ie UNIVERSITY
CENTER FOR THE GOVERNANCE OF CHANGE

CLOSING THE DIGITAL SKILL GAP: THE POTENTIAL OF ONLINE PLATFORM DATA FOR ACTIVE LABOUR MARKETS POLICIES

FABIAN STEPHANY
OXFORD INTERNET INSTITUTE, UNIVERSITY OF OXFORD &
HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY

JUNE 2022

CLOSING THE DIGITAL SKILL GAP: THE POTENTIAL OF ONLINE PLATFORM DATA FOR ACTIVE LABOUR MARKET POLICIES

Fabian Stephany

Oxford Internet Institute, University of Oxford &

Humboldt Institute for Internet and Society

Reference to this paper should be made as follows:
Stephany, F. (2022) "Closing the Digital Skill Gap: The Potential of Online Platform Data for Active Labour Market Policies", The Digital Revolution and the New Social Contract series, Center for the Governance of Change, IE University, June.

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License. To view a copy of the license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/"

ABSTRACT

The global challenge of rapidly changing skill requirements due to task automation is currently overwhelming workers, firms, and governments. Indeed, the digital skill gap continues to widen as technological and social transformation outpaces national education systems, and the precise skill requirements for mastering emerging technologies, such as Artificial Intelligence (AI), remain opaque. For many newly emerging jobs, labour market mismatches occur as training lags behind workforce and industry needs. In this article, we report on how online user-generated data can provide useful foresight about skills requirements and training implications, showcasing how data from online labour platforms could help us to monitor and understand the complex system of skill formation. This data could allow us to establish a taxonomy of skills, understand their application and individual complementarity, and enable automated, individual, and far-sighted suggestions on the value of learning a new skill in a future of technological disruption. Policy recommendations are manifold. First, reskilling institutions, like the European Centre for the Development of Vocational Training, are a beneficiary of this highly individualised data. Workers with the need to reskill could be located in the data-based landscape of skills and would receive a targeted reskilling advice that allows them to switch to more sustainable occupations that are closely related to their existing skill set. Furthermore, official occupational and skill taxonomies could be improved with near real-time data, as conventional taxonomies currently struggle with the EC's ambitious effort to define "AI jobs" and "green skills". The European Commission's 2022 Data Act[1] acknowledges this versatile potential of online generated data. However, opening the Data Act towards data access practices via web-scraping and improving the legal security of data recipients would further facilitate the usage of data in the public interest.

INTRODUCTION

Automation, digital platforms, and other innovations are changing the fundamental nature of work. On the labour market, task automation and rapidly changing occupations (Acemoglu & Autor, 2011) lead to the paradoxical situation of unemployment during a time of labour shortage (Autor, 2015). Professional service or admin, white collar jobs are particularly exposed to this trend that is "hollowing-out" the middle employment spectrum (Baldwin & Forslid, 2020). A conventional policy response has been to align national education systems with changing labour market demand. However, this solution

[1] https://digital-strategy.ec.europa.eu/en/policies/data-act

is becoming increasingly ineffectual as technological and social transformation outpaces national education systems, which usually show slow reaction times to systemic change (Collins & Halverson, 2018). Workers have to some extent begun to assume greater personal responsibility for reskilling, via skill-based online training (Allen & Seaman, 2015; Lehdonvirta, Margaryan, & Davies, 2019). However, the economic benefits and costs of reskilling strategies are often unclear, as they are highly individual, and the precise skill requirements for mastering emerging technologies, such as "AI" or "Big Data" analysis, remain opaque (De Mauro et al., 2018).

Here, we show that online generated data can provide important information for active labour market policies in the domain of reskilling, including data generated by online job vacancy sites and social networking pages. More specifically, in this paper we focus on data from online labour platforms (OLP), global marketplaces that match millions of buyers and sellers of digitally delivered work. While these only cover a small segment of the labour market, i.e., digitalised tasks from jobs in the professional service sector, their data contains useful information on both the demand and supply side of skills. In addition, it is possible to observe the job matching process and price (e.g., hourly rate) for each job with a certain skill bundle attached to it. OLPs therefore present an ecosystem that covers the *entire* economic complexity of the price building mechanism. These properties make OLPs an interesting data source for studying skill formation, skill matching, and the evaluation of individual skills or skill bundles (Stephany, 2021).

