European Commission

JRC Technical Report

# The Professional Lens: What Online Job Advertisements Can Say About Occupational Task Profiles

JRC Working Papers Series on Labour, education and Technology 2021/13

Matteo Sostero, Enrique Fernández-Macías

LABOUR EDUCATION TECHNOLOGY

Joint Research Centre

This Working Paper is part of a Working paper series on Labour, Education and Technology by the Joint Research Centre (JRC) The JRC is the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

**Contact information**
Name: Matteo Sostero
Address: Joint Research Centre, European Commission (Seville, Spain)
Email: matteo.sostero@ec.europa.eu
Tel.: +34 9544-87118

**EU Science Hub**
https://ec.europa.eu/jrc

JRC125917

Seville: European Commission, 2021



How to cite this report: Sostero, M. and Fernández-Macías, E., *The Professional Lens: What Online Job Advertisements Can Say about Occupational Task Profiles*, Seville: European Commission, 2021, JRC125917.

# The Professional Lens
## What Online Job Advertisements Can Say About Occupational Task Profiles

Matteo Sostero (European Commission – Joint Research Centre),
Enrique Fernández-Macías (European Commission – Joint Research Centre)

## Abstract

Data from online job advertisements are increasingly used in the emerging area of "skills intelligence" to describe labour market dynamics and the demand for skills in different occupations. Collecting this data involves gathering unstructured information from the internet and processing it into structured datasets, which may provide a biased description of the labour market. We present a framework for these different sources of bias, in terms of representativeness of occupations and their task content. We analyse the Nova UK dataset of online job advertisements from Burning Glass Technologies, containing over 60m individual job ads for the United Kingdom from 2012–2020. We compare the occupation task profiles embedded in this data with the JRC-Eurofound Task Database, through a new *Skill-Task Dictionary*. The dictionary classifies the rich but unstructured information on "skills" describing individual occupations into the hierarchical Task Taxonomy developed by the JRC and Eurofound, and measured through occupation surveys. In general, we find that the task profile implied in job advertisements is relatively consistent with the EU Task Database across most occupations, especially for intellectual and social tasks, and for tools of work. However, online job advertisements in general (and Nova UK in particular) tend to focus especially on professional occupations, which are relatively better represented in their numbers and in their variety of skills and tasks, relative to less qualified occupations. We enumerate several types of bias that can occur with this data, and discuss possible future applications.

**Keywords:** Skills, Tasks, Online Job Advertisements, Job Vacancies

**Authors:** Matteo Sostero (European Commission – Joint Research Centre),
Enrique Fernández-Macías (European Commission – Joint Research Centre)

**Acknowledgements**: The authors would like to thank Simon Walo for contributing to an early version of the skill-task classification, as well as Martina Bisello (Eurofound) and Jiri Branka (Cedefop) for their helpful reviews and suggestions.

**Joint Research Centre reference number:** JRC125917

**Related publications and reports:**

- Fernández-Macías, E. and Bisello, M., A Taxonomy of Tasks for Assessing the Impact of New Technologies on Work, Seville, European Commission, 2020, JRC120618
- Rodrigues, M, Fernández-Macías, E., Sostero, M., A unified conceptual framework of tasks, skills and competences, Seville: European Commission, 2021, JRC121897.
- Bisello, M., Fana, M., Fernández-Macías, E. and Torrejón Pérez, S. A comprehensive European database of tasks indices for socio-economic research, Seville: European Commission, 2021, JRC124124.
- Fana, M., Cirillo, V., Guarascio, D., Tubiana, M., A comparative national tasks database, Seville: European Commission, 2020, JRC122699.

# Contents

# 1 Introduction

Online job advertisement (OJA) data is gaining popularity as a tool to collect information on the labour market. This type of data can complement labour-force and occupational surveys, including official vacancy statistics, by providing timely signals on the amount of job openings, the trends of specific industries and occupations, and the changing skills required by employers. It has been already used for a variety of purposes: to document the rising concentration of local labour markets in the United States and its effects on wages (Azar, 2018), to track the changing requirement of skills across occupation groups in the U.S. (Deming et al., 2019), or to quantify skill mismatches by occupations (Tijdens et al. 2018), among others.

The data from online job ads seems particularly attractive from the perspective of European skills policy. There is a widespread perception that the skills needed from the economy are changing at an increasingly faster rate, due to the acceleration of technical change in the digital era. This perception goes together with an increasing frustration with the existing data sources on the skills needed by European business, because those sources are perceived as too slow in capturing changing skills needs and somewhat divorced from the real needs of businesses. Most of the existing European sources for measuring skills needs are surveys, which are very powerful tools for capturing data on the social reality but also slow and cumbersome. In this context, online job ads appear like the perfect solution to the data needs of skills policy. Nowadays, a very significant proportion (in some countries or industries, most) of all job positions are advertised online, and in principle they reflect directly the skills required by actual job openings in the economy in real time, according to the employers' own assessment of what skills are required to cover a post. Online job ads thus appear as a direct and timely source of information for measuring the skills needed by businesses, which is precisely the kind of skills intelligence that European policy needs to respond to the challenges of the digital age.

Conceptually, we should clearly distinguish between a *job vacancy*, which denotes an open position that an organisation is actively trying to fill, and a *job advertisement* (or *job ad*, for short) – which publicly announces a position, as described by the employer. This distinction, used by Cedefop and Eurostat, is not always clearly made in the literature, but is important for analytical purposes. Online job advertisements are a timely and flexible tool for employers, which allow them to describe positions – including job titles, task description, salaries and benefits – in as much (or as little) detail as they choose. However, we should bear in mind that not all vacancies are advertised publicly (or online), as some are filled internally or through informal networks. The number or type of unfilled vacancies is typically measured though employer surveys, which can also measure how long a certain position has been vacant, which is generally not feasible with advertisements. Conversely, a single job ad can sometimes refer to an unspecified number of underlying open positions, or even none at all: some temporary recruitment agencies use job ads to build pools of candidates for potential future positions, and (anecdotally) some organisation use them to gauge the size and interest of certain profiles, for potential future hires. While there is certainly some overlap between vacant and advertised positions, analysing only the latter can provide a distorted image of the state of labour demand.

Because OJA data exists and is increasingly accessible (thanks to recent advances in machine learning and cloud computing), and because the policy need exists and is increasingly urgent, the use of online job ads as a source of skills intelligence seems unstoppable. Indeed, in English-speaking countries this data is already being widely used for labour market analysis – and tentatively to inform skills policy. In Europe, the diversity of languages creates an additional technical difficulty for the comparative analysis of online job advertisement data, and there are already some very promising pilot comparative databases of online job ads (Cedefop's Skills Ovate project being without any doubt the most advanced; see Cedefop 2019).

Online job advertisements are scattered across different websites and portals, where they are presented in different formats. Retrieving, extracting and classifying this text into structured, meaningful information on skills needs on a large scale is a complex endeavour, which requires a careful mix of automated algorithms with human supervision and expert domain knowledge, to ensure that the data are interpreted and classified correctly. This process inevitably entails some degree of human judgment and arbitrariness – about what exactly is a skill, or about which job titles or skills are interchangeable or distinct – as well as errors resulting from the usage of algorithms. Algorithms for automatic classification allow to operate at a scale unfeasible for humans and remove one-off mistakes in repetitive classification tasks, but they can introduce biases of their own. Despite widespread enthusiasm for the recent progress in the area of "Artificial intelligence", the state of the art of machine learning for text processing remains relatively crude: algorithms have no built-in understanding of the meaning of words and have a limited ability to account for the context in which they occur. Although job advertisements are a relatively standardised form of writing (at least in English) they are also a subtle form of prose, often carefully crafted in their choice of words, in which context and subtext are often essential to convey information, that can be lost via algorithmic processing. Algorithms dealing with the ever-changing texts of new job advertisements can easily ignore, misclassify or spuriously insert information on a large scale, thus requiring regular human supervision and validation.

Precisely because of the promises and apparent limitations of this type of data, we need to be careful with using it for informing policy until we have a proper understanding of what it measures. In particular, the nature of data collection of online job advertisements presents its own challenges. As often happens with the databases generated with complex artificial intelligence algorithms and big data processing, the skills intelligence generated by online job ad analytics looks rich and useful, but there is still a lot of uncertainty as to how those results exactly came about, and it is very difficult to identify possible sources of bias that surely affects this data as any other. With slow and cumbersome surveys, we know very precisely the possible sources of bias in the sampling of populations and the collection of information, and indeed we can quantify those biases and the uncertainty they generate. With online job ad data, we still know relatively little about those possible biases. We know a bit about the statistical bias that may result from relying on job advertisements that are posted online, because there are some comparisons between the theoretical universe of job openings and the observed sample of job ads. However, we do not know so much about the extent to which the job ad itself is a reliable source of information for the job position that it is in principle intended to fill. Perhaps there are systematic biases in the type of information that employers want to include in the ads they post online which may give us a wrong impression of some key labour market dynamics. Furthermore, we do not know much either about the possible biases that the AI-driven processing of vast amounts of job advertisements can introduce in the data. To transform the raw scattered and unstructured information into usable knowledge, there are a number of assumptions and inferences that must be made, which can also bias the information. In our view, online job advertisement data will not be fully usable for policy purposes until those potential sources of bias have been identified and systematised to a significant extent.

In this paper, we will try to contribute to a better understanding of the informational value of online job ads. We will focus on assessing to what an extent online job ads can be used to infer what people do at work and what skills they need. We will do this by first sketching a conceptual framework on how online job ads are produced and processed, and linking the task contents of jobs with the information contained in online job ad databases; second, by analysing in detail the internal structure of online job ads data to assess its plausibility according to our initial conceptual framework; third, by comparing the distribution of tasks across occupations that we can infer from job ads with the distribution that we have independently measured in a survey-based occupational tasks database. After these assessment exercises, we will discuss the informational content of online job ads and the implications of our findings for European skills policy.

# 2 Literature review and policy context

The term skill, which in the English language simply means the acquired ability to do something well, is very frequently used in the academic and policy debates around work and education, particularly when discussing the challenges of recent technological change. Skills (or more specifically, "high" skills) are often considered a pre-requisite for finding good jobs, and it is often assumed that those who lose their jobs or want to change jobs will require different sets of skills, to be acquired through so-called "re-skilling" or "up-skilling". This is because many believe that technological change – and digital technology in particular – as well as changes in business environment and practices – such as globalisation, changes in value chains, outsourcing, but also regulation and governance practices – require an increasing number of new skills, and are making existing ones obsolete. Labour market imbalances, such as unfilled job vacancies and structural unemployment, are described in terms of "skill shortages" or "skill mismatches" (McGuinnes and Pouliakas, 2018). This understanding is based on a certain reading of the academic literature, which finds that skills contribute to higher salaries, better profession-al development, and benefit from technological change (Acemoglu and Autor, 2011). In this view, the role of the education system – both vocational and compulsory education – is seen as providing students with the correct skills that they will need in the labour market.

This view is echoed with the European Commission (2020)'s European skills agenda for sus-tainable competitiveness, social fairness and resilience identifies skills as "crucial for long-term and sustainable growth, productivity and innovation [...] and competitiveness". In particular, it cites a number of benefits resulting from developing the correct skills.

> *Providing people with the right skills allows them to work more effectively and take advantage of advanced technologies, eliminates the major obstacle identified to business investment, prevents labour market mismatches and lays the ground for research and development (R&D) and firm-based innovation.*

To that end, the Skills Agenda sets out a number of actions to improve the development and matching of individual skills with the skills that employers need. These actions are coordinated across Members States, including initiatives targeting specific sectors ("Blueprint for Sectoral Cooperation on Skills") and aimed at collecting more information on skill needed by employers ("Skill Intelligence", see Cedefop 2019a, 2019b).

However, this overarching narrative on the importance of skills relies on different implicit definitions of what "skills" actually are, leading to confusion in the academic debate across disciplines, and ultimately in policy practice. There are different theoretical conceptualisations of the term (see e.g., Attewell, 1990). For the purposes of policy-oriented discussion, we will distinguish between five related but distinct usages of the term *skill,* depending on the context.

The first usage of the term is as *craft of a specific trade*, namely the ability to master the tools and methods of a production process. The word was probably first applied to denote the manual dexterity and know-how of artisans and craftspeople, but the usage has expanded beyond manufacturing, to the service sector. By extension, it is used to describe the essential, defining characteristics of a given job. For instance, Burning Glass Technologies calls skills the "DNA of jobs". In particular, in the Information and Communication Technology sector, the term "skill" has taken on a very specific meaning, namely proficiency with a given software or framework, such as programming languages, as tools of the trade.

The second usage of skill is as *occupation classifier (or ranker)*, as a distinguishing attribute of jobs that define occupational hierarchies. It is a relevant dimension of international occupational classifications such as the International Standard Classification of Occupations (ISCO) or the UK Standard Occupation Classification (SOC). For instance, ISCO distinguishes between four *skill levels* based on the degree of complexity and prior training required to execute a task in a given occupation (International Labour Organisation, 2012). However, the resulting expressions "high-skilled" and "low-skilled" are also often misused in common parlance to distinguish between jobs that are highly paid or not, or even to describe the people doing them.

A third, distinct usage of the term is as *education learning objective*. It is common to describe the goal of a given education curriculum in terms of the skills that learners are expected to master, such as reading and writing, or solving equations. A related but imprecise usage is "skill" to describe a level of education attainment, like a high-school diploma or university degree.

A fourth usage of skill is the *human capital view*, which uses "skill" as a synonym for "ability": an individual attribute that is valuable in the labour market, which is often thought to be unobservable directly. This attribute describes a latent characteristic of the individual (which can be cultivated and sometimes certified by education) and correlates with salient traits like educational attainment and earnings, which are difficult to disentangle empirically (see e.g., Cunha and Heckman, 2007). This conceptualisation as a latent individual trait results in an expansive definition of skill, covering also so-called "soft", "non-technical", or "non-cognitive skills", which, depending on the definition (for a systematic review see Cinque et al., 2021), overlap with individual personality traits.

The fifth and final usage of the term skill is as a *metonym for professional qualification, occupation, or industry*. This expansive usage is sometimes employed in policy discussions in relation to shortages of qualified personnel (e.g., "shortage of nursing skills"). It can also be used to refer to a broad range of occupations and industries (e.g., "green skills", or the jobs and processes affected by environmental policy).

These different definitions have sufficient overlap between them that they are usually mutually intelligible, but are also sufficiently distinct that they do not always coincide. At a minimum, they share some essential common features in their conceptualisation of "skill": it is an attribute of the individual, valued in the world of work, which can be learned. This makes "skill" a positive, empowering concept, but whose boundaries are sometimes imprecise, which in turns limits its use as a measurable form of human capital that can be accumulated and expended. How can individual credibly claim to possess any individual skill – distinct and unbundled from education or professional qualifications – and even have it recognised and certified, when the definition is so blurry? To address these problems recent developments in the private sector and in public policy attempt to reify the concept of skill, as anything that can be vouched by co-workers and previous employers (as LinkedIn does), and which can be certified officially through "micro-credentials".

In this perspective, it is natural to measure skills based on the way that employers describe them, and hence to collect and analyse data on online job advertisements. This new data source, regularly collected by aggregators, can serve as a useful complement to fill the gaps of official labour market surveys, for instance as labour market conditions changed rapidly during the early stages of the COVID-19 crisis (Forsythe et al., 2020). However, perhaps the main appeal of such data for researchers is that it can provide "skills intelligence" – rich information on the evolution of jobs, in terms of their task content, or skill and education requirements. Online job ads generally describe in broad terms what the job on offer entails in terms of tasks and responsibilities and what skills and qualifications are needed or expected from candidates.

In the past few years, several academic contributions have used online job advertisements for assessing the evolution of job contents and requirements, using datasets compiled by Burning Glass Technology. Grinis (2017) used it to compare the changing skill requirements between jobs requiring a background in science, technology, engineering, and mathematics (STEM). Hershbein and Kahn (2018) relied on this data to track changes in skills required across occupations, to uncover evidence of increased job polarization. The breadth and depth of this data has also been used to measure the spread of new technology, in particular Artificial Intelligence (AI), and its effects on the changing nature of jobs. Furman and Seamans (2018) and Felten et al. (2019) used job descriptions observed from Burning Glass US data to measure how different occupations are exposed to AI. Acemoglu et al (2021) also used BGT data in the US to measure whether companies advertised positions in the field of AI, finding that so-called "AI-exposed" establishments went on to expand "AI positions" further, at the expense of "non-AI" ones – though on a modest scale. The authors detect no aggregate labour market consequences, considering that AI adoption remains limited. More recently, Cedefop has been collecting online job advertisement data from across the European Union through its Skill-OVATE platform, significantly expanding the number of countries and languages for which this type of data is systematically aggregated (Cedefop, 2019d).

However, despite its growing usage in the literature, it is still unclear whether online job advertisements data provides an accurate measure of what people actually do at work. Although De Pedraza et al (2019) and Cammeraat and Squicciarini (2021) found that the trends of online job advertisements data broadly reflect that of official vacancy statistics, they need not represent occupations uniformly. Some sectors and occupations – like information and communication technology – are more likely to advertise on online platforms, as discussed further in Section 5. Moreover, job advertisements are a one-sided and aspirational description of what a job entails: employers construct them to appeal to prospective employees by describing the attributes an ideal candidate should possess. Employees who end up filling the positions may not possess all the required qualifications or skills, and their duties may be different from those described in the advertisements. We will discuss all these possible sources of bias in the extent to which online job advertisement data reflect the actual contents and skills requirement of jobs in the next section.

