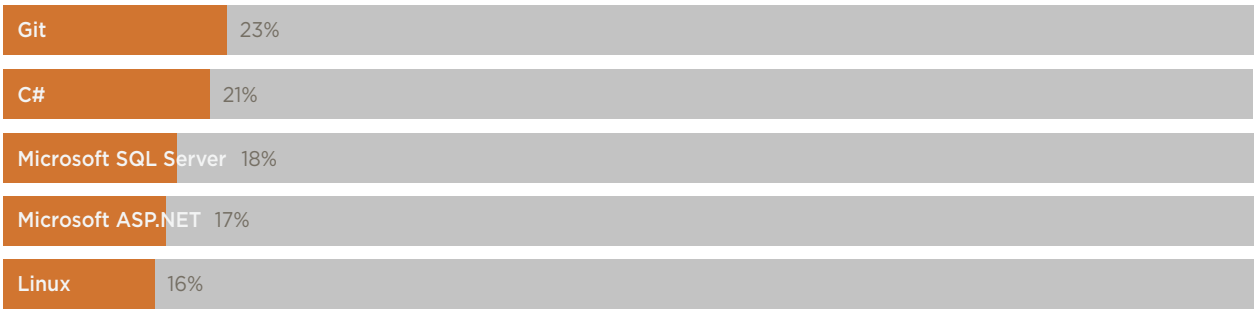



**Figure 3. Git system software is the most frequently required tool and technology for computer programmers and interactive media developers (NOC 2174) in Winnipeg (2019).**




## Advantages and Limitations

Data collected from online postings offer several key benefits, the most obvious being insight into the work requirements sought by employers in near real-time. This makes it possible to identify the knowledge, skills, tools and technologies, etc. that job seekers need in order to be successful in today's labour market. Since this information comes from vacancies posted on public job boards, there are no restrictions that would limit the level of detail accessible.

Online job postings are also useful for the large volume of data and localized information they provide. An analysis of the 2019 Vicinity Jobs data, for example, shows that 91% of postings are available at the city or town level, making it the most local source of work requirements information currently available. Furthermore, we can obtain this volume of data at a fraction of the cost of traditional survey methods. Machine learning algorithms used to scan webpages, for example, can be implemented with open source (freely and publicly available) programming languages. The millions of data points available potentially allow for greater statistical reliability and for clustering observations as needed for customized data analysis.

This information is valuable to a broad range of stakeholders, including educators developing course curricula, students choosing their training and educational pathways, policy makers allocating funding for upskilling, and organizations engaging in workforce planning. Moreover, examining the change in demand for work requirements over time makes it possible to identify emerging trends early, such as the recent surge in demand for knowledge of Artificial Intelligence for actuaries.

Using online job posting data also comes with some important limitations, including its representativeness. Previous studies comparing job posting data to nationally representative labour market surveys have identified several differences. Large employers, **for example**, are more likely to post all available jobs online, whereas smaller businesses post only executive positions. Similarly, occupations such as healthcare or IT are more likely to be overrepresented while those such as construction are underrepresented.

Second, employers in different sectors of the economy recruit in different ways. Small firms and those working in sectors like agricultural are more likely to hire by word of mouth than to post online. To further complicate things, some online sources like LinkedIn utilize anti-scraping technology to prevent third parties from extracting their data at scale. In addition, many data analytics companies do not analyze French language postings. All of these limitations can make it difficult to conduct trend analysis and comparisons over time as the sample size and composition of collected postings can vary year-over-year.

Third, natural language processing methods that structure the raw posting text differ across data providers. Since many of these methods are proprietary, they are closed to public scrutiny. Providing open, transparent methodologies is an important step to creating trustworthy, quality data but this remains a challenge.

Fourth, there is no way to tell which work requirements are critical for the position in question. The data allow us only to observe that certain requirements are more frequently stated by employers across online job postings.

Finally, online job postings may suffer from various biases, such as omitted information or skill-inflation. Omitted information is especially likely if employers expect certain work requirements to be obvious to applicants. For example, numeracy is clearly essential for engineers yet is unlikely to be mentioned in a job posting. Alternatively, some employers may specify work requirements in excess of what is truly needed for the position. In fact, such biases will skew the results of work requirements analyses if online job posting data is used exclusively.

## The Way Forward

Given its timeliness, granularity, localness and frequency, data obtained from online job postings present an immense opportunity to complement existing data sources. This is evidenced by the wide use of this type of data across stakeholders such as government agencies, workforce planning boards and private organizations to improve career planning, skills training, program development and more. Additionally, the data give us a glimpse of employer demand for particular skills and other work requirements in the labour market.

Yet, as with all data sources, the important limitations of lack of representativeness, over or understating the skill requirements and so on mean that enhanced strategies are needed. One solution that LMIC, ESDC and STC will explore in a forthcoming *LMI Insight Report* the advantages and limitations of asking employers directly about their skill requirements as is now done in the United Kingdom and Australia.

## Acknowledgements

This *LMI Insight Report* was prepared jointly by the staff of the Labour Market Information Council, Employment and Social Development Canada (Labour Market Information Directorate) and Statistics Canada (Centre for Labour Market Information). We would like to thank David Ticoll (University of Toronto), Jacob Loree (**Ryerson University**), and Ron Samson and Austin Hracs (**Magnet**) for their comments.

Your feedback is welcome. We invite you to provide your input and views on this subject by emailing us at **info@lmic-cimt.ca**.

## End Notes

1 Vicinity Jobs, one of the leading Canadian vendors of online job posting data, has partnered with LMIC to leverage this data source for labour market analytics. Although our analysis here is based primarily on Vicinity Jobs data, the limitations noted throughout this report are common to all data providers (Burning Glass Technologies, Talent Neuron, etc.).

2 Vicinity Jobs first links job postings to work requirements using its own proprietary taxonomy, classifying the raw text data into five groups: 1) General/Soft Skills, 2) Specialized Skills, 3) Technologies, 4) Tools and Equipment, and 5) Other. In partnership with Vicinity and ESDC, LMIC reclassified each work requirement based on ESDC's Skills and Competencies Taxonomy: 1) Knowledge, 2) Skills, 3) Tools and Technology, and 4) Other.