The skill and task taxonomies emerging from the analysis of large-scale OLP data allow researchers to investigate the complementarities of particular skills. A specific skill, for example data mining, machine learning, or 3d design, can incur different costs and leverage different benefits depending on how it sits within the learner's existing skill set. These skill trajectories, as learners develop their skill bundles, illustrate what online labour market data allow us to say about the complementarities of learning a new skill. The data allow us to evaluate the economic benefit of individual skills based on the *existing* skill bundle of a person, to ultimately sketch optimal individual re-skilling pathways. Early investigations on the complex ecosystem of skill formation show that online data can indeed be a valuable tool for designing sustainable reskilling policies. In its Data Act (EC, 2022), the European Union has identified that private sector data could be demanded to administer societal change for the public good. However, to leverage the full potential of online generated data, e.g., from OLPs, amendments in data access, legal security, and strategic partnerships need to be implemented. If platform providers could be integrated into this renewal of the social contract, the true value of online generated data could finally be released for the benefit of society.

THE EU SKILL GAP AND STRATEGIES TO CLOSE IT

Technological Change Drives Skill Mismatches

Technology is changing the way we work, in a fundamental way. Technological and social transformation change the necessary skill composition of work (Acemoglu & Autor, 2011), leading to the paradox of simultaneous unemployment and labour shortage (Autor, 2015). Often, the precise skill requirements for mastering emerging technologies remain opaque (De Mauro et al., 2018), and – despite the growing demand for them – new occupations, in fields like data management, digital design, and autonomous systems, are not yet acknowledged by official employment taxonomies. This is bad news for both firms and workers, as professional training providers find it hard to "speak" with the same language as market demand.

History suggests that this skills gap, even more than the elimination of jobs per se, increases economic inequality (Card & DiNardo, 2002) and causes a lag in firm growth (Krueger & Kumar, 2004) during times of technological and social transformation. In a situation of rapidly changing skill requirements and nameless new occupations, systematic oversight is essential. However, individual workers often lack foresight regarding the occupations that are about to emerge, or which skills are rising or falling in demand. They might get locked into path dependencies that may result in dead ends, preventing them from reskilling into new areas (Escobari et al., 2019).

Skill mismatch can also negatively impact earnings, as individuals might accept a less desirable job as a result of higher competition, as is shown by the example of high-skilled knowledge workers, who run multiple, short-term and often precarious online jobs. This trend might also lead to permanent effects in the form of human capital depreciation. Skill mismatches and skill shortages distort the optimal allocation of resources, and thereby reduce average productivity. In terms of GDP loss, Mavromaras et al. (2007) proxy the individual productivity loss with the estimated wage penalty associated with over-skilling, and multiply this penalty by the number of overskilled workers by educational attainment level, concluding that the costs of over-skilling amount to about 2.6% of GDP in the EU.

The Beveridge curve, which describes the relation between unemployment and job vacancy rates, allows us to study the impact of business cycles on skill mismatches and their development over time. In times of economic contraction, job vacancy rates drop and unemployment rises. In times of skill mismatches or increased search intensity, shifts in the curve can occur such that unemployment rises given a specific level of vacancies. Figure 1 shows that in the EU, the Beveridge curve has shifted outwards during the 2008-

2021 period. After 2013, job vacancy rates increased and unemployment rates declined. Currently, vacancy rates are significantly higher than 10 years ago, with a lower level of unemployment, indicating that despite economic recovery many firms are not able to fill their job offerings.

Figure 1 *In the EU, the Beveridge curve has shifted upwards during the 2008-2021 period. Job vacancy rates have increased and unemployment rates declined. Source: Eurostat and author calculations.*

In times of economic downturn, matching efficiency, defined as the time taken to match unemployed workers with unfilled vacancies, declines. Active labour market policies can help to improve the matching efficiency by re- and up-skilling measures, which improve the matching prospects for the long term unemployed. Re- and up-skilling policies will become increasingly relevant, as job polarisation increases as a result of automation and digitisation, and as re-employment prospects worsen as a result of longer unemployment and rapidly changing skill requirements in response to technological change.

For the EU, recent skills and job forecasts by CEDEFOP (Pouliakas, 2021) indicate that job market polarisation continues to rise, with rising employment shares for professionals, managers and technicians on the one hand and a decline in labour market demand for clerks, craft workers and plant and machine operators. Typically, jobs at low risk of automation require professional training or tertiary education. Unfortunately, workers, and similarly employers, in domains that are *most* exposed to high automation risk are also least likely to invest in training (see Nedelkoska and Quintini, 2018) and often have limited access to it. A central question for active labour market policies targeting a

reduction of the skill mismatch is therefore **what type of training can effectively allow people to upgrade skills and move to jobs that are less automatable** (Tamm 2018, Schmidpeter and Winter-Ebmer 2018)?