In our understanding, a *skill* is *the ability to do a specific task* (Rodrigues et al. 2021), which is closest to the first usage documented above. Especially when skills are measured at a high level of detail – as is mostly the case in online job ads databases – there can be a one-to-one correspondence between the concept of skill and the concept of task at work. In other words, if a job ad requires the skill of “calculating prices, costs or budgets”, we can infer that the job in question involves “calculating prices, costs or budgets”. When we refer to specific skills or tasks, the two concepts can almost be considered as synonymous. We can thus generally expect data on skills compiled from job advertisement databases to reflect the task content of occupations as well, though it may also suffer from similar biases, which we discuss in this paper.

The tasks approach to labour market analysis is still a relatively new field in labour economics, which uses some concepts long established in the sociology of work literature – such as *routine*, *deskilling* or *job polarisation*. Analytically, *tasks* are defined as discrete units of work in economic processes. Different tasks require different skills, and tasks can be classified accordingly. The task approach is especially useful to analyse the impact of technology on work (Autor 2013; Fernández-Macías and Bisello 2020). Because of this, this approach is central to some key academic and policy debates on the changing nature of work and the recent impact of technical change on employment. For instance, the debates on job polarisation and upgrading (Goos, Manning and Salomons 2014; Fernández-Macías and Hurley 2017), and on the debate on the impact of artificial intelligence on work (Frey and Osborne 2017; Tolan et al. 2021).

In a recent paper, Fernández-Macías and Bisello (2020) make a detailed review of the literature on tasks and employment to identify the main types of task content analysed, and to propose a comprehensive taxonomy of tasks for labour market analysis (the *JRC-Eurofound tasks taxonomy* from now on). This JRC-Eurofound tasks taxonomy connects the material *content* of work (*what* people do at work, with tasks classified according to the type of object and transformation processes), with its *methods* (which determine *how* people coordinate their work, classified according to the main axes of work organisation), while also considering the *tools* used (namely machinery or digital technologies, classified by specificity and complexity). In a subsequent paper, they construct an European Tasks database of task indicators, which implements the JRC-Eurofound tasks taxonomy across the European Union, using several surveys on working conditions, skills and occupational contents (Bisello et al. 2021). This includes standardised scores of task intensity for each of the types of tasks classified in the hierarchical taxonomy, computed for detailed occupations (3-digit ISCO codes) and detailed occupation-by-sector combinations (2-digit ISCO combined with 2-digit NACE).

In this paper, we map the detailed skill indicators of the BG database with the lower-level indicators of the JRC-Eurofound tasks taxonomy through a *Skill-Task Dictionary* developed for this purpose. The basic idea behind this mapping is that, as previously mentioned, there is a one-to-one correspondence between skills and tasks when they are defined at a sufficiently detailed level (as is the case). This allows to map all the relevant (detailed) skills contained within the Burning Glass database to the fundamental dimensions of labour input according to the JRC-Eurofound hierarchical taxonomy of tasks. Thus, the JRC-Eurofound tasks taxonomy can be used as a bridge to compare the information contained in the online job ads database and the information contained in the JRC-Eurofound European Tasks Database. These two databases are from independent sources: the BG database comes from the processing of free text contained in online job ads, whereas the JRC-Eurofound European Tasks Database comes from the collection of information via survey questions made to workers. To compare the attributes of occupations across different sources, we converted and aggregated the 4-digit UK Standard Occupation Classification into 2-digit ISCO codes, which entails some loss of precision, and disregards the sectoral dimension present in the European Tasks Database.

Overall, comparing the task content of the same occupations across these two different datasets should provide a good method for externally validating the information contained in both. To better understand the task information contained in the online job ads database, identify its potential biases, we first need to consider the information-exchange process in which advertisements are constructed, as explained in the next Section.

# 3 A theoretical framework on the information content of online job advertisements

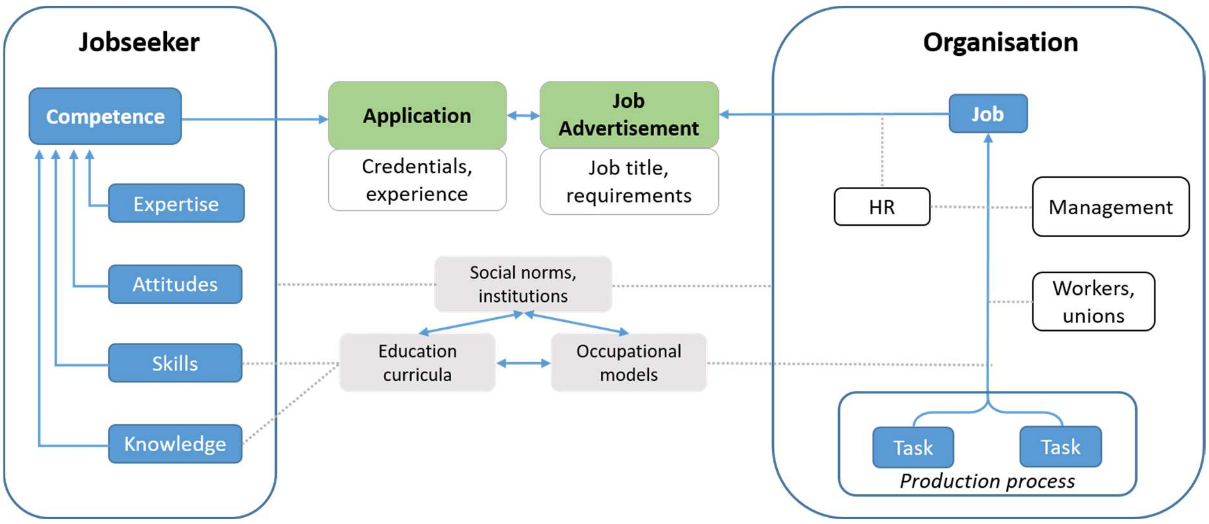
To understand the value and limitations of job advertisements as a signal for the demand for skills and task contents of the jobs advertised, it helps to deconstruct the single posting as communication medium. Job advertisements, whether published online or not, are a form of communication by which a prospective employer sends a signal to an audience of potential employees. Candidates who respond do so spontaneously, based on their desire to work for the employer in question, and their perceived affinity with the required level of skills and ability to carry out the tasks described in the posting.

Adopting the material perspective proposed in Fernández-Macías and Bisello (2020; see also Rodrigues et al. 2021), the diagram in Figure 1 shows a stylized illustration of the process of producing and responding to job advertisements. The right-hand side describes the relevant concepts at play within an organisation producing a job advertisement. At the technical level, a job consists of a coherent bundle of tasks (discrete units of labour) contributing to a production process, each task requiring one or more skills to be executed. This description is consistent with the evolutionary theory of the firm, where skills are understood as the individual equivalent of "organisation routines", or "smooth sequences of coordinated behaviour", effective in the context of the organisation (Nelson and Winter, 1985). This quasi-programmatic definition of skill, far from reducing individuals to cogs into an organisational machine, actually recognises the individual "tacit knowledge" (Polanyi, 1967), context-awareness, and ability to make the subtle choices required to carry out tasks in a production process and fulfil organisation capabilities. While the procedure of advertising and filling a position clearly varies depending on the size and complexity of the organisation, it is helpful to consider a case where such processes are relatively structured, to describe all the potential agents that can shape a job advertisement. First, although the bundling of tasks into a job is surely driven by the technical concerns of the production process, it is also shaped by organisational social factors, including the vision and needs of management, those of human resources departments, and of other employees. Potentially none of these actors has a complete and unbiased conception of what the job actually entails, because skills in complex production processes are often tacit. In some cases, the precise set of tasks that are bundled in a job is formally negotiated and shaped by worker's unions – either at the organisation or at the sector or national level – setting out the tasks that workers in a particular job are (and in some cases, are not) expected to do (Damelang et al. 2019).

Organisations do not define jobs in isolation; they search and compete for employees in a broader labour market, so their definition and bundling of a job are also influenced by occupational models defined at the social levels. These are abstract, idealised conceptions of what an occupation (like "Accountant", "Teacher", "Cook") entails. They are also standardised at varying level of detail in occupation classifications like ISCO or the UK SOC. These occupational models are affected in turn by educational curricula – which contribute to defining occupations by instructing them – and by broader social norms and institutions, including traditional historical conceptions of occupations, but also class, gender or race dynamics (see for instance Cohen 2016; Witz 1990). When creating an advertisement for a particular job, Human Resources (HR) departments may take into account all these factors, within the organisation and outside it. The resulting advertisement is a message that needs to communicate the requirements and context of the organisation to jobseekers, using a terminology that both understand. To do so, job advertisements prominently include a job title – which can already be quite informative on its own – and additional information on the task contents of the job, and the skills that these require (they may also include information on the conditions of employment, the location, etc.).

Clearly, the message contained in a job advertisement cannot be a complete and unbiased description of the contents of the job in question. In most cases, a full description of the task contents and skill requirements of a job would be too long for a job advertisement, and thus it will necessarily focus on those aspects considered most salient or important for the job, leaving out many implicit aspects. In particular, we expect this description to focus on the contents of the tasks (what the job does, materially) and its tools (how it is done), but to underplay specific methods of work that employees find undesirable, such as repetitiveness and standardisation, control or lack of autonomy. Job ads may also present a biased view of their tasks because Human Resources wants to curate and preserve the image of the firm as a desirable employer for all future positions, for instance emphasizing desirable attributes (such as communication, creativity or teamwork) irrespective of how important they are to the job at hand. Job advertisements are also economic instruments in markets characterised by imperfect information, so they can act as tools of screening and selection, for instance by inflating the education and skill requirements, in order for prospective employees to self-select based on their competence. Job ads can also disclose relevant information selectively, such as salary, or even the identity of the employer, to allow employers to retain an informational advantage with respect to prospective employees and the competition. All these aspects would make the information contained in job ads a biased indicator of the contents of the associated jobs. If we want to use job ads as a source of information on the attributes of jobs (such as task content, skills demand or working conditions), these possible sources of bias have to be taken into account.

*Figure 1: Task-based deconstruction of job advertisements*




On the other side of the transaction, individual jobseekers are characterised by a certain type and level of competence, which in our understanding is a function of skills, knowledge, attitudes, and expertise (Rodrigues et al. 2021). Among these attributes, a significant part of the skills and knowledge is determined by education curricula, which are influenced in turn by occupational models. For example, vocational education and training institutions often aim to provide the bundle of skills that is best suited for – or defines – a specific trade or occupation. Social norms and institutions can also shape the individual directly, most notably through attitudes. When applying for a position, job seekers are assessed based on their competence, as attested by their credentials, which certify their professional experience, their educational and professional qualifications. However, credentials are an imperfect signal of competence – and ultimately skill in the job at hand – which is why employers rely on multiple sources of information to assess it. For our purposes, we are interested in the fact that prospective employers and employees may conceptualise “skills” differently – coming from the demand and supply perspective, respectively – and thus need to agree on a common terminology. An ideal job advertisement should thus provide sufficient information to describe a job, such that prospective employees can decide whether it is appealing and suited to their competence.

Probably the most informative part of a job advertisement is the occupational title (or “job title”). In some cases, an advertisement may include only a few additional details on the position beyond the job title. In other cases, the job title may be an original – or even unique – expression to describe a particular position, which may or may not depart from a conventional occupation model. On balance, the level of detail of the job ad will vary across occupational titles, depending on two main factors. First, the level of detail of job ads beyond the occupational title will depend on how "standardised" the specific occupational title is. Occupations that are clearly linked to a particular educational curriculum or that are neatly demarcated in public imagination can be better described by the occupational title. This compares to occupations that are not linked to an educational title, because they are newer or less well defined (compare “nurse” with “sales assistant”, for instance). This degree of standardisation of occupational titles not only varies across occupations, but also across countries, which makes it a potential source of cross-country bias in the information content of job ads. Countries with more institutionalised occupational models (for instance, where the vocational education system standardises trades occupations; or where collective bargaining covers occupational demarcations) will have more standardised occupational titles and thus the job ads may tend to be less detailed. On the contrary, countries whose educational system is less focused towards pre-defined occupational titles or where professional associations are weaker may have less standardised occupational titles and thus the job ads may need to be more detailed. Secondly, the level of detail of the job ad beyond the occupational title will also vary depending on how close the match between the specific position to be covered and the relevant occupational title is. In some cases, the position may be so close to the standard contents and skill requirements of an occupational title that it is not necessary to add any further detail. But in other cases, even if the position broadly matches a specific job title, it may involve a significant number of tasks that are not typically associated to that title and thus require separate specification (to the extent that the organisation thinks specifying those tasks or the associated skill demands is important). For instance, even if nurse is in most cases a well-defined occupational title, associated with a well-defined educational curriculum, a specific organisation may want to hire a nurse that can also provide health and safety training to co-workers, prepare reports, or manage part of the company website. To the extent that those tasks require skills that are not typically expected from a nurse, the employer may want to specify them separately in the job ad. Since the definition of jobs are ultimately organisation-specific, the degree of match between society-wide occupational titles and the specific positions being advertised will mostly depend on organisational factors. In small organisations, the division of labour is shallower, and positions tend to be less strictly demarcated according to occupational titles. In big organisations, the labour process is more subdivided and positions are more likely to fit existing occupational models, although in some big organisations they may have sui generis occupational models that may not be known outside and thus may need detailed description in job ads.

Overall, our framework assumes that job advertisement are fundamentally a one-sided form of communication. They express the preferences of employers, based on the specific requirements of the position within the organisation, but are also influenced by prevailing market conditions, as well as social norms and conventions, for instance about the specific bundle of tasks that make up a given job. As such, advertisements are inevitably "wishful" messages: they purport to describe the requirements that prospective employees should have – in terms of education, qualifications, skills and previous experience – and often include a prospective salary. However, by observing the advertisements in isolation we have no way of knowing whether the position is eventually filled or not, and whether it is filled by a person with all the required attributes, nor whether the salary is the same as advertised. In addition, job advertisement may often omit important requirements because they are implicit in the context, or can be safely assumed based on social norms. For instance, job advertisements typically do not explicitly require basic literacy as a skill, even when it is crucial to the job in question. This is because literacy can safely assumed – not least, because the advertisement is itself in writing – or because other educational or professional qualifications require literacy themselves. In many instances, advertisements tend to include the most advanced skill or education requirements necessary and omit those that are implied or subsumed by the former; for instance, a university degree implies secondary education. This feature is efficient in an information-theoretic sense, insofar as messages convey useful information by including content (such as skill, education or professional qualification requirements) which is "surprising", meaning that it cannot reasonably be assumed from context. This tendency is not universal, however, and the degree to which advertisements explicitly mention requirements can vary by employer, occupation, industry, and country. Because of this variability, job advertisements are not always well suited to establish the prevalence of minimum skill or qualification requirements, but are more informative about the most advanced skills. The next section explains how information is extracted from the text of online job advertisements and turned into structured data.

# 4 Data and methodology

## 4.1 The Nova UK database from Burning Glass Technologies

Online job advertisements are found across many different websites in a variety of formats. They are addressed to job seekers, and are thus presented in a format – plain-language text – that is fundamentally unstructured and difficult for machines to collect, parse, and store in structured datasets. Commercial providers like Burning Glass Technologies (BGT) develop complex technical infrastructures and commercial agreements to scour online job portals, capture the text of job advertisements, and compile them in structured datasets, containing information on job titles, occupations, employers, skills required and job descriptions with varying degrees of standardization. In general, the quality of the resulting data depends on the combination of two factors: the amount of relevant information available in the original advertisement text, and the ability of providers like BGT to identify and classify relevant data correctly on a large scale. This paper examines the Burning Glass NOVA UK dataset, which covers over 60 million individual job advertisements in the United Kingdom over the period January 2012–January 2020, but most of the issues discussed in this section also apply to other large-scale collection of online job advertisements.

The crucial feature of online job advertisement data is its scale. The sheer volume of online job advertisements makes it infeasible to examine and process them manually, which means that the bulk of the processing is done automatically by a collection of purpose-built algorithms. In the case of BGT, some parts of this classification uses supervised machine learning, which is trained (and periodically subsequently validated) on a subset of advertisements classified by human experts, and then deployed algorithmically on the rest of advertisements collected. To process online job advertisements at scale in a timely manner, Burning Glass thus combines human knowledge and judgment with algorithmic decision-making (which can be described as "Artificial intelligence"). However, this combination introduces some technical challenges that can affect the quality, resulting from limitations either of the original advertisements, of the technology available, or of human expertise. An example of data limitation is the fact that job advertisements may omit valuable information like the salary, the identity of the employer, or describe the position in few words. This is a flaw in the original data that no amount of further data collection – or indeed human or artificial intelligence – can reasonably overcome, and leads to missing values for many important fields, as presented in *Table 1* below. Another, perhaps unexpected technical challenge is ensuring that advertisements are not accidentally counted twice during the so-called “ingestion” process of online text, when they are posted on different platforms or even observed at different times. Detecting duplicates would probably not be an issue for a human analyst operating on a small sample, but for technical reasons it is non-trivial to automate at scale, because in practice even duplicate advertisements are rarely identical – though robust tools can ultimately be developed.[1] In practice, automatic de-duplication relies on carefully calibrated text similarity metrics, which are complicated by the fact that some advertisements include a lengthy description of the employer in the advertisement itself, which is often identical across advertisements.

A different technical challenge is to identify correctly the relevant data fields from job advertisements, such as job title, industry, salary, contractual conditions, education and professional qualifications, or skills. Again, these would already be difficult tasks for human classifiers, but are particularly challenging to automate, in part because algorithms have no built-in understanding of the meaning and context of words. Job titles typically appear in the header (title) of advertisements, though that may also include extraneous information, like date of posting or industry, so the text requires some cleaning and standardisation (like removing some punctuation and normalising capitalisation and spelling). Despite this effort in cleaning job titles, the NOVA UK dataset reports approximately 15 million distinct job titles[2] over 60 million advertisements, meaning that a plurality of job advertisements have a globally unique job title, and that the most common job titles has a frequency of just 0.23% across all advertisements.

Some fields like dates or salary can be identified by their distinctive formats using so-called *regular expressions*, which identify patterns of digits, letters, or symbols, including currency symbols. Most other fields, like the type of contract on offer, education requirements or skills, need to be more explicitly indexed by providing a list of keywords to look for in the text. Although this index can be complemented with the use of natural-language processing techniques (as discussed in the next subsection) at a fundamental level, any information extracted from the text of job advertisements has to be explicitly looked for in terms of a curated index (or *ontology*), and anything not explicitly included in the index is ignored. This methodology has profound implications for the information content and representativeness of the resulting data because indexing inevitably entails some forms of bias, for instance when specific keywords for skills are not included in the index (intentionally or otherwise), or because the phrasing in the advertisement makes them harder to detect, as discussed in the next subsection.

Finally, an issue that combines limitations in data, technology, and human judgment is the classification of the advertisement in terms of standard occupational categories, such as the International Standard Classification of Occupations (ISCO) or the UK Standard Occupation Classification (SOC). Standardised classifications of occupations are needed for analytical purposes, especially considering the sheer variety of job titles observed “in the wild”. In principle, the job title alone should be sufficient to classify a job title into a standard occupation category. Indeed, the UK National Office for Statistics curates an occupational coding index[3] and an interactive coding tool[4] based on job

[1] Different online job platforms may display the same text slightly differently (in terms of layout as well as content), report different publication timestamps, or embed distinct platform-specific internal references identifiers. Therefore, one cannot rely on two advertisements captured across different websites (and sometimes in the same website, observed at different times) to be identical in terms of their binary or text content, which complicates automated de-duplication.

[2] This figure is partially inflated by the exacting requirement of distinguishing unique text strings, where even minor variations in spelling or punctuation imply that two otherwise similar text strings are counted as different. This occurs in the NOVA UK dataset, despite the cleaning process.

[3] https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010/soc2010volume2thestructureandcodingindex#the-job-title-coding-index

[4] https://onsdigital.github.io/dp-classification-tools/standard-occupational-classification/ONS_SOC_occupation_coding_tool.html

titles. We understand that Burning Glass uses a custom machine-learning algorithm to classify the large number of job titles collected online. As discussed in Section 6, this algorithmic approach, although necessary to operate at this scale, is also prone to errors, different in kind to those that a human analyst would make. Although this may introduce some biases, for the purposes of this paper, we will assume that the occupational classification in the NOVA UK is mostly correct, but we invite greater transparency and independent validation on this matter.

*Table 1* provides an overview of the completeness and data quality of the NOVA UK dataset. Among the 60 million advertisements included between January 2012 and January 2020, all of them have a job title and nearly all are classified in terms of a standard occupation code. By contrast, only a third of advertisements report the name of the employer, but the industry is still identified in half the cases. This is consistent with common practice for job advertisements in the United Kingdom, which often omit the name of the employer, but may still specify the industry. The salary offer (either starting salary or range) is detected in around 60% of advertisements, which compares favourably to the figures from the United States, but is still missing for a plurality of observations, possibly correlated with industry and occupation, which limits the usefulness of this variable. Information on education requirements, qualifications, or experience is detected only in around 16% of advertisements. This may be under-reporting stemming from the inability of Burning Glass' algorithms to detect these attributes in the original text. However – assuming the magnitude is broadly correct – these figures suggest that UK employers often do not explicitly require minimum education levels, qualifications, or prior experience. Finally, the variable "Skills" (present in around 90% of all ads in the database) contains additional information on the attributes required in the advertisement – not included in educational or professional qualifications – and is described in the next section.

*Table 1: Data completeness in Burning Glass NOVA UK dataset, January 2012– January 2020*

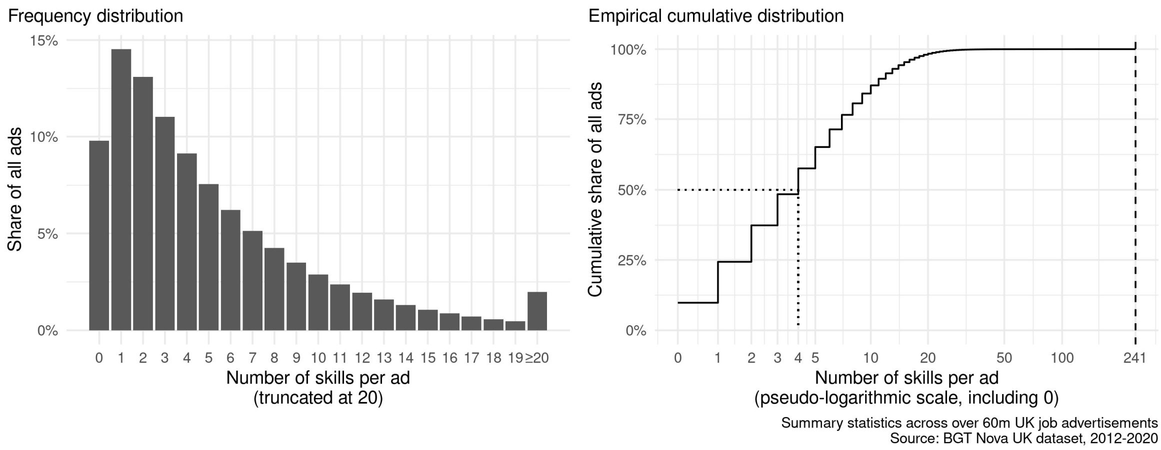
| **Variables in data (*topics*)** | **Coverage (% of ads)** | **Comment** |
|---|---|---|
| Job title | 100 | Fairly clean, 15m distinct |
| (*Occupation*): UK SOC (4-digit), BGT code: group/career area | 98.7 | Unknown accuracy |
| (*Employer*) name | 33 | Fairly Clean |
| (*Location*): City, Language, Country, Nation, Region, Travel-to-work Area, Coordinates, Local authority | 78 | |
| (*Sector/Industry*): UK SIC Code | 51 | Fair: 1–4 digits |
| (*Education*): Min/Max/Preferred ed. level, major/license | 17 | |
| (*Certifications or licences*): UK NQF levels, years of experience required (min/max) | 16 | |
| (*Contract/conditions*): permanent/temporary contract, working hours, internship, work from home | 75-100 | |
| (*Salary*): (min-max range salary range; hourly salary) | 62 | |
| (*Misc.*): Company stock ticker, local enterprise partnership | Very low | |
| (*Skills*)<br>▷ BGT skill classification (cluster/family);<br>▷flags: foreign language, baseline, specialized, software | ~4/ad<br>70 inst.<br>100 inst. | |

## 4.2 Extracting data on skills from job advertisements

This section presents an overview of the procedure for identifying and classifying skills from job advertisements, using the Burning Glass Nova UK data as an example. Fundamentally, extracting skills from online job advertisements means comparing the text of the advertisement with a curated listing of keywords that denote different skills. In the case of Burning Glass, this “Skill Taxonomy” is developed through a combination of keywords initially provided by analysts, and others detected algorithmically as salient keywords. The latter group can be identified, for instance, using Natural Language Processing (NLP), which generally involves splitting text into unstructured chunks of words by disregarding grammar, sentence structure, punctuation and capitalisation. Comparing the statistical properties of word co-occurrence between the text of job advertisements and a reference corpus of natural text (e.g., like Wikipedia pages) reveals which keywords are more often used in job advertisements, and are more likely to describe something important, like skills. Ultimately, this algorithmic approach also requires some manual review to exclude irrelevant or uninformative terms, and to group together keywords spelled differently, or even synonyms. This last step inevitably entails a form of value judgment regarding which keywords are relevant or not, and which are distinct or not. In the Nova UK data, “detail-orientated” (sic.) encompasses similar phrases, including the more common spelling “detail-oriented”, “focus(ed) on detail(s)” or “eye for detail”. Hypothetically, “spreadsheets” may be considered equivalent to “Microsoft Excel”, but is “office organisation” always the same thing as “office administration”? It may depend on the definition and context. In summary, the procedure to extract skill information from online job advertisements requires both that the skill be explicitly mentioned in the advertisement and that it (and its variant spellings) be deliberately inserted in a list of keywords to look for.

The Nova UK dataset on average identifies only a handful of skill keywords in each advertisements. The distribution of number of skills per advertisement over the period 2012-2021 in Figure 2 shows that the most common case is an advertisement listing a single skill, and nearly 6 million advertisements (or 10% of the total) feature no skills at all. The cumulative distribution (right) shows that the median number of skills identified for each job ad is less than four. This can be explained either by the fact that the original advertisement did not in fact list many skills, for instance because employers consider that job title or qualification convey sufficient information, or by the inability of BGT’s algorithms to detect any skill keywords in the text. At the other end of the spectrum, a small number of advertisements feature over 100 skills, up to a maximum of 241.

*Figure 2: Distribution in number of skills per advertisement in Burning Glass Nova UK data, January 2012-January 2020*




To some extent, the low average and extreme skewness in the number of skills per advertisement may reflect some genuine differences in the length and level of detail of the underlying original advertisement text. It may really be the case that some jobs require a larger number of skills than others, and that this would be reflected in the text of the advertisement, and hence in the compiled data. However, the distribution observed in the dataset may also be distorted by the methodology used to index skills, which may ignore some keywords, such as those describing less glamourous jobs. It may also inflate the count in some cases, for instance when skill keywords appear in the description of the company, rather than in the description of the position being advertised.

From a purely statistical perspective, this distribution means that each observation in the data, representing a single job advertisement, contains relatively little information on skills, which is certainly a limitation of the dataset. However, we may obtain more information about skill on aggregate, by pooling together all the many advertisements of specific occupations of job titles, across employers and over time. This approach requires identifying consistent occupation classes, which has limitations discussed in section 5.

## 4.3 The Burning Glass Skill Taxonomy

The Nova UK distinguished between over 13,000 distinct "skills" keywords. However, many of these do not necessarily correspond to our definition of *skill* – understood as the ability to carry out a specific task, as defined earlier. Some keywords refer **to modes of communication and interactions**, such as "People management", "Building Effective Relationships", "Persuasion", or "Listening". Others are umbrella terms for **broad sets of skills** (presumably the result of BGT aggregating keywords that it considered equivalent**: "**Communication Skills", "Organization Skills", "Computer Literacy". Again, these do not point to a specific task, but at most to a broader task domain. A third collection entails **personal attitudes or character traits**, like "Energetic", "Self-starter", "Positive disposition", "Self-Motivation", or "Detail-orientated". There are also terms referring to **knowledge**, whether **broad industry knowledge** such as "Accounting Industry Knowledge", "Media Buying Industry Knowledge", or **specific knowledge** of organizational processes or types of work organization, like "Key Performance Indicators (KPIs)", "Lean Manufacturing", "Six Sigma", "Scrum", "Cisco", or "ISO 9001 Standards". Among these, those in the healthcare sector are particularly numerous, and can refer to specific conditions, like "Dementia knowledge", or broad areas of practice like "Midwifery", "Paediatrics", or "Elder care. Some keywords refer to **task domains**, close to being occupational titles, such as "Sales", "Budgeting", "Teaching", "Administrative support", "Customer Service", or "Engineering Design". Although each of these concepts refer to duties, which, in turn, may require one or more skills, they are not point to atomic tasks in themselves. Others, however, refer directly to the content of work, like "Scheduling", "Filing", "Planning", "Weeding", "Editing", "Budget Forecasting". Then there are some terms related to methods of work, like "Teamwork / Collaboration", "Time Management", "Multi-Tasking", or "Prioritising Tasks". A significant share of items mentions specific work **tools**, especially software applications like Microsoft Excel or Microsoft Office, and a panoply of programming languages and frameworks like JavaScript, Java, .NET, or Python. They also include other non-digital office tools (Printers, Scanners) and machinery (Lathes, Machine Tools, Grinders, Milling Cutters).

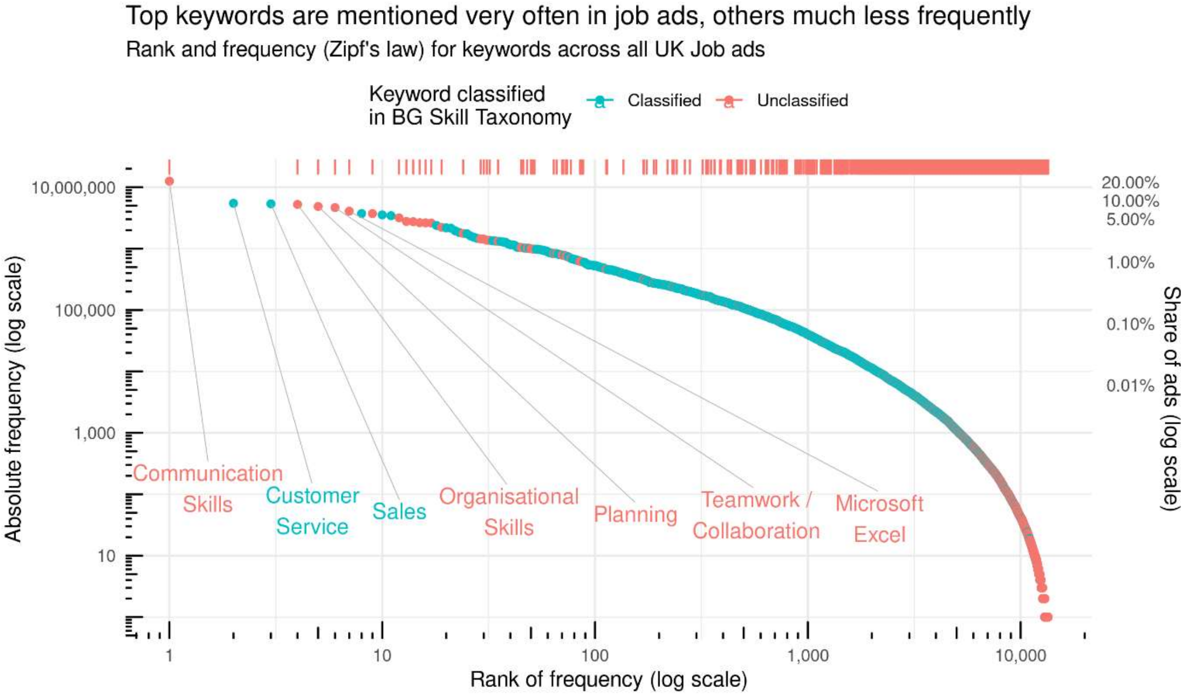
Notwithstanding these conceptual distinctions, BGT classifies the so-called "skills" individually – that is, in separate tables, independent from the context of the job ad in which they appear – in terms of different attributes, into what it calls the Burning Glass Skill Taxonomy:

- **Baseline or Specialized**: are two distinct binary flags indicating whether, according to Burning Glass' assessment, the skill in question is *basic* – 130 of them in total, which are not necessarily among the most common – while the remainder are flagged as *specialized*. There are about 60 skills flagged as neither baseline nor specialized, likely due to error.
- **Software**: this binary flag indicates whether the skill refers to a specific software or programming language. About 1,200 skills are flagged as software skills. Among these, there is still the distinction between *baseline* (e.g., *Microsoft Excel*) and *specialized* (e.g., *JavaScript*)
- **Language**: a binary flag indicating whether the skill in question is a (foreign-)language skill, like *Arabic*, for a total of 33.

- **BGT Skill Cluster** and **BGT Skill Cluster Family**: this is Burning Glass' own hierarchical skill classification, developed in two levels (see Annex 5). The broadest levels distinguishes between 29 *skill cluster families*, further divided in 539 *skill clusters*. However, this classification only covers about 7,300 items in the data, about half the total.

These additional classifications are helpful, especially as it helps to identify software skills. However, as Figure 3 below demonstrates, many common skill keywords in the data are unclassified. The figure plots the overall frequency distribution of skill keywords across all advertisements in the period 2012-2020. On the horizontal axis, it ranks the frequency of skills from most to least common, on a logarithmic scale. On the vertical axis, it shows the number of different advertisements that mention that specific skill (left vertical axis) or the share of total ads in which the skill in question appears (right vertical axis). *Communication skills* are the most common ones, mentioned in over 10 million different ads, or about 20% of all ads. The frequency distribution of skills is heavily skewed, meaning that about a third of them (4,700 skills) are mentioned in fewer than 100 advertisements, out of 55 million.