With increasing international labour market competition and rapid technological change, skill requirements will continue to shift. Research shows that skills that are complementary to new technology requirements facilitate adaptation to changing job requirements, and carry additional value for employees. Respectively, for firms that want to adopt new technologies and the practices that come with them, employees' skills will become increasingly important (EIB 2018).

Existing Policy Solutions to Try to Close the Skill Gap

Most policies targeting the skills mismatch try to enhance the responsiveness of the education and training sectors to newly emerging demand from labour markets. This includes for example an enhancement of youth employability via reforming vocational training or better forecasts of future skills needed to meet labour market demands. Likewise, policy can target skill mismatches by reducing the information asymmetry between jobseekers, workers, and firms, for example if they are using different taxonomies to describe skill requirements and existing capabilities (Colahan et al, 2017).

Unfortunately, the conventional policy response – aligning training programmes with changing labour market demand – is becoming increasingly ineffectual as technological and social transformation outpaces national training systems (Collins & Halverson, 2018). Likewise, large employers are struggling to keep their workforces' skills up to date (Illanes et al., 2018). At the same time, the COVID-19 pandemic has tightened company budgets, forced employees to work remotely, and further driven the global need for reskilling (Stephany et al., 2022). Workers have begun to assume greater personal responsibility for their reskilling, via online courses, distance education tools, and entrepreneurial approaches to work (Allen & Seaman, 2015).

The responsibility for developing "far-sighted" skills by schools and employers is unclear. For schools it is often difficult to foresee the labour market demands of the next few years (Cappelli, 2014), and existing constraints on what should be taught can lead to inflexibility. In the absence of skill demand forecasts, pupils might invest in more academic skills as they want to cushion the potential costs of skill obsolescence in the medium to long run (Brunello and Rocco, 2017). Skill mismatches that are not solved by market mechanisms can thus be targeted by policies adjusting the under-provision of education or training. In Europe, levy-grant schemes, tax deductions, and co-financing

programs targeted at individuals are examples of public policies encouraging adult training that include co-financing programs targeted at firms (see Brunello & Wruuck, 2019).

Most recently, just-in-time skills development, motivated by perceived market shifts, has emerged, particularly where formal training courses remain unaffordable for many workers (Kester et al., 2006). Another factor driving this trend is the fact that female participation is still hindered by cultural aspects in traditional STEM education (Kahn & Ginther, 2017). Indeed, research shows that independent professionals, including women, prefer informal, digital, social learning resources like Stack Overflow and tutorial videos to develop new skills (Yin et al., 2018).

Lastly, on a European level, skill policies can and must proactively respond to skill mismatches caused by structural trends such as technological change (e.g., digitalisation and automation), which lead to labour market polarisation or inequality. It will be increasingly important to ensure a smooth transition between existing jobs that are exposed to structural change, such as automation pressure, and the development of new and sustainable qualifications. This is a challenge for both firms requiring workers skilled with new capabilities, and for job seekers, who enter the European labour market or try to find a new job via adult re-skilling.

Recently, the European Commission's "Pact for Skills",[2] launched in 2020, recognises both the need for a data-driven and targeted approach to reskilling, and the inclusion of public–private partnerships in the process. The overall goal of the Pact is to maximise the impact and effectiveness of skills investment, with a particular focus on upskilling and reskilling in the vocational training sector. For a successful implementation of the Pact, two aspects will be crucial. Firstly, industry needs for specific skills must be made explicit, and secondly, the unique training history of workers needs to be acknowledged. This paper presents an approach to monitoring occupation taxonomies and skill requirements via online labour platform data, in order to offer targeted and near-real time reskilling advice to workers, regarding both industry needs and the worker skills required to fulfil them.

ONLINE GENERATED DATA: A NEW WAY TO IMPROVE SKILL MATCHING?