*Figure 3: Frequency and rank of keywords across posting*




Thus, the BG Skill Taxonomy is essentially a set of criteria for classifying the very long list of specific skill keywords (over 13,000) into some broad categories by their specificity and content similarity. Aside from its limited coverage (less than half the available skills are thus classified), this taxonomy has the problem of not being driven by any theoretical model or conceptual framework. Therefore, this taxonomy cannot be used to systematically describe the skills demanded by employers according to the online job advertisement data. Taking advantage of the conceptual correspondence between skills and tasks, in this paper we propose using an existing theory-driven taxonomy of tasks (the JRC-Eurofound Task Taxonomy) to give some structure to the long list of skills identified in the BG database. As previously discussed, this taxonomy is built upon a conceptual framework that connects the material content of work with its methods, while also considering the tools used. In principle, this taxonomy can be used to classify all the possible types of tasks that can be performed within productive organisations, and thus also all the skills which are necessary to perform those tasks. The next section discussed the representativeness of occupations found in online job advertisements, while Section 6, explains how we systematically mapped over 1,600 of all the skills contained in the BG data (corresponding to 87% of the skills mentioned across all ads) into the corresponding categories of the tasks taxonomy, at the most detailed level possible.

# 5 Online job ads as representation of the labour market

The value of the information contained in online job advertisements data depends on whether it correctly describes the composition of the labour market. Because not all positions are advertised, and many are not posted online, some occupations or sectors may end up being less represented in online job advertisement data. Of course, we do not expect the number of online job advertisements to represent a scaled representation of the current employment structure, for several reasons. The first is that job advertisements, as signals of vacant positions, represent prospective incoming occupation flows, or changes in the composition of the existing employment stock. Jobs with short-term contracts or with higher turnover are advertised more frequently than other, longer-tenured ones. The second reason is that not all jobs are advertised online. The third is that not all online job advertisement are necessarily captured by Burning Glass, which may miss some positions posted on lesser-known portals, or on websites that are harder to capture. The fourth and final reason is that Burning Glass may misclassify job titles into the wrong occupation code, which artificially inflates the number of jobs attributed to certain occupations at the expense of others.

Several studies have compared online job advertisements in general – and those from Burning Glass Technologies in particular – with traditional survey-based vacancy statistics, which ask employers about open and vacant positions. They generally show that although online job advertisement data shows time variations similar to official vacancy statistics, online sources do not capture all job openings, and may represent a biased sample of new open positions, emphasising white-collar occupations, or those with higher education requirements. Carnevale et al (2014) compared data from Burning Glass Technologies from the US with official Bureau of Labor Statistics (BLS) Job Openings and Labor Turnover Survey (JOLTS) and concluded that online job advertisements represented 60-70% of all job openings at the time in the US. However, they also noted that the jobs advertised online tended to "over-represent industries that demand high-skilled [i.e., more educated] workers" and that the majority of online ads consisted of "white-collar office jobs", such as sales and customer support and managerial and professional occupations. Moreover, jobs with higher education requirements are also more likely to be posted online: over 80 percent of jobs for requiring Bachelor's degrees or better are posted online.

These properties are not specific to online job advertisements collected by Burning Glass. De Pedraza et al. (2019) compared a different dataset of online job advertisements in the Netherlands, by the company Textkernel, and compared it to vacancy data collected via survey by the national office for statistics. They found similar co-movement in the two series, which indicated that both data sources reflect the same underlying phenomenon, though online job ads were always less numerous than those sampled through traditional surveys. More recently, Cammeraat and Squicciarini (2021) found that although the trends of online job advertisement data[5] collected in 2012–2018 by Burning Glass broadly reflect that of official survey-based vacancy statistics, they need not represent occupations uniformly in six, mostly English-speaking countries (Australia, Canada, New Zealand, Singapore, the United Kingdom and the United States). They caution that data from Burning Glass Technologies is difficult to compare directly with official vacancy data and suggest comparing it with data on total employment volumes instead, even though this compares employment stocks and flows. In particular, the comparison with official data for the UK is especially flattering, as Burning Glass appears to include nearly the same number of advertisements at the ONS Vacancy Survey, and in some periods even more. They similarly conclude that the time trends of BGT data follow reasonably closely those in official job openings but are better at representing so-called "high-skilled" occupations, with higher educational requirements.

There are several possible explanations why some occupations or sectors are better represented in online job advertisements. Following the framework in Section 3, we can suppose that larger and more complex organisations have more structured job positions (especially of white-collar profes-

[5] These advertisements are called "vacancies" in the article, which does not seem to distinguish between the two.

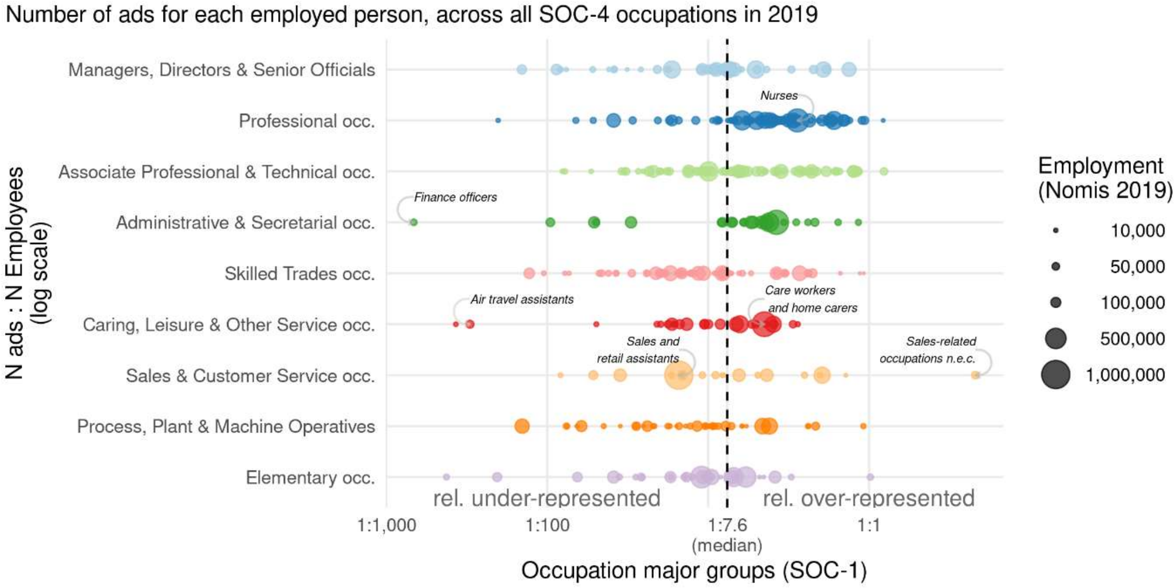
sional occupations), and more established procedures for hiring personnel. As a result, they may be more likely to advertise online, leading to professional occupations being well represented in online job advertisement. Another – partly distinct – explanation is technology. Advertisements for information and communication technology positions were among the first to be advertised online because of the digitalisation of employers and prospective employees, and to this day have a number of dedicated online portals. For this reason, we expect digital occupations, and digital skills to be better represented in the data, relative to the rest. Moreover, we also expect that employers may not find it necessary, or even desirable, to post online advertisements for all positions. Candidates for some job profiles are harder to find, which pushes employers to advertise as widely as possible, while others are easier to fill within a specific work organisation. If employers can readily tap a pool of suitable candidates, they do not need to advertise in the first place. Finally, it may also be the case that many positions are not advertised because they offer sub-standard pay and conditions (such as below minimum wage or with excessive working hours). This is most likely to be the case in smaller or family-run businesses, but likely applies to some extent to entire industries like restaurants and hospitality, and to positions for entry-level or manual jobs, such as cleaners. All these factors combined should result in online job advertisements over-representing white-collar, professional occupations with high ICT content, compared to manual and less prestigious ones.

To explore whether online job advertisements indeed favour professional occupations, we can compare the number of job advertisements in the Nova UK dataset to labour market statistics collected by the UK Office for National Statistics for each 4-digit occupation in the UK Standard Occupation Classification, through the NOMIS platform.[6] This gives us a range of values for the sampling rate of different occupations in the data, which is influenced by a combination of the factors listed above. Figure 4 below summarises, for the year 2019, the distribution of rates between the total number of job advertisements for any SOC 4-digit occupations in the Nova UK dataset and the number of people employed in the same occupation according to the NOMIS survey. This is admittedly a crude measure, which compares job flows with stocks. It is likely influenced by different rates of job churn, and by the different rates at which newly open positions are advertised compared to those that replace previous workers who left their employment. Nevertheless, in the absence of granular, economy-wide commensurable data between advertisements and vacancies, this approach allows for comparisons at detailed occupation categories, and is also used by Cammeraat and Squicciarini (2021). There is a wide range of magnitudes, so the horizontal axis is expressed on a logarithmic scale. The median value is 1:7.6, meaning that, for each job advertisements for a given occupation in a year there are around eight people employed in that occupation. The occupations to the right of the median are relatively better represented in Nova UK, meaning that there are comparatively more advertisements for each employer person. Professional, associate professional and administrative occupations tend to be well represented – relative to the median – with ratios for some occupations in these groups approaching one advertisement for each person employed. Conversely, elementary occupations, those in the skilled trades, as well as process, plant and machine operatives are relatively worse represented, with ratios of one advertisement every 100 or 1,000 people employed. The extremes of the distributions – occupations that are especially well (or poorly) represented – are probably driven by job title misclassification. For instance, the highest value – purportedly showing an occupation with four advertisements for each person employed – corresponds to SOC 7129: “Sales related occupations not elsewhere classified”. In reality, many of these are more likely to be SOC 7111: “Sales and retail assistants”, which themselves appear poorly sampled in job advertisements. This is also suggested by comparing side by side the employment structures according to the labour force survey and job ads in Annex 4, within the “Sales and Customer Service occupations”. Such a misclassification, between nearly adjacent occupation codes, is a relatively minor category error, but cautions against relying uncritically on the standardised occupational classifications included in the Nova UK dataset, which is discussed more in detailed in the next subsection.

[6] https://www.nomisweb.co.uk/

Overall, if we consider that SOC-1 major occupation groups are roughly ordered by salary and "skill" (in the derived sense of "education attainment and training"), higher-skilled and better-paid occupations seem relatively over-represented in online job advertisements.

*Figure 4: Relation between number of ads in Nova UK and employed persons in Nomis survey*




## 5.1 Job title misclassification

In general, classifying job titles observed in advertisements into standard occupation codes is challenging, in some cases impossible (see, for instance, Tijdens and Kaandorp, 2019). Burning Glass relies on an algorithmic approach to classify the job titles observed in job advertisements into the nearly 400 different categories of the 4-digit UK Standard Occupation Classification (SOC). In principle, this task is well suited to so-called *supervised machine learning*. This family of statistical procedures works by "training" an algorithm (i.e., optimising its parameters) on a curated set of job title strings previously classified validated by experts, testing it on a different set of as-yet-unseen job titles strings while optimising for accurate classification. This automated approach is necessary to code job titles on a large scale but is far from perfect. For instance, we understand that the algorithm used by Burning Glass does not account for the hierarchy of the SOC. This means that a category error between two adjacent occupations (as the case of two sales-related occupations above: 7129 and 7111, within the group SOC-71) is treated the same way as a misclassification across completely different occupation major groups (e.g., SOC-1 "Managers, Directors and Senior Officials" vs SOC-9 "Elementary Occupations"). Moreover, the professional "distance" between adjacent occupation codes varies across the hierarchy of classifications like SOC, ESCO or ISCO. For instance, there is a substantial professional dive between Civil engineers (ISCO 2142) and Mechanical engineers (ISCO 2144), although they share the same three-digit ISCO code. These fine-grained distinctions tend to be more common in occupation groups requiring advanced qualifications, and less common in so-called "low-skilled" occupations.[7] Overall, the two types of category errors clearly have different degrees of severity, but that is not something the algorithm is designed to avoid.

Moreover, by design, the algorithmic classifier assigns a 4-digit SOC code even when there is insufficient information or too much ambiguity in the job title to assign a specific code (e.g., "Team Member" or "Assistant"). This means that, although nearly all advertisements (around 98%) are assigned a 4-digit SOC, some of these may be inaccurate – i.e., their occupational code is incorrect –

[7] We are thankful to Jiři Branka from Cedefop for illustrating this point.

or spuriously precise –i.e., are assigned a specific SOC 4-digit code, when in fact the job title describes a more generic occupation grouping, like SOC 3, 2, or 1 digit.

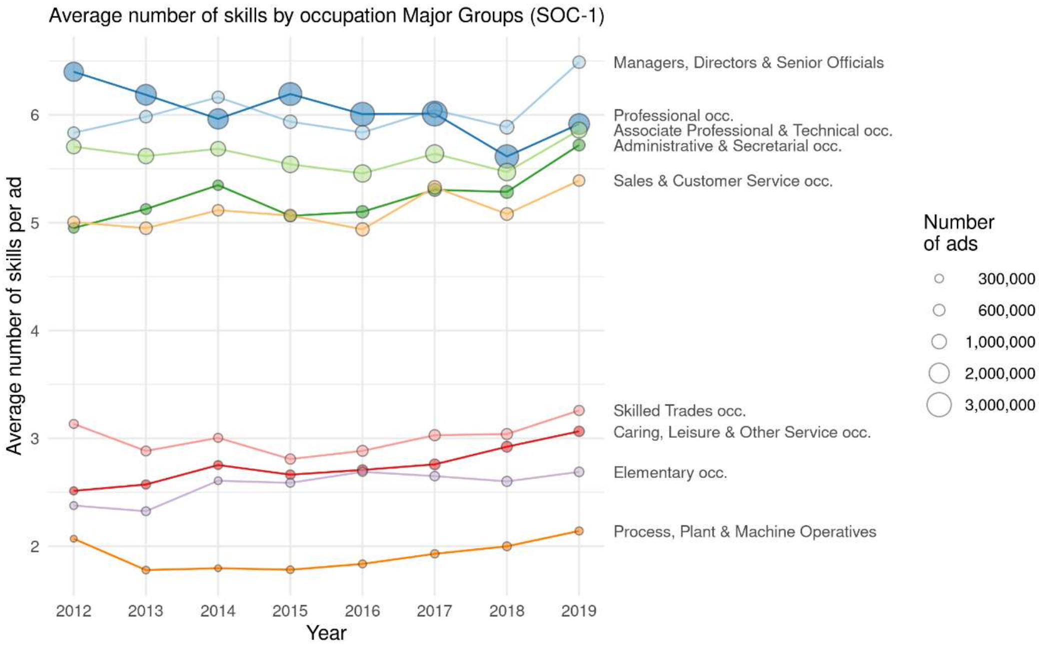
It is beyond the scope of this article to systematically validate the occupational coding of the Nova UK dataset. In many cases, it would require access to the original advertisement text, which is not included in the dataset. However, we can illustrate some of these issues, by examining a particular subset of apparent errors that we have encountered in our analysis. In the Nova UK dataset, we observe nearly 300,000 cases of the same job title being coded to different SOC codes across different observations (or coded only some of the time). This affects job titles appearing across over 10 million advertisements, or a sixth of the total. Annex 6 shows a summary of the most common job titles split across different SOC codes.

In some cases, this split classification may be warranted by the fact that the job title reported in the original advertisement is too generic and can indeed be used to describe occupations across the SOC hierarchy. For instance, "Assistant Manager" can describe occupations in different industries, each with their designated SOC code. Indeed, the NOVA UK data classifies it across 42 different SOC codes, mainly SOC 4159: "Other administrative occupations n.e.c."; 7130: "Sales supervisors"; 1223: "Restaurant and catering establishment managers and proprietors", which seem plausible. However, the same job title is also classified in less plausible codes, including 2431: "Architects", 2231: "Nurses", 5431: "Butchers". Even unambiguous job titles can be coded to radically different positions of the SOC classification. For instance, the title "Management Accountant" is sometimes classified (correctly) SOC 2421 "Chartered and certified accountants Professional occupations" but also into SOC 9120 "Elementary construction occupations". In general, the algorithmic classifier seems to be too easily swayed by the presence of modifiers like "manager", "assistant", "administrator" in job titles. Overall, without access to the original advertisement text, we rely on the occupational coding by Burning Glass, but remain mindful of the possibility that occupations may be misclassified.

## 5.2 Occupation skill quantity

A To better understand how occupations differ in terms of their task content, Figure 5 shows the average number of skills (keywords) that are identified in each posting, for different occupational groups, grouped by one-digit UK Standard Occupational Classification. Occupation groups like managers, professionals and administrative occupations tend to feature more skill keywords in the BG database than skilled trades, caring professions, plant and machine operators, and elementary occupations. The gap, and overall trends, remain relatively stable across the period 2012–2019, albeit with a minor reduction in the average number of skills required in professional occupations, though all groups show a slight uptick in 2019.

*Figure 5: Occupational trends in number of keywords mentioned per advertisement*




What can be behind these differences in the number of skills identified across different occupational groups? More specifically, what can explain the clear differentiation between the mid-upper occupational groups (five or six skills identified on average) and the mid-lower ones (2 to 3 skills identified on average)?

The most obvious difference between those two groups of occupations concerns the nature of the tasks/skills they involve. The occupations with fewer skills identified in the BG database are manual occupations in manufacturing or services, which typically do not require formal academic qualifications. Skilled trades do require credentials in some cases, but they tend to be specific vocational ones rather than formal academic qualifications. In contrast, the occupations with more skills identified in the BG database are managerial, professional or clerical occupations involved in information-processing tasks, and which typically are associated with specific formal academic qualifications.

It could be that the manual occupations with fewer skills identified in the Nova UK database indeed tend to have fewer specific skill requirements in real organisations. By extension, we could infer that they have less specific task contents. However, it could also be that the skills (and credentials) of the upper occupational categories are more formalised and standardised in the labour market. This would reflect a broader, subtle social bias in favour of so-called "high-skills": the tasks associated with white-collar employment are more valued, thus more often explicitly mentioned, and ultimately easier for Burning Glass to process. Moreover, a more detailed skill content signals the distinction between ostensibly adjacent occupations (such as the different specialisations in engineering or ICT, which can sometimes relate to a very specific technology). By contrast, it may be that the skills required in many manual tasks is less appreciated or even have undesirable connotations, and hence are less likely to be mentioned and classified.

In general, it is difficult to know whether some of the skills in question – such as "Communication skills" or "Organisational Skills" – are actual formal requirements that employees can demonstrate, and which employers actually use to decide which candidate to hire. It is also possible that these are just formulaic expressions, part of the standard filler text in job advertisements. In some instances, they may be polite euphemisms for other attributes (such as gender, class, education), which the employers would rather not write explicitly.

In any case, this systematic difference suggests that online job advertisement data is better suited to identify and monitor the skills demand of managerial, professional and clerical occupations than for manual occupations in manufacturing and services. Indeed, this is consistent with the underrepresentation of manual occupational groups in the sample of online job advertisement data that we discussed in previous pages.

# 6 Task profiles of occupations according to online job ads

In this section, we will present a novel perspective for the analysis of online job advertisement data. We will classify the full list of specific skills identified in the BG database into the corresponding categories of a theory-driven and comprehensive taxonomy of (work) tasks. This will give a meaningful structure to the very detailed but unstructured skills information of the Nova UK database, and allow to compare this database with external and independently generated data on tasks within occupations, to assess its validity.

There are some assumptions underpinning this exercise, and it is important to make them explicit. First, this exercise assumes that there is a fundamental correspondence between skills and tasks, especially at the most detailed level (see Rodrigues et al. 2020). Second, it also assumes that the JRC-Eurofound tasks taxonomy can accommodate all types of task content and skills demand (see Fernández-Macías and Bisello 2020). Third, it assumes that the occupational classifications used in the BG data and in the JRC-Eurofound tasks database are comparable (see section 5). We believe that these assumptions are reasonable, and thus this exercise is justified.

## 6.1 Mapping the detailed BG skill keywords to the JRC-Eurofound tasks database

The starting point of this exercise is the creation of a Skills-Tasks Dictionary, which maps a significant number of skill keywords identified in the Burning Glass database into the corresponding branch of the JRC-Eurofound tasks database, at the most detailed level possible. This was done manually by expert coding by the authors. In the majority of cases, the mapping is straightforward and unambiguous, because the terms used in the BG skills database are semantically close (in many cases, identical) to a specific branch of the JRC-Eurofound tasks database. However, in some cases, the mapping is not so clear, and thus we conducted an iterative process of parallel coding to solve any coding discrepancy by interactive consensus.

The process was as follows. First, we listed the BG skill keywords by order of frequency in which they appear across all advertisements in the database. Then, we randomly selected a sample of keywords (with more frequent keywords having a proportionally higher probability of selection), and coded them independently, ensuring that each keyword was independently coded by at least two different persons. Once the first coding was carried out, those keywords where there was a discrepancy were classified again after discussion among the coders. Finally, there was a third round of re-coding for the few remaining cases without consensus.

Overall, we classified 1,649 skill keywords from the Nova UK database, corresponding to 12.34% of the total number of distinct keywords identified in that database but covering 82.9% of all the skills actually mentioned in the online job advertisements available in the database, since the sampling procedure followed tended to classify the most frequently mentioned keywords. In other words, our skills-tasks dictionary only misses 17% of the skills mentioned in all the job ads available in the Nova UK between January 2012 and January 2020, many of which are very idiosyncratic skills that only appeared in a single job ad throughout the entire period.

Among the 1,649 individual skill keywords that we examined, the majority of (1,456, or around 85% of them) corresponded to clear categories of the JRC-Eurofound tasks taxonomy, either because they explicitly referred to a task (e.g., “Planning”, “Teaching”) or because they referred to a skill clearly linked to a specific task category (e.g., “Communication Skills”, which maps into Social tasks, but no further). The remaining 184 “skills” we examined did not actually refer to tasks or skills, but to other attributes of the jobs included in job advertisements that Burning Glass classified as skills. We also assigned these other keywords to a number of categories for possible later use, but we do not discuss them in this paper. The most important of these “other” categories are: job titles (“Midwifery", "Lecturer"), attitudes ("Positive Disposition", "Energetic", "Self-Starter"), knowledge ("Product Knowledge", "Physics"), productive processes ("Procurement", "Logistics"), sectors ("Civil Engineering", "Property Management"), languages ("Welsh", "Chinese") and experience ("Financial Services Industry Experience", "Industrial Engineering Industry Expertise"). Table 2 in the Annex shows the total number of different skill keywords that map in each of the categories of the Task Taxonomy, and how of often they appear in the Nova UK dataset. [8]

It is important to note that our mapping also achieved a respectably uniform coverage of the skills across different occupations: among all the ISCO 2-digit occupational codes, at least 57% of the volume of all skills mentioned for those occupations are classified strictly in terms of the taxonomy, while up to 16% we determined not to be skills, conceptually.

## 6.2 Comparing average task profiles inferred from job ads and measured in surveys

Using the Skill-Task dictionary, we can use the BG online job advertisement database to infer the task content of the different occupations and compare it with the task content for the same occupations as measured in surveys and occupational databases.

Before doing this comparison, it is important to emphasize the different nature of the measures of tasks that are contained in the two types of data. The JRC-Eurofound tasks database contains a set of standardised measures of "task intensity" for each of the elements in the taxonomy. These measures come from survey variables that measure directly the extent to which a given occupation (or job) involves doing a given type of task content. Although the metric of those indices is relative rather than absolute – it is intended for cross-occupation comparative purposes – the indices for each category can be interpreted as measures of task intensity. In contrast, with the online job advertisement data we can only count the relative frequency with which keywords linked to a given task category appear in specific occupations, and use that as an indicator of the extent to which those occupations involve that type of task content.

To illustrate, Figure 6 below shows the average values for all occupations, weighted by their employment levels, of each of the task indicators constructed from BG online job ads data alongside the values of the same indicators from the JRC-Eurofound tasks database (constructed from workers surveys and occupational databases), as well as their standard deviation. If we look at the index of "physical dexterity", we can see that it has a value of 0.096, which means that 9.6% of all job ads (measured within specific occupations and then averaged for the whole economy) contain keywords that have been mapped to "physical dexterity" (e.g, "wiring", "packaging", "cash handling"). If we look at the column for the JRC-Eurofound tasks database, we can see that the value is in this case 0.455, which means that the average occupation requires a level of physical dexterity (precise movements of and operations with hands and fingers) which is about 45 out of 100. Thus, according to the JRC-Eurofound tasks database physical dexterity is a much more prevalent type of task content in UK occupations than according to the (inferred) task content from BG online job ads data. This is probably explained by the fact that dexterity, though certainly valuable in a range of manual tasks, from using tools to typing on keyboards, both widespread and implicit. Therefore, it can be quantified in surveys that explicitly measure it, but is not referred to explicitly by employers in job advertisements

Comparing the first two columns of Figure 6 we can see how the task profile of the average worker in the UK differs according to the two sources. The task profile inferred from job ads looks rather different from the one measured via surveys. The BG task scores are generally much lower and more diverse, with a few values being as high as the corresponding JRC-Eurofound ones (even

[8] The full Skill-Task dictionary is available at https://doi.org/10.5281/zenodo.6488230 and an interactive visualisation of the hierarchy and frequency of the skills in the Nova UK database is available a https://observablehq.com/@m-sostero/skill-task-mentions

higher in the case of social tasks) but most of them being much lower. This implies that in the JRC-Eurofound tasks data there is much more overlap in the task content of the different occupations (i.e., most occupations involve many different types of task content simultaneously), whereas in the BG data the (inferred) tasks profiles are more concentrated around a few types of tasks content. In particular, the types of tasks more frequently mentioned in online job advertisements are: *social* (especially managing and selling), *intellectual* (especially generic information processing and problem-solving tasks) and use of *ICT tools*. By contrast, *physical* tasks and *work organisation* practices are much less frequently mentioned in ads than expected according to their importance in survey measures.

Note that, in the job advertisement data, the lower (more detailed) branches of the tasks taxonomy are less populated than expected. In other words, generic task categories tend to be more frequently mentioned than specific ones. It is unclear whether this is a real feature of job ads – that the skill/task descriptors tend to be generic – or an artefact of the pre-classification of the original free text of the ads into skill keywords by Burning Glass. In this case, the uncertainty in the classification of a given term can be resolved by moving up a level of generality. It may even be an undesired outcome of our mapping of the original skill keywords to the tasks taxonomy, because this uncertainty can also be resolved in some cases by decreasing the specificity of the mapping.

Figure 6 also shows the standard deviation of each task index, as well as the coefficient of variation – which simply rescales the standard deviation to the mean, to make it more comparable across indices. We can immediately see that the variability of the task indices across occupations in the BG data is much larger in most cases than in the corresponding indices in the JRC-Eurofound tasks database. In fact, in many cases the CV for the BG inferred task indices is higher than one (meaning that the standard deviation is larger than the mean), indicating a high variability in the task scores of the different occupations.

*Figure 6: Task profiles for the average job, online job advertisement (BG) vs. survey data (JRC-EF)*

| task indicator | mean | | std deviation | | coefficient of variation | |
|---|---|---|---|---|---|---|
| | BG | JRC-EF | BG | JRC-EF | BG | JRC-EF |
| ● Physical tasks | 0.114 | 0.321 | 0.127 | 0.18 | 1.108 | 0.561 |
| └strength | 0.015 | 0.187 | 0.026 | 0.159 | 1.779 | 0.850 |
| └dexterity | 0.096 | 0.455 | 0.129 | 0.204 | 1.346 | 0.448 |
| └navigation | 0.002 | 0.303 | 0.006 | 0.234 | 4.130 | 0.772 |
| ● Intellectual tasks | 0.516 | 0.532 | 0.194 | 0.191 | 0.376 | 0.359 |
| └○ information processing | 0.337 | 0.457 | 0.183 | 0.198 | 0.543 | 0.433 |
| \|—○ literacy | 0.245 | 0.49 | 0.125 | 0.215 | 0.511 | 0.439 |
| \| └business | 0.110 | 0.551 | 0.091 | 0.261 | 0.829 | 0.474 |
| \| └technical | 0.060 | 0.369 | 0.077 | 0.171 | 1.289 | 0.463 |
| \|—○ numeracy | 0.174 | 0.423 | 0.149 | 0.195 | 0.854 | 0.461 |
| \| └calculation | 0.152 | 0.468 | 0.140 | 0.226 | 0.920 | 0.483 |
| \| └analytical | 0.034 | 0.378 | 0.034 | 0.186 | 0.996 | 0.492 |
| └○ problem-solving | 0.375 | 0.608 | 0.150 | 0.196 | 0.400 | 0.322 |
| └○ information gathering | 0.053 | 0.619 | 0.071 | 0.231 | 1.326 | 0.373 |
| \| └search | 0.045 | 0.601 | 0.070 | 0.306 | 1.549 | 0.509 |
| \| └conceptualisation | 0.008 | 0.639 | 0.008 | 0.171 | 0.944 | 0.268 |
| └○ creat. and resolution | 0.328 | 0.596 | 0.127 | 0.184 | 0.388 | 0.309 |
| └creativity | 0.072 | 0.648 | 0.042 | 0.185 | 0.594 | 0.285 |
| └planning | 0.285 | 0.545 | 0.120 | 0.194 | 0.423 | 0.356 |
| ● Social Tasks | 0.618 | 0.423 | 0.192 | 0.16 | 0.312 | 0.378 |
| └serving | 0.167 | 0.53 | 0.123 | 0.256 | 0.733 | 0.483 |
| └teaching | 0.117 | 0.37 | 0.127 | 0.174 | 1.084 | 0.470 |
| └sell/influen | 0.198 | 0.471 | 0.148 | 0.191 | 0.748 | 0.406 |
| └managing | 0.280 | 0.341 | 0.130 | 0.166 | 0.464 | 0.487 |
| └caring | 0.098 | 0.409 | 0.098 | 0.239 | 0.995 | 0.584 |
| ● Autonomy | 0.041 | 0.554 | 0.025 | 0.203 | 0.618 | 0.366 |
| └latitude | 0.020 | 0.559 | 0.014 | 0.166 | 0.701 | 0.297 |
| └control | 0.023 | 0.549 | 0.015 | 0.252 | 0.654 | 0.459 |
| ● Teamwork | 0.135 | 0.59 | 0.050 | 0.112 | 0.373 | 0.190 |
| ● Routine | 0.039 | 0.413 | 0.030 | 0.125 | 0.770 | 0.303 |
| └standardisation | 0.039 | 0.486 | 0.030 | 0.139 | 0.770 | 0.286 |
| ● Machines | 0.031 | 0.191 | 0.066 | 0.227 | 2.151 | 1.188 |
| ● ICT | 0.223 | 0.42 | 0.172 | 0.249 | 0.771 | 0.593 |
| └basic ICT | 0.102 | 0.614 | 0.100 | 0.213 | 0.976 | 0.347 |
| └programming | 0.139 | 0.177 | 0.148 | 0.158 | 1.060 | 0.893 |

## 6.3 Comparing tasks across occupations

We correlate the occupation-level task scores inferred from job ads and the task scores measured in surveys, from the BG and JRC-Eurofound database respectively, in Figure 7 and Figure 8 below. They show one scatterplot for each of the task indicators of the taxonomy. The vertical axis represents the value in the JRC-Eurofound task database, and the horizontal axis the value inferred from counting the number of mentions of each task category in online job ads in the NOVA UK database. Each circle represents a specific occupation, and its size is scaled to its share of overall employment. The blue lines represent a fitted regression line, with shaded confidence intervals. A close fit between the bubbles and the line, and the width of the grey area around them, reflect the extent of the association: the consistency between the task scores inferred from job ads and measured through surveys. Additionally, the distribution of the bubbles (occupations) around the line show the dispersion and allow to identify outliers. Conversely, a flat slope with a wide grey confidence interval around it, as for instance in the index of “navigation”, indicates a low consistency between the two measures of task content. Looking at the bubbles, we can see that this is because most of the values for online job ads are extremely close to zero and uncorrelated with the values measured in surveys, with just a couple of outliers being outside the cloud of values near zero. In Annex 3, we include a complementary table with correlation coefficients.

The correlograms presented in in Figure 7 allow us to discuss the consistency in the task indicators of the two sources across ISCO 2-digit occupations, following the structure of the taxonomy:

- **Physical tasks**: in this case, the consistency between the two indices is rather poor, especially in the case of navigation as we have already mentioned. Even if dexterity and strength look slightly more consistent (with steeper slopes and narrower grey areas around the regression line), most of the values are extremely low in the online job ads inferred scores and seemingly uncorrelated from the JRC-Eurofound values. As we will also see in other cases, the upper branch (the overall physical tasks indices) is more consistent than the lower (more specific) ones.
- **Intellectual tasks** are clearly much more consistent across the two sources than physical tasks. The regression lines are generally steeper and the bubbles are more spread across the full range of values, clustering around the regression lines. This is particularly the case for the three headline indices (the three upper branches of this dimension of the taxonomy). The overall index of intellectual tasks and the indices of information processing and problem solving seem very consistent in the two sources. The lower levels of consistency between the two sources can be found in the specific task indices of technical literacy, information gathering and conceptualisation. It is important to note that in most cases, the scores for the online job ads inferred task contents are much lower than those from surveys, especially for business and technical literacy, analytical numeracy, information gathering, conceptualisation and creativity. In many cases, despite the very different ranges of scores, the relative positions of occupations remain broadly consistent in the two sources.
- **Social tasks** are quite consistent at the upper level (headline index), and in some of the specific indicators, especially managing and caring. Serving and attending shows a low level of consistency across the two indicators, while teaching and training is very concentrated in very low values according to online job ads data, with some big outliers (the biggest one corresponding to teaching professionals, who get a value of 1 in the online jobs ads data).

Figure 8 on p. 33 shows the same graphical correlation analysis for the dimensions of methods (forms of work organisation) and tools (machines and ICT) used at work:

- **Methods** are remarkably *inconsistent* in the two sources. With the partial exception of standardisation, the values for occupations are almost randomly located across the space defined by the two measures, as if they referred to completely different phenomena. Given that in principle the JRC-Eurofound tasks database includes good measures of forms of work organisation across occupations, we can only understand this strong inconsistency as an indication that online job ads are not a good source for inferring the forms of work organisation prevalent across different occupations.
- As for **tools**, the values are more consistent, though not as much as for intellectual tasks. Advanced ICT is the indicator most consistently measured in the two sources, followed by mechanical machinery and computing devices. Although use of basic ICT at work is reasonably consistent, there are several outliers that make the uncertainty area around the regression line quite big.