Recent research shows that, in the absence of institutional support, independent professionals today develop new skills incrementally, adding closely related skills to their existing portfolio (Lehdonvirta, Margaryan, & Davies, 2019). That particular study

[2] https://ec.europa.eu/social/main.jsp?catId=1517&langId=en

examined the skill development of freelancers using online labour platforms, that is, global marketplaces that match millions of buyers and sellers of digitally delivered work in various occupational domains (Horton, 2010). The sellers of work on these platforms are either people in regular employment who are earning additional income by "moonlighting" as online freelancers, or they are self-employed independent contractors. The buyers of work range from individuals and early-stage startups to Fortune 500 companies (Corporaal & Lehdonvirta, 2017). OLPs can be divided into microtask platforms (for example, Amazon Mechanical Turk), where payment is on a piece-rate basis, and freelancing platforms, such as UpWork, where payment is on an hourly or milestone basis (Lehdonvirta, 2018). Between 2017 and 2020, the global market for online labour grew by approximately 50% (Kässi & Lehdonvirta, 2018), with more than 160 million workers engaged worldwide (Kässi et al., 2021). In light of the COVID-19 pandemic and its significant economic repercussions across industries (Stephany et al., 2020a), OLPs continue to increase in popularity due to a general trend of work at distance (Stephany et al., 2020b).

The idea of studying OLP data for skill monitoring is still very novel, and a few social data science scholars have explored alternative sources of online generated data for investigating skill formation. De Mauro et al. (2018), for example, have examined the skill complexity of the new profession of data science with data retrieved from various job boards. Similarly, Calanca et al. (2019) demonstrate the increasing relevance of soft skills in a large body of online job vacancies. Bastian et al. (2014), on the other hand, make use of data from LinkedIn, the world's most popular professional online social network, to compare the relevance of certain "hard skills" across industry domains. In addition to OLP data, information from online job vacancy portals, such as Indeed or Glassdoor, or professional social network sites, like LinkedIn, could be used to inform skill development and warn of the emergence of novel occupational domains. However, all these data avenues have particular advantages and shortcomings in terms of their ability to inform the question of skills development, as summarised in Table 1.

Data Source	**Demand Side**	**Supply Side**	**Price Information**	**Broad Coverage**
Online Job Vacancies	✅	❌	?	✅
Networking Sites	❌	✅	❌	?
Online Labour Platforms	✅	✅	✅	❌

Table 1 *Compared to data from online job vacancy sites and career portals, online labour market data allow for the study of both the demand and supply side of work, including relevant information on prices. Source: Stephany & Luckin (2022).*

Online Labour Platforms: Marketplace for Skills

With regard to skill formation on OLPs, Stephany (2021) shows that online labour markets provide relevant data for active labour market policies in the domain of reskilling, as they are characterised by a high level of skill elasticity. This means that, in contrast to employees, for whom market competition and skill premia are, to a large extent, mitigated by or absorbed by the firm, online freelancers have several strong and immediate incentives to acquire a new capability once it becomes marketable. On the one hand, online freelancers who quickly acquire a newly demanded skill can cash in the global market premium attached to this new capacity. For standard employees, however, this positive incentive to develop a new skill might be weaker, as it is mediated by a firm that relies on a small set of locally defined customers only. On the other hand, online workers have a strong incentive to constantly re-skill, as they are exposed to global competition and have little labour protection. Here, again, employees are shielded by their company, which mitigates market competition on the employee level.

Stephany (2021) argues that the high level of skill elasticity has made OLPs a marketplace for skill rebundling – that is, the versatile recombination of previously unrelated individual skills into profitable and novel bundles of competencies. He compares the recombination of various skills to the de- and rebundling of songs:

"During the early days of the commercial Internet, download platforms allowed music lovers to access songs individually without having to acquire the band's entire album. The single item (song) was unbundled from the original bundle (album). Later, at a second stage, streaming platforms, like Spotify, reversed the trick by allowing the (re)bundling of previously unrelated items. Users could listen to songs from different artists for one single price. The mastery of this strategy has made digital entertainment companies superstars firms (Eriksson et al., 2019). Similarly, this paradigm has affected the way we learn new skills. Initially, digital technologies allowed education providers to offer topical online courses (Wulf et al., 2014). In a second stage, e-learning platforms like Udacity performed the rebundling and offered a whole set of topical courses for one single price (Bates, 2019). The acquisition of individual skills (programming in Python or designing a logo) has been detached from its original domain of training (studying informatics or graphic design)."

It could be argued that, for work, OLPs have turned into what streaming providers are for music: Freelancers can sell previously unrelated skills in one single portfolio for one hourly price, as illustrated by Figure 2. The role of the data scientist is a prime example

of how the rebundling of skills from different domains, i.e., visualisation, programming, and statistics, can result in an economically profitable offer.

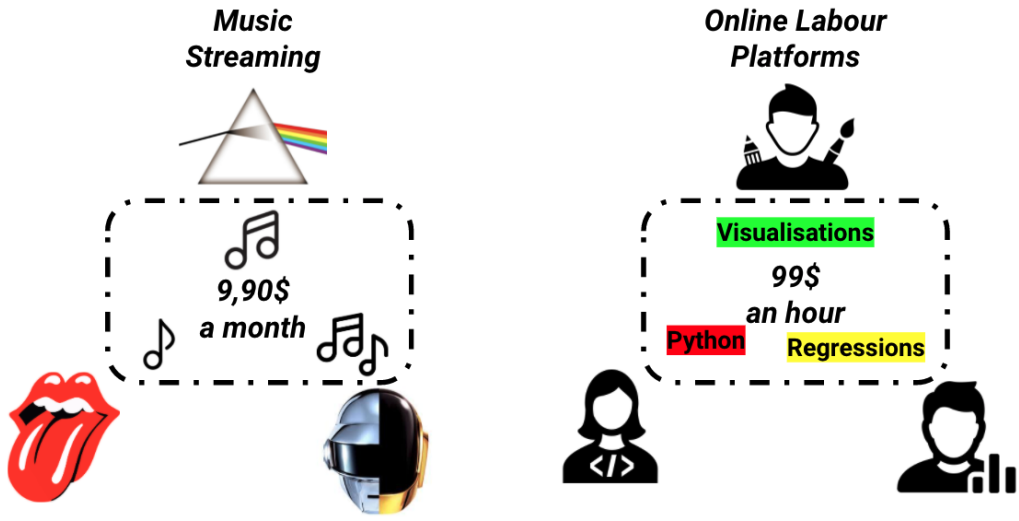

Figure 2: *Music streaming platforms allow users to rebundle individual songs from different artists for a monthly charge (left hand side). Via online labour platforms, previously unrelated domain specific skills appear now rebundled under a new profession, e.g. data scientist (right hand side). Reproduced from Stephany (2021).*

The general success of the data scientist skill bundle is only one of many examples of the profitability of skill rebundling and cross-skilling strategies; i.e., the combination of skills from different occupational domains. Anderson (2017), for example, shows that freelancers with diverse skill portfolios are able to command higher wages, on average. Similarly, Stephany (2020, 2021) shows that the acquisition of a new skill from a different, but adjacent skill domain is related to higher asking wages of online freelancers, but that the benefit of that individual skill depends on its complementarity with the skill bundle the freelancer already has.

The Network Perspective: Capturing the Complexity of Skill Formation

In the situation of rapidly changing skill requirements, systematic oversight is essential. However, individuals often lack foresight about which skills are rising and falling in popularity, which skills are most valuable, and, most importantly, which skills complement their existing portfolio. Individual workers need to avoid getting locked into path dependencies that result in dead ends and that prevent them from re-skilling into new areas (Escobari, Seyal, & Meaney, 2019).

The work by Anderson (2017) and Stephany (2021) shows how OLP data allow us to monitor skill rebundling in a global workforce in near real-time with up-to-date skill

bundles on a granular level. They use the rich toolbox of network analysis for the characterisation of skill relationships. Given a sample of freelancers with multidimensional skill portfolios, the authors construct a human capital network in which skills are nodes and two skills are connected by a link if a worker has both. This skill network provides the researchers with an endogenous categorisation of skills based on their application, as shown by Stephany (2021) (see Figure 3).