*Figure 7: Comparing indices of task content across occupations*

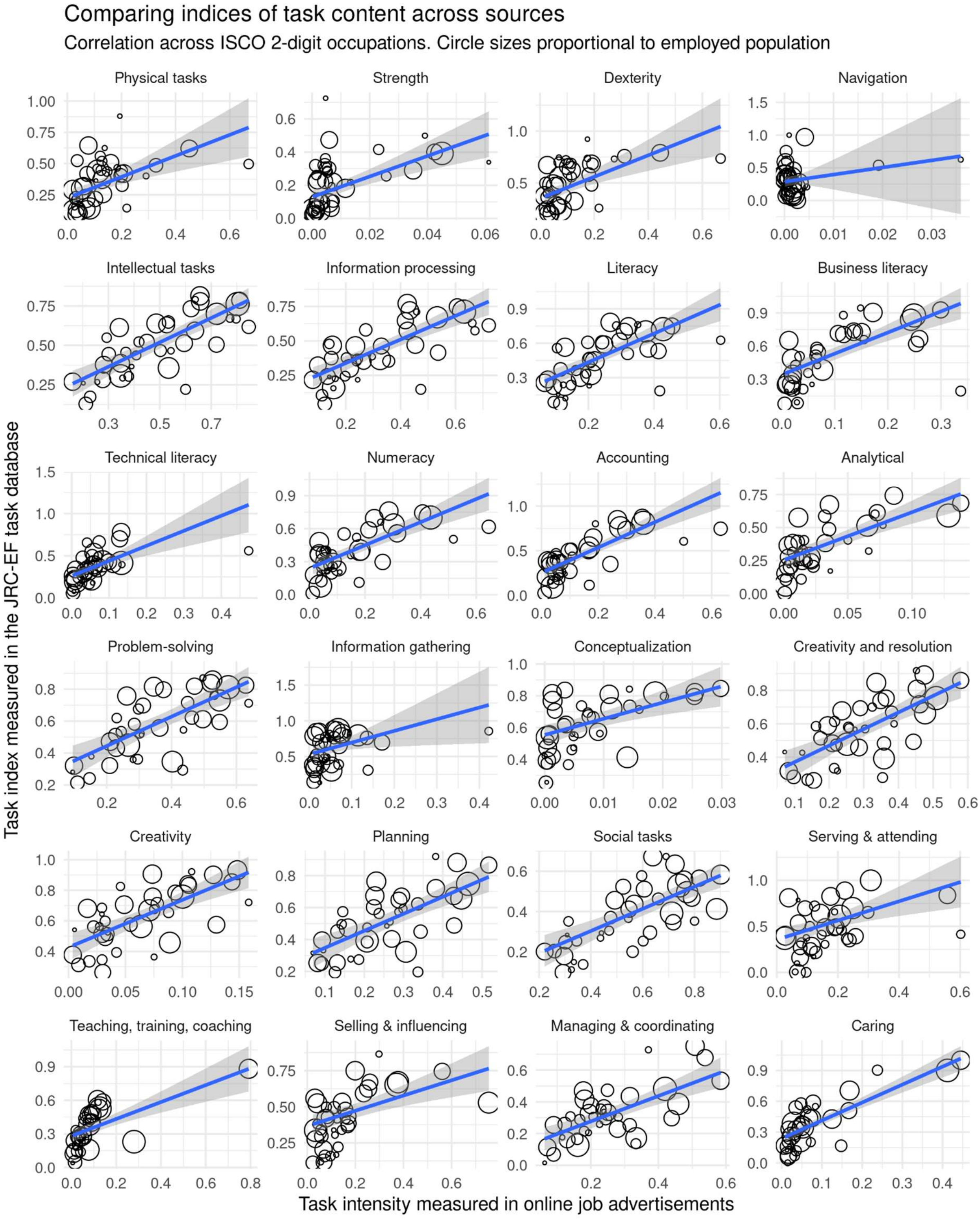
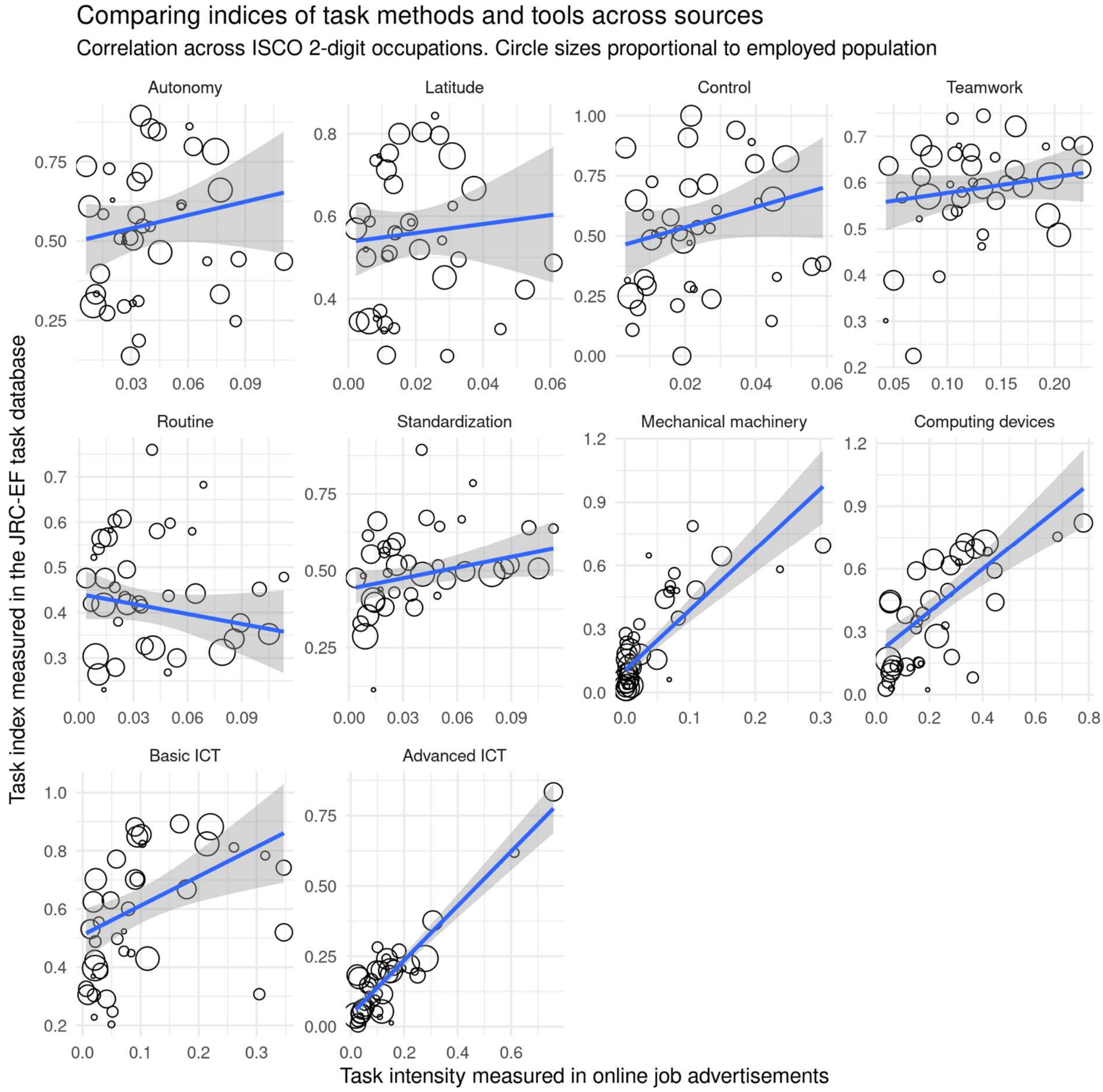
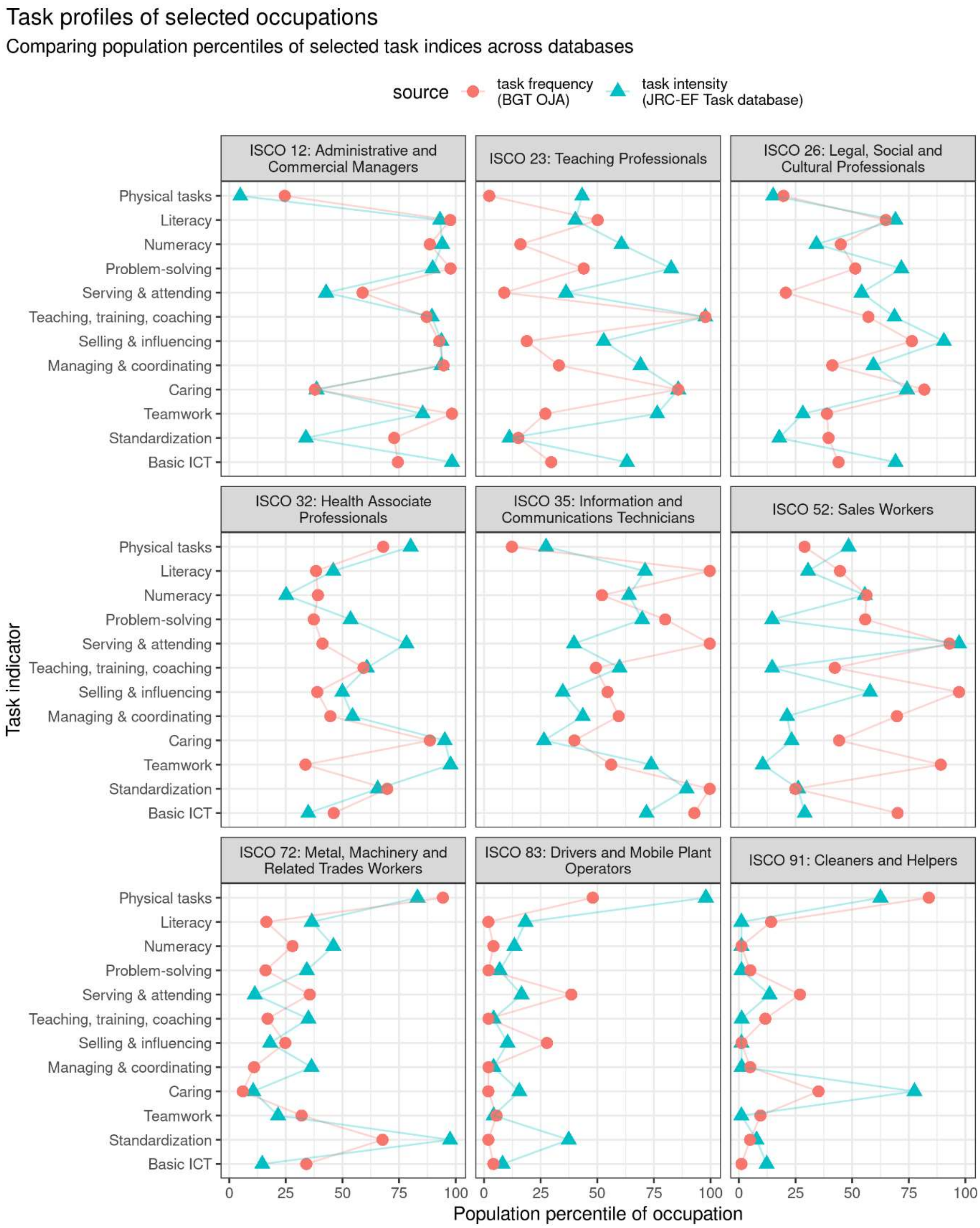


*Figure 8: Comparing indices of task methods and tools across occupations*




A different but complementary way to assess the consistency between the task measures inferred from online job ads and measured in surveys is to look at the task profiles of specific occupations (ISCO 2-digit codes). This is shown in *Figure 9* below. To facilitate the comparison at this level of detail and remove the distorting effect of the different value ranges we have repeatedly observed, we have transformed the scores for both sources into percentiles. In other words, the score for a given task indicator and occupation in this case represents the position that the occupation has in the percentile distribution of all occupations ranked by that task indicator. For instance, if an occupation gets a value of 97 in the indicator of "physical tasks'' in a given source, we know that (for that source) that occupation has one of the highest values of "physical tasks" of the entire economy (because it is located at the 97th percentile in terms of its physical task content). The percentile values in *Figure 9* are represented with red circles for BG data and with blue triangles for the JRC-Eurofound tasks database. For each of the sources, we have also added lines connecting the task indicator values. Since the task indicators are structured according to the JRC-Eurofound taxonomy, the zigzagging lines connecting the values for each source can be interpreted as a representation of the task profile of that occupation (for instance, the line of a more cognitively demanding occupation would go down for physical tasks and up for intellectual tasks with middling values for social tasks, etc.). For assessing the consistency in the task profiles according to the two sources, we can therefore not only look at the distance between the scores for specific task indicators, but also at the overall shape of the line representing the task profile. Two lines which have a similar profile and do not cross, even if the actual scores are different, would indicate consistent task profiles in the two sources.

*Figure 9: Comparing task profiles for selected occupations*




The comparison of task profiles according to online job ads and survey data suggests a high consistency in some occupations and a low consistency in others:

- For **administrative and commercial managers**, the task profiles are very consistent in the two sources, with the exception of standardisation (higher for BG data) and basic ICT (higher for JRC-Eurofound data). However, the overall pattern of the two profiles looks remarkably similar.

- For **teaching professionals**, there is some consistency in the overall profiles but not in terms of scores: it seems that BG data tends to assign lower values to teachers for most of the task categories (except for general literacy and caring). This may reflect the fact that, as we have repeatedly argued, the inferred values of task content from online job ads tend to be more concentrated in the few task categories which are more idiosyncratic of each occupation. By contrast, task categories that are generally higher in survey data tend to be less frequently mentioned in job ads.
- **Legal, social and cultural professionals** are quite consistent in the overall profiles although there are some exceptions (again, BG data suggests higher standardisation and lower ICT than JRC-Eurofound data).
- **Health associate professionals** (nurses) are broadly consistent in their profiles across the two sources, although there is more serving in the JRC-Eurofound database and less teamwork in BG data.
- **ICT technicians** are also broadly consistent in profiles and scores with some exceptions: surprisingly high values for *serving and attending* and general literacy in BG data compared to the JRC-Eurofound data.
- **Sales workers** seem less consistent, both in terms of scores and in terms of overall profiles, although there is some broad similarity in both patterns.
- But where the two sources give less consistent pictures of task content, methods and tools is for the three manual occupations included in *Figure 9*. The three occupations at the bottom of the figure (**metal, machinery and related trades worker**s; **drivers and mobile plant operators**; and **cleaners and helpers**) show profiles which often cross in inconsistent ways, with very low values overall. That said, the two sources show some consistency in the specific task categories that get higher values. Both sources assign very high scores for “physical tasks” for the three occupations. Additionally, both also coincide in giving high values of standardisation to metal, machinery and related trades workers and in giving high values of caring to cleaners and helpers (a category which includes many low-paid carers). It should be noted that the lowest levels of consistency mostly apply to task categories which receive very low values in both sources, and lower values imply less precision in the specific ordering of the scores that determines the shape of the profiles (and whether the lines cross or not).

Overall, we conclude that online job ads data can be used for inferring the task content of occupations, but with some cautions and limitations. The frequency of mentions of a given type of task content in online job ads is generally a good proxy for task intensity. However, it is important to be aware that most of the values tend to be lower than they would be if we would observe task content directly. More specifically, the mentions tend to cluster around those specific task categories which are most salient for a given occupation (according to the judgment of those producing the online job ads), while other types of task content, which may be also important in practice but are less specific of that occupation, tend to get fewer mentions. This is why the statistical distribution of task mentions in job ads is considerably more dispersed than the statistical distribution of task content observed in surveys. Another important caveat is that not all occupations and task categories are equally well covered. Our analysis suggested that manual occupations tend to be less well described in online job ads (or less well processed by the analysis, it is hard to say), and thus their task content is probably not well inferred from online job ads. This is consistent with (and compounds) the observation we made earlier that manual occupations are less well covered in online job ads. Moreover, our analysis showed that online job ads data is relatively good for inferring intellectual task content, and to a lesser extent social task content and use of technologies at work, but not very good for inferring physical task content and forms of work organisation.

# 7 Conclusions

Data from online job advertisements (OJA) present a number of advantages for the study of labour markets, in particular for the emerging area of "Skills intelligence". Compared to traditional surveys, these data sources are detailed and timely, and reported directly from the source – employers posting advertisements online. By signalling movements in the type and quantity of jobs created, these data collected "in the wild" may allow us to detect changing trends in demand for job profiles, education, skills, and the task content of occupations. This paper analyses the extent to which online job advertisement data accurately represent the composition of the labour market, and describe what people do at work and the skills they need.

We present a simple framework to deconstruct online job advertisements in the context of an organisation structure in the job-seeking process and relate it to an existing unified framework that articulates the concepts of task, skill, and competence. We argue that job advertisements can describe the different tasks bundled into a job, within the labour process of an organisation. However, this description may also be biased by some factors within the organisation – such as the different perspective that management, human resources, employees and unions have of the tasks bundled in a specific profile, which also includes a tacit dimension. Moreover, since the goal of job ads is to communicate with candidates from outside the organisation, they need to be formulated in a common language shared by job seekers, which is influenced by social norms and institutions, and by archetypal occupational models. All these elements contribute to define an idea of "competence" for well-known occupations, which are often described metonymically based on the job title: a cook should be competent in cooking, and a teacher in teaching. As a result, job advertisements may sometimes omit some skills, which – while ultimately important – can be taken for granted by employers.