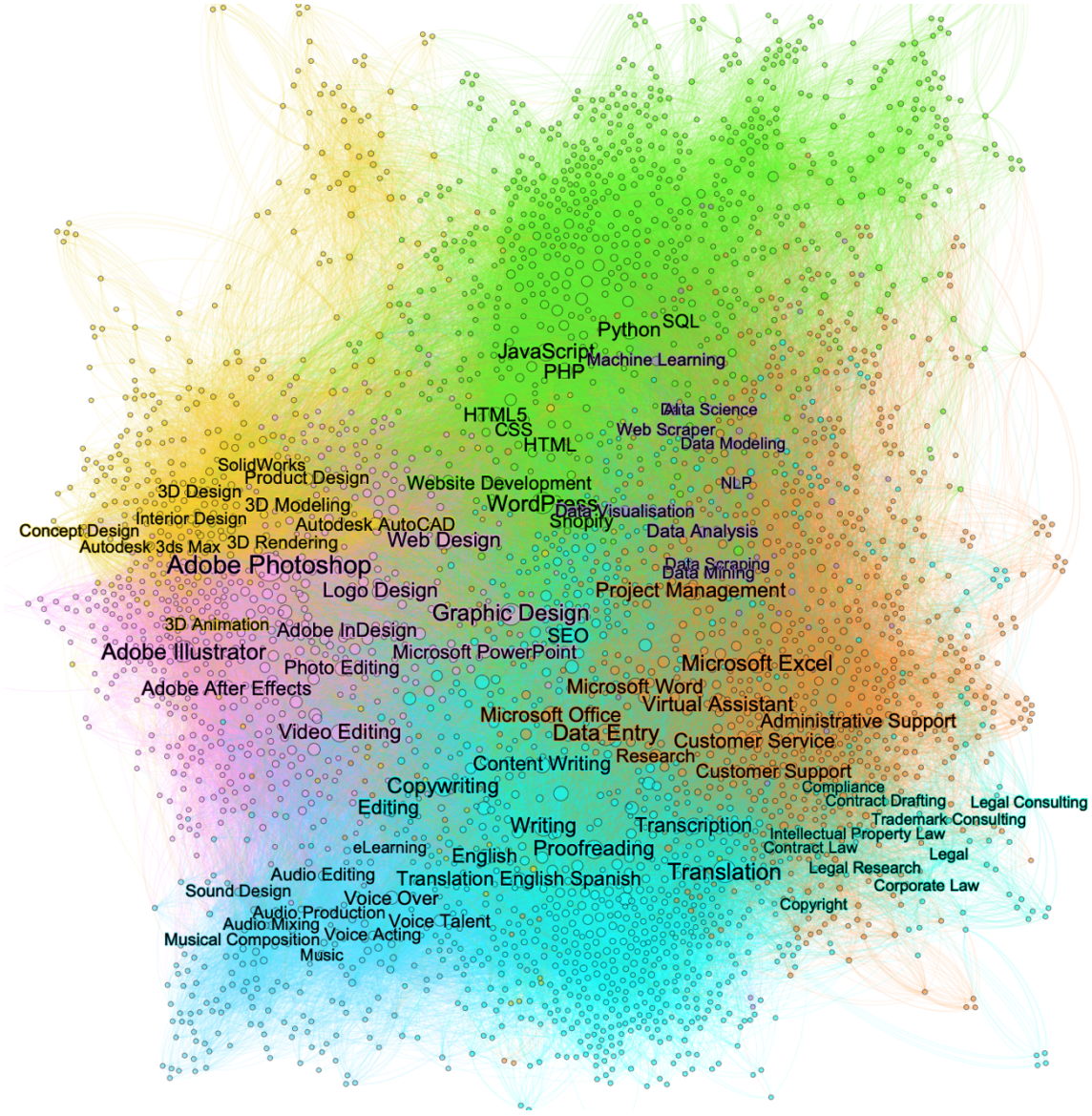

Figure 3: *In a skill network, 3,525 different skills are connected if jointly advertised by the same worker. Skills group into eight clusters with different degrees of centrality (node size), namely: 3D Design (yellow), Admin Support (orange), Audio Design (blue), Data Engineering (green-orange), Graphic Design (pink), Legal (blue), Software and Technology (green), and Translation and Writing (blue). Reproduced from Stephany (2021).*

Most importantly, the taxonomies emerging from the analysis of large-scale OLP data allow researchers to investigate the complementarities of skills. A particular skill could incur different costs and leverage different benefits depending on how it complements the learner's existing skill set. Using the example of programming languages, Stephany (2020) shows that learning Java is of limited economic benefit in the field of data engineering, whereas learning how to program in Python, on the other hand, increases worker wages significantly, as shown in Figure 4. For the field of 3D design, however, the picture flips, and Java yields a much higher contribution to worker wages than the so-called "super star" programming language Python (Grus, 2019). Similarly, costs in learning can be reduced as previously acquired skills may lower the entry barrier into new skill domains, for example via the underlying similarity of language logics across programming languages.