The paper examines the Nova UK dataset from Burning Glass Technologies, which collects over 60 million online job advertisements for the United Kingdom from January 2012 to January 2020. This structured dataset represents posted positions through a number of fields but does not include the original text of the advertisement. We assess it in terms of how accurately it represents the structure of the labour market, and on how closely its description of different occupations match existing task data. To do so, we developed a Skill-Task Dictionary, which maps the skill keywords indexed in the Nova UK data in terms of a hierarchical Task Taxonomy used to consistently describe occupation task profiles in the JRC-Eurofound EU occupational Task Database. We find that the Nova UK data set provides reasonably good description of occupation task profiles, but that online job advertisements in general, and those processed by Burning Glass in particular, suffer from three types of potential biases.

The first is a structural representation bias in online job ads, because not all types of jobs are equally likely to be advertised online. Many jobs are filled via informal networks rather than through formalised communication channels such as the internet, or they may only be advertised offline, or filled via internal promotion channels. To the extent that these factors are more frequent for some types of jobs, this can make OJA data a biased source of information for the composition of the labour market. For instance, low-paid jobs in small firms may be more likely to be filled via informal networks and thus underrepresented in OJA data; or high-paid occupations in large firms filled via internal promotion channels may never be publicly advertised. Indeed, in the Nova UK dataset, we observe an over-representation of managerial, professional and clerical occupations, compared to the actual distribution of employment according to labour force survey data. Conversely, manual occupations requiring low or generic academic credentials tend to be under-represented in this data.

A second category includes informational biases in online job ads, because they are necessarily a partial and intentional description of their respective jobs, derived from the limitations of advertisements as a communication medium. This includes incompleteness, because advertisements tend to be concise in their description of the respective jobs: many job ads only include a handful of skill keywords, or even none at all. That may indeed be because some jobs do not need much description beyond the job title or even the name of the employer. It may also be because some jobs are difficult to describe, have a significant tacit dimension, or the person drafting the text of the online job ad may not have a clear understanding of them. Again, to the extent that this is systematically associated with some types of jobs and not others, this can make OJA a biased source of information on the attributes of the respective jobs. Indeed, we found significantly more descriptors (skill keywords) for highly qualified occupations, and a better representation of intellectual and social task content compared to physical and manual. Moreover, the Nova UK dataset was considerably better in its coverage of use of technology at work than in measuring forms of work organisation.

Another instance of informational bias is the one-sided perspective of advertisements. Their primary purpose is to find a suitable candidate for the job and to preserve the image of the employer, not to faithfully describe the position in every detail. Some aspects of the job may not be disclosed for reasons of privacy or economic interests (such as the salary or the name of the employer), or for being unpleasant or unattractive. Conversely, some aspects of the job may be over-emphasized or even spuriously inserted because they may make the job and the prospective employer more attractive. We did find some hints of social desirability bias in the task descriptions of occupations within Nova UK data. For instance, positive forms of work organisation such as teamwork were more frequently mentioned than negative ones such as routine (which is never mentioned, in fact). In general, the poor coverage of work organisation may hint at social desirability bias, because work organisation can describe aspects like control or routine, which are not particularly attractive. By contrast, positive "soft" aspects of work such as communication are more frequently mentioned.

A third category includes analytical biases resulting from the processing of the online job ads scraped from websites into structured data for analytical purposes. This processing favours those attributes of jobs that are easier to process, for instance towards those that are standardised by common usage or institutionalised certification. More importantly for our purposes, a lack of standardisation certainly applies to the skills indexed in the data. This is partly an ontological question, as there are different, sometimes conflicting definitions of "skills" in the academic literature, in industry, and in policy practice, which makes it difficult to identify and classify them consistently. In processing online job advertisements, sometimes "skills" are just salient keywords, partly determined by the frequency with which they are mentioned in advertisements, which have to be explicitly indexed and looked for in the advertisement text, and possibly classified in terms of a curated taxonomy. As a result, the Nova UK dataset favours some types of skills, such as those related to ICT, that are clearly enumerated in the advertisement text and have a standard nomenclature. Not all the skills included in the dataset correspond to our definition of "skill" – the ability to do a task – nor are they always descriptors of demonstrable requirements. Sometimes they can also be formulaic buzzwords used to fill advertisements or describe certain occupations. There can also be a bias towards those jobs and attributes that the analysts themselves understand better, in their own image: professionals, highly educated, technical, software. The processing of the data can also force the observed information into standardised categories, thus removing variability that may not fit into existing taxonomies. For instance, the Nova UK dataset attempts to fit all job titles into the UK Standard Occupation Classification, with an accuracy which is difficult to assess (indeed, we occasionally found it to be inconsistent or plainly inaccurate), and which may in practice conceal the emergence of new job titles and occupational categories.

Based on this conceptual and empirical analysis of the potential sources of bias in online job ads data, we find that they can be a valuable additional source of information for the composition of the labour market, for what kinds of skills are demanded over time, and even for what are the contents of the different types of jobs, though with some biases and caveats. This type of data has certainly some very important benefits, such as its large scale, timeliness and granularity for at least some aspects of work, such as ICT skills demand. Nevertheless, there are some important limitations that should be carefully considered when using this type of data, to avoid falling prey to hidden biases and generating misleading intelligence. Online job ads codify and reinforce a trend of looking at labour markets and skills through the lens of professional occupations, such as office-based, white-collar jobs. Indeed, the data over-represents managerial, professional and clerical occupations; tends to detect relatively few skill or task descriptors in each advertisement; measures poorly skills and tasks that are of a manual, low-qualified or unstandardized nature; is a particularly unreliable source of information for forms of work organisation. It probably suffers from some "social desirability" biases emphasizing the most attractive attributes of jobs, and minimizing or omitting the least appealing ones. As a result, it offers a somewhat distorted picture of the underlying demand for task, skills, and occupations. We thus see it as a valuable complement, but certainly not a substitute, of survey-based data such as official vacancy statistics, the European Skills and Jobs Survey, the European Working Condition Survey, or the Labour Force Survey. We could also consider conducting follow-up surveys to the employers concerned, to compare the job description and skill requirements as presented in the advertisements with the actual position that has been filled. This would establish the representativeness of this data by directly comparing the advertisement with the actual task profiles and skills needed to employers.

Doubtless, this data has a lot of potential beyond its ability to faithfully describe occupation task content and skill demand. We plan to use this source of data to map technological diffusion, in particular for digital skills, which are especially well indexed, and characterise the evolution of skills and technologies used in Artificial Intelligence. This data may also provide some insight into occupational change, in terms of the changing structure and content of occupations, though it spans less than a decade, which limits the ability to observe long-running trends. Furthermore, the structure of job titles and their correspondence to standard occupational classifications provides an interesting challenge in itself, which may point to new occupational categories.

# References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics* (Vol. 4, pp. 1043-1171). Elsevier.

Attewell, P. (1990). What is Skill? *Work and Occupations*, 17(4), 422–448. DOI: 10.1177/0730888490017004003h

Azar, José A., Ioana Marinescu, Marshall I. Steinbaum, and Bledi Taska (2018). *Concentration in US labor markets: Evidence from online vacancy data*. No. w24395. National Bureau of Economic Research, 2018.

Bisello, M., Fana, M., Fernández-Macías, E. and Torrejón Pérez, S., (2021). A compre-hensive European database of tasks indices for socio-economic research, *JRC-Eurofound*, JRC124124.

Carnevale, A., T.Jayasundera and D.Repnikov (2014). "Understanding online job ads data: A Technical Report", Washington, D.C.: Georgetown University Center on Education and the Workforce.

Cammeraat, E. and M. Squicciarini (2021). Burning Glass Technologies' data use in policy-relevant analysis: An occupation-level assessment. *OECD Science, Technology and Industry Working Papers*, No. 2021/05, OECD Publishing, Paris, https://doi.org/10.1787/cd75c3e7-en

Cedefop (2019a). *Online job vacancies and skills analysis: a Cedefop pan-European approach*. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/097022

Cedefop (2019b). *The online job vacancy market in the EU: driving forces and emerging trends.* Luxembourg: Publications Office. Cedefop research paper; No 72. http://data.europa.eu/doi/10.2801/16675

Cedefop (2019c). *The Skills Employers Want! Cedefop develops real-time skills intelligence by analysing online job vacancies*. Cedefop briefing note 9137EN; ISBN 978-92-896-2763-4, DOI: 10.2801/79736

Cedefop (2019d). Online job vacancies and skills analysis: a Cedefop pan-European approach. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/097022

Cinque, M., Carretero, S. and Napierala, J., "Non-cognitive skills and other related concepts: towards a better understanding of similarities and differences." *JRC Working Papers on Labour, Education and Technology* 2021/09, Seville: European Commission, 2021, JRC123827

Cohen, L. E. (2016). "Jobs as Gordian knots: A new perspective linking individuals, tasks, organizations, and institutions." In *The structuring of work in organizations*. Emerald Group Publishing Limited.

Cunha, Flavio, and James Heckman. (2007). "The Technology of Skill Formation." American Economic Review, 97 (2): 31-47.

Damelang, A., Stops, M., & Abraham, M. (2019). "Occupations as labour market institutions. Occupational regulation and its effects on job matching and occupational closure." SozW Soziale Welt, 69(4), 406-426.

Deming, David, and Lisa B. Kahn (2018). "Skill requirements across firms and labor markets: Evidence from job postings for professionals." *Journal of Labor Economics* 36, no. S1 (2018): S337-S369.

De Pedraza P, Visintin S, Tijdens KG, Kismihók G (2019) "Survey vs scraped data: comparing time series properties of web and survey vacancy data", *IZA Journal of Labor Economics*, DOI: 10.2478/izajole-2019-0004, 8 (1)

European Commission (2020) *European Skills Agenda for sustainable competitiveness, social fairness and resilience*, COM(2020) 274

Felten, Edward W. and Raj, Manav and Seamans, Robert (2019), "The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization." *SSRN*. DOI: 10.2139/ssrn.3368605

Fernández-Macías, E. and Bisello, M., A Taxonomy of Tasks for Assessing the Impact of New Technologies on Work, Seville, European Commission, 2020, JRC120618

Grinis, Inna. (2017). The STEM Requirements of 'Non-STEM' Jobs: Evidence from UK Online Vacancy Postings and Implications for Skills & Knowledge Shortages."

Hershbein, Brad, and Lisa B. Kahn. (2018). "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings." *American Economic Review* 108.7: 1737-72.

International Labour Office & International Labour Organization. (2012). International standard classification of occupations: Isco-08. International Labour Office.

McGuinness, S., Pouliakas, K., & Redmond, P. (2018). Skills mismatch: Concepts, measurement and policy approaches. Journal of Economic Surveys, 32(4), 985-1015.

Nelson, R. and S. Winter (1982). *An Evolutionary Theory of Economic Change*. Belknap Press of Harvard University Press: Cambridge MA.

Rodrigues, M., Fernández-Macías, E., Sostero, M., *A unified conceptual framework of tasks, skills and competences*, Seville: European Commission, 2021, JRC121897.

Tijdens KG, Beblavý M, Thum-Thysen A (2018) Skill mismatch comparing educational requirements vs attainments by occupation, International Journal of Manpower, 39(8), 997-1009

Tijdens, K.G., Kaandorp, C.S. (2019) Classifying job titles from job vacancies into ISCO-08 and related job features – the Netherlands., Deliverable 8.05 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221.
https://www.researchgate.net/publication/335703270_Classifying_job_titles_from_job_vacancies_into_ISCO-08_and_related_job_features_-_the_Netherlands

Witz, A. (1990). Patriarchy and professions: The gendered politics of occupational closure. Sociology, 24(4), 675-690.

# Annexes

## Annex 1. The Skill-Task Dictionary

Table 2 below shows an overview of the number of skills identified in the NOVA UK database, classified in terms of the JRC-Eurofound Task Taxonomy. For each level of the taxonomy, it reports the number of distinct skill keywords, and their frequency across all job advertisements. The full dictionary is available at https://doi.org/10.5281/zenodo.6488230

*Table 2: Size and coverage of the skill-task dictionary*

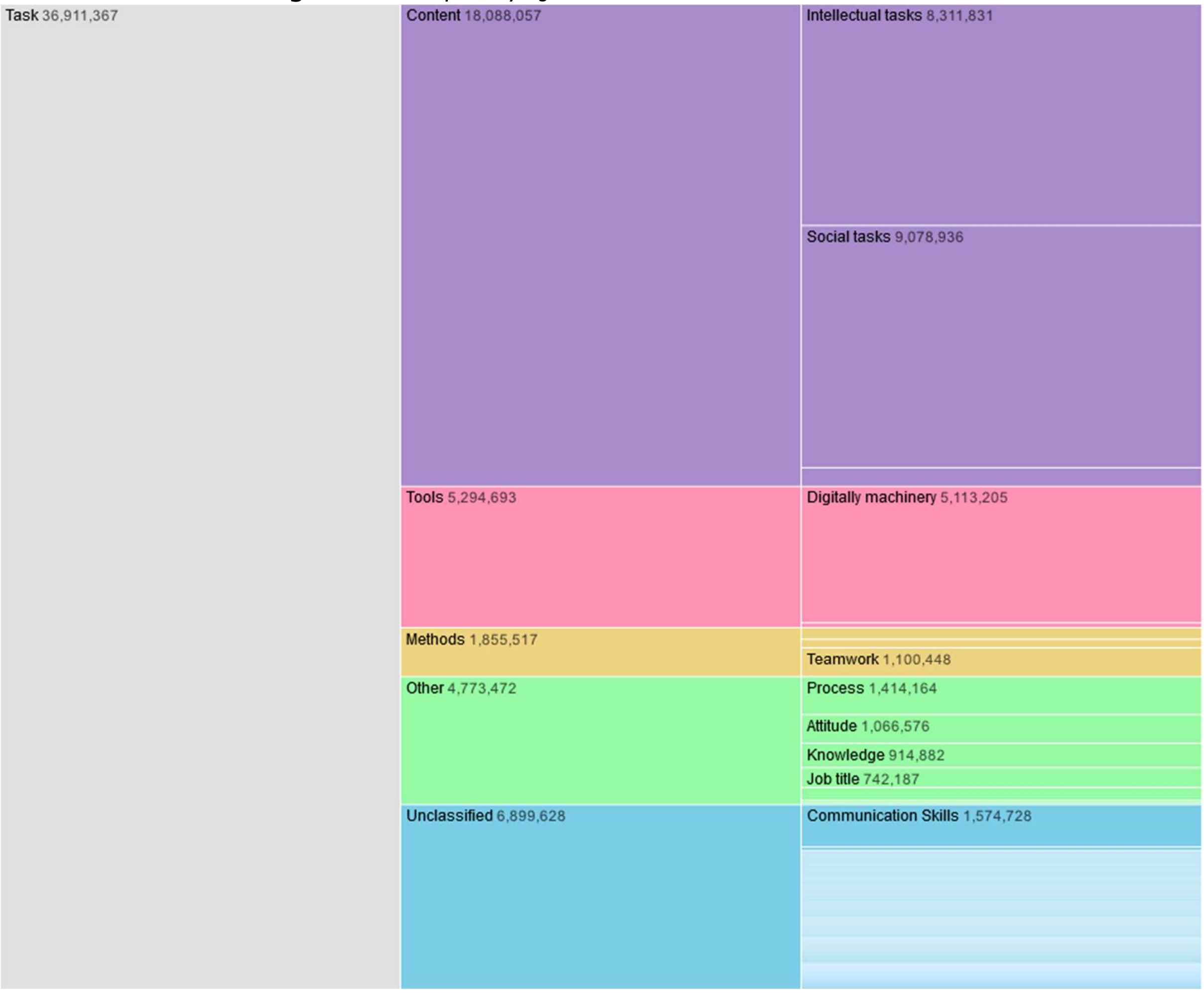
| | *Task category* | *Number of skills* | *Frequency in ads* |
|---|---|---|---|
| *Content of work* | ● Physical tasks | 6 | 215,406 |
| | └ strength | 3 | 126,359 |
| | └ dexterity | 92 | 4,528,548 |
| | └ navigation | 2 | 73,898 |
| | ● Intellectual tasks | | |
| | └○ uncodified information | 1 | 4,298 |
| | └○ information processing | 1 | 72,237 |
| | —○ literacy | 16 | 8,258,108 |
| | └ business | 84 | 8,219,623 |
| | └ technical | 10 | 3,084,263 |
| | └ humanities | 0 | - |
| | —○ numeracy | 1 | 984,394 |
| | └ calculation | 41 | 13,156,140 |
| | └ analytical | 127 | 2,589,023 |
| | └○ problem-solving | 11 | 5,530,790 |
| | └○ information gathering | 1 | 125,813 |
| | └ search | 23 | 2,649,583 |
| | └ conceptualisation | 9 | 559,652 |
| | └○ creativity and resolution | 6 | 363,781 |
| | └ creativity | 22 | 4,415,356 |
| | └ planning | 37 | 22,709,549 |
| | ● Social Tasks | 4 | 14,029,964 |
| | └ serving and attending | 20 | 11,998,406 |
| | └ teaching, training and coaching | 14 | 7,173,664 |
| | └ selling and influencing | 72 | 22,113,344 |
| | └ managing and coordinating | 37 | 15,793,202 |
| | └ caring | 437 | 5,925,763 |
| *Methods of work* | ● Autonomy | | |
| | └ latitude | 2 | 1,041,348 |
| | └ control | 1 | 1,369,584 |
| | ● Teamwork | 3 | 5,467,780 |
| | ● Routine | 0 | - |
| | └ repetitiveness | 0 | - |
| | └ standardisation | 12 | 1,420,841 |
| *Tools of work* | ● Machines | 17 | 1,296,364 |
| | ● ICT | | |
| | └○ autonomous | 1 | 16,543 |
| | └○ non-autonomous | 24 | 1,473,395 |
| | └ Computing devices | 3 | 1,894,022 |
| | └ Basic ICT | 17 | 10,160,726 |
| | └ Specialised ICT | 159 | 10,652,462 |
| | └ Advanced ICT | 136 | 29,093,896 |
| | └ Other ICT | 13 | 1,325,632 |
| *Other* | ● Job title | 12 | 6,150,250 |
| | ● Attitude | 10 | 7,747,994 |
| | ● Knowledge | 78 | 7,620,748 |
| | ● Process | 58 | 12,055,902 |
| | ● Sector | 17 | 4,079,901 |
| | ● Language | 4 | 450,951 |
| | ● Experience | 5 | 866,545 |

## Annex 2. Occupation task profiles

The figure below represents the relative frequency of skill keywords, in terms of the JRC-Eurofound Task Taxonomy, for all job advertisements in the NOVA UK dataset. The different columns represent the hierarchical levels of the taxonomy, and the height of each divisions is proportional to the frequency of the task across all advertisements.