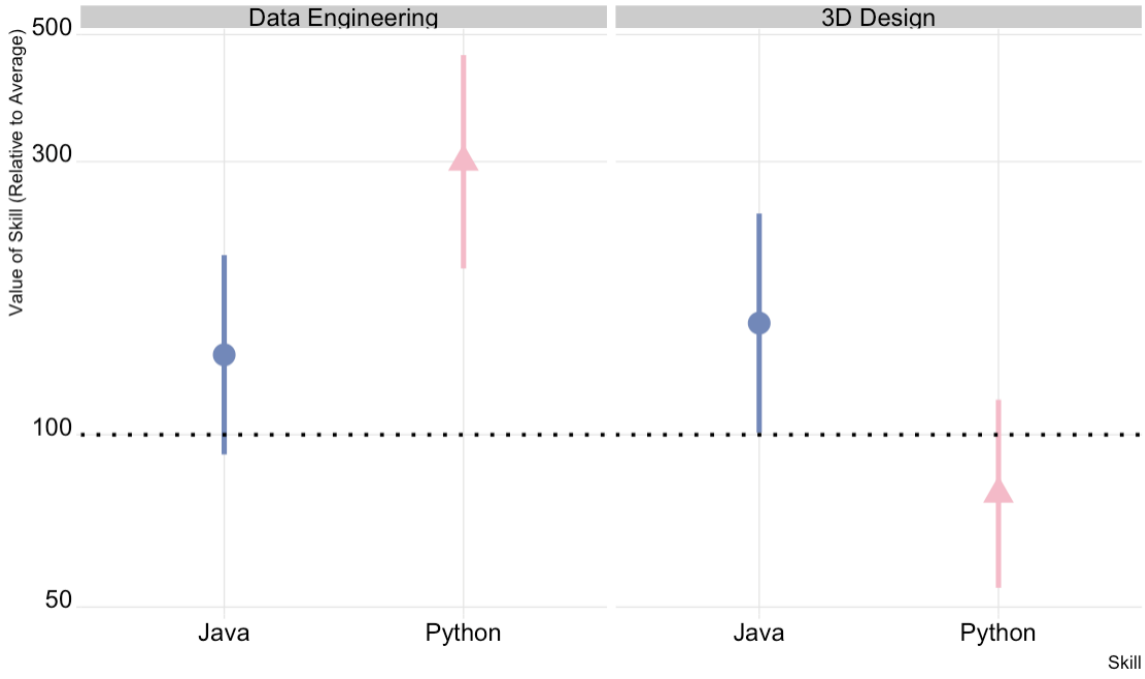

Figure 4: *The benefit of knowing a particular programming language, for example Java or Python, will depend on the specific part of the software sector the worker is in. Knowledge of Python will result in higher wages, on average, in data engineering, but not in 3D design, where Java is favoured. - Source: Stephany (2021).*

3-D designers who know Python can't expect their wage to be higher as a result, but workers that are skilled in Java can more than double their wage, on average. These skill trajectories provide an illustration of what online labour market data allow us to say about the complementarities of learning a new skill. The data we obtain from OLPs can thereby enable us to evaluate the economic benefit of individual skills – based on that person's existing skill bundle – to ultimately sketch optimal individual re-skilling pathways.

Recent analysis of the economic complexity of skill formation (Stephany, 2022) has also questioned the "one-size-fits-all" benefit of certain digital skills. His analysis of 393 popular digital skills across 12 online freelancing occupations shows that only a small set of skills have a clear and positive impact on worker wages. For many skills, the variance in additional value is sizable, and therefore the added value of those skills is not particularly clear (this includes, for example, X and Y). This variance is determined by a skill's features of complementarity and positioning in the "skill space", that is, the complementarity network of skills (Figure 3). The variance is much smaller for skills that have a smaller set of complements and that are more central in the network of skills. The added value of a skill, on the other hand, increases for skills with a small but diverse set of complements, which are applied within a small set of occupational diversity. Examples of such skills include X, Y and Z.

This recent analysis of OLP and online generated data, in general, shows the value of big social data for the purpose of understanding the economic complexity of skill formation, and thereby improving the design of individual reskilling advice (Stephany, 2022). While researchers and policy makers can rely on the growing toolbox of big data processing, statistical analysis, and complexity modelling to process and analyse data, a crucial bottleneck in the further development of online data analysis is (still) data access. The platform providers (i.e. OLPs and social networking sites) that gather useful data on skills, jobs, and occupations often hide information behind paywalls, restrict automated access by researchers and, at times, even issue legal threats after retrieval of data via web-scraping. The integration of platforms into a renewed social contract will be crucial if we are to develop rigorous, data-driven advice on addressing the challenges of the digital skill mismatch.

REFORMING ONLINE DATA ACCESS FOR THE PUBLIC GOOD

Building Better Policies – Individual Reskilling and Real-Time Taxonomies

The review of current research in the field shows that online generated labour market data could allow for multiple advances in understanding labour market developments. It would help us to establish a taxonomy of skills, understand their application and individual complementarity, and enable automated, individual, and far-sighted suggestions on the value of learning a new skill in a future of technological disruption. Hence, policy recommendations are manifold. Efforts should at least include reducing the complexity of individual reskilling and improving occupational taxonomies.