The full interactive version, which includes all hierarchical levels of the Task Taxonomy, and the corresponding skill keywords observed in the NOVA UK dataset is available at https://observablehq.com/@m-sostero/skill-task-mentions

*Figure 10: Frequency of tasks across all advertisements*



## Annex 3. Comparing task indices across databases

The table compares the values of the task indicators for ISCO 2-digit occupations measured in the NOVA UK dataset in 2019 to those measured in the JRC-Eurofound Task Database for the United Kingdom. The p-values test for the null hypothesis that there is no correlation between the indicators measured from different databases. Although the magnitude of the different indicators is not strictly comparable across database, we nevertheless report the Pearson correlation coefficient between the two sources. We also report the rank-based Spearman correlation coefficient, expressed both as a simple of rank across occupations, and weighted for the population size of each occupation, ad measured in the UK labour force survey.

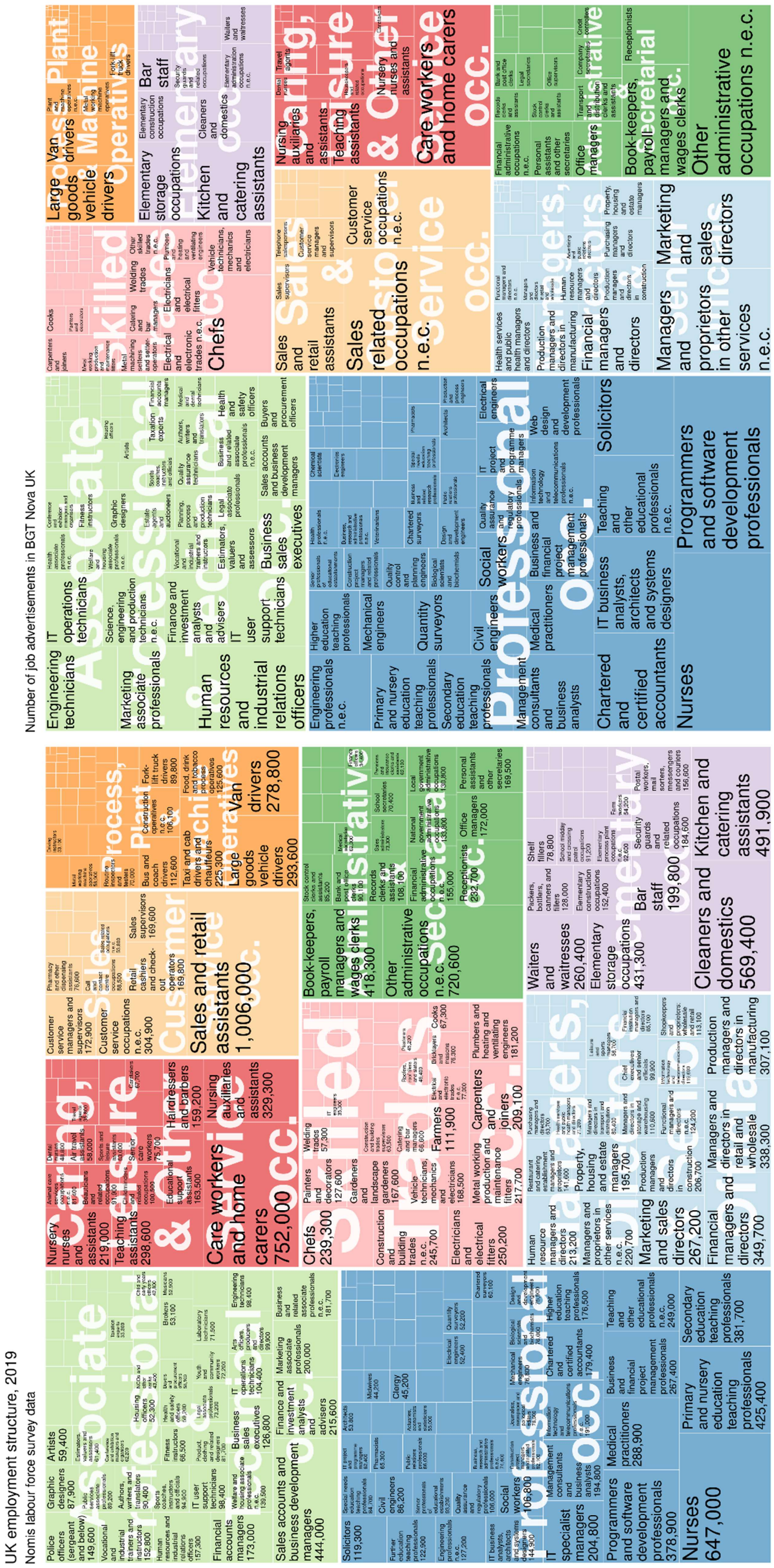
*Table 3: Correlations between JRC-Eurofound task database and task intensity measured computed from NOVA UK*

| | Pearson correlation | | Spearman correlation | | | |
|---|---|---|---|---|---|---|
| | Magnitude, unwgt | | Rank, unweighted | | Rank, weighted | |
| Task indicator | estimate | p-value | estimate | p-value | estimate | p-value |
| ● Physical tasks | 0.440 | 0.005 | 0.559 | 0.000 | 0.615 | 0.000 |
| └ strength | 0.344 | 0.032 | 0.640 | 0.000 | 0.670 | 0.000 |
| └ dexterity | 0.490 | 0.002 | 0.600 | 0.000 | 0.582 | 0.000 |
| └ navigation | 0.236 | 0.149 | -0.178 | 0.280 | -0.138 | 0.404 |
| ● Intellectual tasks | 0.763 | 0.000 | 0.767 | 0.000 | 0.749 | 0.000 |
| └○ information processing | 0.720 | 0.000 | 0.697 | 0.000 | 0.689 | 0.000 |
| —○ literacy | 0.642 | 0.000 | 0.682 | 0.000 | 0.679 | 0.000 |
| └ business | 0.607 | 0.000 | 0.688 | 0.000 | 0.641 | 0.000 |
| └ technical | 0.490 | 0.002 | 0.729 | 0.000 | 0.713 | 0.000 |
| —○ numeracy | 0.687 | 0.000 | 0.662 | 0.000 | 0.698 | 0.000 |
| └ calculation | 0.718 | 0.000 | 0.700 | 0.000 | 0.722 | 0.000 |
| └ analytical | 0.721 | 0.000 | 0.719 | 0.000 | 0.724 | 0.000 |
| └○ problem-solving | 0.716 | 0.000 | 0.738 | 0.000 | 0.699 | 0.000 |
| └○ information gathering | 0.345 | 0.032 | 0.447 | 0.004 | 0.438 | 0.005 |
| └ search | 0.367 | 0.022 | 0.474 | 0.002 | 0.477 | 0.002 |
| └ conceptualisation | 0.615 | 0.000 | 0.665 | 0.000 | 0.677 | 0.000 |
| └○ creativity and resolution | 0.696 | 0.000 | 0.683 | 0.000 | 0.693 | 0.000 |
| └ creativity | 0.654 | 0.000 | 0.655 | 0.000 | 0.674 | 0.000 |
| └ planning | 0.686 | 0.000 | 0.671 | 0.000 | 0.679 | 0.000 |
| ● Social Tasks | 0.729 | 0.000 | 0.744 | 0.000 | 0.647 | 0.000 |
| └ serving | 0.369 | 0.021 | 0.417 | 0.008 | 0.410 | 0.009 |
| └ teaching | 0.668 | 0.000 | 0.808 | 0.000 | 0.805 | 0.000 |
| └ selling & influencing | 0.583 | 0.000 | 0.651 | 0.000 | 0.686 | 0.000 |
| └ managing | 0.692 | 0.000 | 0.605 | 0.000 | 0.670 | 0.000 |
| └ caring | 0.798 | 0.000 | 0.739 | 0.000 | 0.771 | 0.000 |
| ● Autonomy | 0.086 | 0.602 | 0.173 | 0.292 | 0.162 | 0.325 |
| └ latitude | 0.008 | 0.961 | 0.071 | 0.669 | 0.071 | 0.667 |
| └ control | 0.164 | 0.319 | 0.225 | 0.168 | 0.221 | 0.177 |
| ● Teamwork | 0.373 | 0.019 | 0.315 | 0.051 | 0.306 | 0.058 |
| ● Routine | -0.061 | 0.713 | -0.037 | 0.824 | -0.025 | 0.879 |
| └ standardisation | 0.366 | 0.022 | 0.435 | 0.006 | 0.445 | 0.005 |
| ● Machines | 0.750 | 0.000 | 0.788 | 0.000 | 0.741 | 0.000 |
| ● ICT | 0.286 | 0.078 | 0.328 | 0.042 | 0.321 | 0.047 |
| └ basic ICT | 0.427 | 0.007 | 0.578 | 0.000 | 0.568 | 0.000 |
| └ programming | 0.905 | 0.000 | 0.682 | 0.000 | 0.690 | 0.000 |

## Annex 4. Comparing volumes of job advertisements with employment surveys

Comparing occupational composition of job advertisements with employment surveys

*Figure 11: UK occupational structure, 2019*



## Annex 5. The Burning Glass Skill Taxonomy

*Table 4* below reports tallies of skills in the data by Burning Glass' 29 *skill cluster families* (the broadest taxonomic level). These are further divided in a variable number of *skill clusters*, which in turn contain a number of individual skill items. The final column reports the number of ads that mention skills in each of the *skill cluster families.* A plurality of ads (over 81 million of them) mention skills not covered by the classification (over 7,000 or half the total).

*Table 4: Burning Glass Skill classification, coverage in UK data*

| BGT Skill cluster family | N. clusters | N. Skills | N. mentions |
|---|---|---|---|
| *(Unclassified)* | (Unclassified) | 7,335 | 81,516,511 |
| *Information Technology* | 81 | 1,106 | 54,953,830 |
| *Business* | 28 | 232 | 23,486,290 |
| *Sales* | 24 | 183 | 19,227,398 |
| *Finance* | 30 | 365 | 15,125,529 |
| *Health Care* | 65 | 941 | 13,638,507 |
| *Marketing and Public Relations* | 23 | 295 | 11,464,809 |
| *Customer and Client Support* | 5 | 67 | 8,663,947 |
| *Administration* | 8 | 66 | 6,819,810 |
| *Supply Chain and Logistics* | 18 | 241 | 6,229,674 |
| *Engineering* | 27 | 271 | 5,554,606 |
| *Education and Training* | 17 | 128 | 5,401,181 |
| *Industry Knowledge* | 94 | 288 | 4,876,317 |
| *Manufacturing and Production* | 15 | 184 | 4,089,015 |
| *Maintenance, Repair, and Installation* | 16 | 214 | 3,272,191 |
| *Personal Care and Services* | 5 | 39 | 3,233,948 |
| *Human Resources* | 10 | 159 | 3,231,580 |
| *Analysis* | 17 | 202 | 3,203,374 |
| *Design* | 10 | 123 | 3,035,018 |
| *Media and Writing* | 9 | 95 | 1,993,284 |
| *Science and Research* | 14 | 196 | 1,852,677 |
| *Architecture and Construction* | 17 | 150 | 1,805,742 |
| *Legal* | 8 | 95 | 1,445,153 |
| *Energy and Utilities* | 21 | 136 | 732,756 |
| *Environment* | 13 | 127 | 716,383 |
| *Economics, Policy, and Social Studies* | 5 | 28 | 363,735 |
| *Public Safety and National Security* | 8 | 52 | 351,670 |
| *Agriculture, Horticulture, and the Outdoors* | 3 | 36 | 66,937 |
| *Religion* | 1 | 5 | 9,414 |

## Annex 6. Job titles classified in different SOC codes

The table lists the most common instances of job titles that have different SOC occupational classification codes across the 60 million advertisements in the NOVA UK dataset, 2012-2020. The column *splits* reports how many different SOC codes are observed, and the next column shows the five most common codes, including "none" if the job title is sometimes not coded.

*Table 5: Most common examples of job titles classified in different SOC codes*

| Job title | N. of ads | splits | Most common SOC codes |
|---|---|---|---|
| Administrator | 212,599 | 2 | 4159, 9219 |
| Project Manager | 176,387 | 4 | 2424, 7130, 2436, none |
| Sales Executive | 153,520 | 4 | 7129, 3534, none, 3542 |
| Business Development Manager | 135,789 | 3 | 1132, none, 3545 |
| Care Assistant | 133,437 | 2 | 6145, 9120 |
| Accounts Assistant | 127,582 | 2 | 4122, 5330 |
| Account Manager | 106,847 | 6 | 7129, 3534, 7130, 5436, 3538, ... |
| Assistant Manager | 100,143 | 41 | 4159, 7130, 1223, 5436, 2431, ... |
| Management Accountant | 97,947 | 2 | 2421, 9120 |
| Store Manager | 90,621 | 5 | 1190, 7130, 5436, 1223, none |
| Registered Nurse | 79,553 | 2 | 2231, 9120 |
| Finance Manager | 71,668 | 3 | 1131, 7130, 9120 |
| Maintenance Engineer | 69,931 | 2 | 3113, 5314 |
| Site Manager | 67,387 | 10 | 1122, 1133, 7130, none, 1123, ... |
| Healthcare Assistant | 67,289 | 2 | 6141, 9120 |
| Sales Assistant | 65,198 | 3 | 7129, 3534, none |
| Sales Administrator | 64,262 | 4 | 4151, 3534, none, 9120 |
| Sales Manager | 60,859 | 5 | 1132, 7130, 3534, none, 3545 |
| Sales Advisor | 57,233 | 3 | 7129, 3534, 7111 |
| Teaching Assistant | 55,547 | 2 | 6125, none |
| Marketing Manager | 54,198 | 2 | 1132, 7130 |
| Operations Manager | 52,994 | 2 | 1259, 7130 |
| Electrician | 52,349 | 2 | 5241, 9272 |
| Marketing Executive | 51,939 | 2 | 3543, none |
| Team Leader | 51,615 | 13 | 7130, 6146, 1161, 3531, 6212, ... |
| Deputy Manager | 50,782 | 5 | 1259, 7130, 1223, 5436, 1133 |
| General Manager | 50,773 | 8 | 1259, 7130, 1223, 2436, 1221, ... |
| Assistant Accountant | 49,479 | 2 | 2421, 9120 |
| Field Service Engineer | 46,478 | 2 | 5249, 3131 |
| Customer Service Assistant | 45,823 | 2 | 7219, 9272 |
| Supervisor | 42,618 | 5 | 1259, 7130, 1181, 5436, 9120 |
| Financial Accountant | 41,659 | 2 | 2421, 9120 |
| Nursery Nurse | 41,618 | 2 | 6121, none |
| Project Engineer | 41,468 | 5 | 2121, 2122, 2111, 2136, 2129 |
| Sales Consultant | 39,259 | 4 | 3542, 3534, none, 7111 |
| Design Engineer | 38,011 | 7 | 2126, 2122, 2121, 2123, 3113, ... |
| Accountant | 37,503 | 2 | 2421, 9120 |
| Area Sales Manager | 37,392 | 2 | 3545, none |
| Field Sales Executive | 34,642 | 3 | 7129, 3534, 3542 |
| Contracts Manager | 34,124 | 3 | 3545, 7130, none |
| Sales Representative | 32,256 | 4 | 7129, 3534, none, 3542 |
| Driver | 29,449 | 2 | 8239, 7123 |
| Production Manager | 28,690 | 2 | 1121, 7130 |
| Account Executive | 27,273 | 2 | 7129, 3542 |
| Product Manager | 27,255 | 3 | 3545, 7130, 9120 |
| Office Manager | 27,136 | 3 | 4161, 7130, 5436 |
| Quality Manager | 26,998 | 3 | 2462, 9120, 7130 |
| Commercial Manager | 26,969 | 5 | 1259, 1133, 7130, 5436, 2317 |

**GETTING IN TOUCH WITH THE EU**

**In person**

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

**On the phone or by email**

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

**FINDING INFORMATION ABOUT THE EU**

**Online**

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

**EU publications**

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

**The European Commission's science and knowledge service**

Joint Research Centre

## JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.