First, reskilling institutions, like the European Centre for the Development of Vocational Training, could be the main beneficiaries of this highly individualised data. The high granularity of online generated data allows us to describe skill profiles of individual workers and track their development over time. It also enables reskilling institutions to assess the individual complementarities of learning a new skill, as shown in Figure 4. Via online generated labour market data, workers with the need to reskill could insert their current skill profile, be located in the landscape of skills, and receive targeted reskilling advice. This allows them to switch to more sustainable occupations that are closely related to their existing skill set with minimal reskilling effort. Via these individualised reskilling recommendations education providers and vocational training organisations could address the urgent need for individualised solutions in adult reskilling. Furthermore, the continuous "pricing" of skills over time, as introduced by Stephany (2021) allows reskilling practitioners to monitor the development of skill values and advise workers on which reskilling to "invest" in.

Secondly, official occupational and skill taxonomies could be improved with near real-time online generated data. As technology creates the demand for novel skills, new occupational clusters can quickly emerge and official taxonomies, such as the European Skills, Competences, and Occupations (ESCO) begin lagging behind. This is bad news for both firms and workers, as professional training providers find it hard to "speak" with the same language as market demand. Online generated data, on the other hand, stems from most recent market development and allows an identification of new occupational clusters including in-demand skills, as shown in Figure 3. Data-driven and near-real time taxonomies could complement conventional classifications. An immediate contribution to current policy efforts would be the continuous (re-)classification of "AI" and "green" skills or jobs, as the "twin-transition" has been identified as a catalyst for active labour market policies (OECD, 2021).

The Data Act – A Step in the Right Direction

In light of the tremendous potential of online generated data, the European Commission's 2022 Data Act (EC, 2022) has identified the importance as well as the complications of accessing business (and platform) data in the interest of the public, while acknowledging the protection of businesses' interests. The Commission intends the Data Act to improve *"means for public sector bodies to access and use data held by the private sector that is necessary for specific public interest purposes. For instance, to develop insights to respond quickly and securely to a public emergency, while minimising the burden on businesses"* (EC, 2022). In detail, Article 15 of the Data Act provides the legal basis for

making private sector data available based on “exceptional needs”. The article clarifies the circumstances of “exceptional needs” as follows (EC 2022, Art. 15):

a. *“where the data requested is necessary to respond to a public emergency;*
b. *where the data request is limited in time and scope and necessary to prevent a public emergency or to assist the recovery from a public emergency;*
c. *where the lack of available data prevents the public sector body or Union institution, agency or body from fulfilling a specific task in the public interest that has been explicitly provided by law;”*

Under this recent proposal for legislation at the time of writing, it would seem that public sector research entities have the right to make use of online generated private sector data, but only under the condition of “exceptional needs”. Indeed, in order to significantly extend the work described in this paper, there would have to be proof that the growing skill mismatch, with all its associated negative impacts for social and economic conditions in the EU, would qualify as a scenario of "public emergency".

While it remains open for legal debate whether a widening skill gap in the EU qualifies as a case of *public emergency,* reducing skill inequalities and creating a digitally literate labour force is certainly of *public interest*. Past examples have shown that independent agreements between public sector body data recipients and private sector data holders are possible, but that they require lengthy negotiations with unnecessary legal risks for both sides. Alternatively, researchers have turned to the practice of accessing platform data without explicit consent via automated modes of data access, such as web-scraping (a term that is, interestingly, not mentioned at all in the EC’s Data Act). As automated data retrieval is not necessarily but often an infringement of a platform’s terms of service, the legal uncertainties involved with web-scraping practices often deter researchers and policy analysts from accessing data that could be used in the interest of the public. Even the Data Act, the EC’s most recent legislative proposal, de-facto allows to rule out web-scraping as a form of data retrieval, as it states, under Article 11, that data recipients who *“deployed deceptive or coercive means or abused evident gaps in the technical infrastructure of the data holder designed to protect the data… shall without undue delay… destroy the data made available by the data holder and any copies thereof; end the production, offering, placing on the market or use of goods, derivative data or services produced on the basis of knowledge obtained through such data, or the importation, export or storage of infringing goods for those purposes, and destroy any infringing goods."* (EC, 2022, Art. 11).

Amending the Data Act in the Interest of the Public

The legislation changes proposed by the Data Act are a step in the right direction, and they demonstrate that the Commission has realised the value of private sector data for the public interest. However, the retrieval and usage of private sector data, such as online labour market or job vacancy data, by public body institutions, is not necessarily enabled under the new legislation. Enforced sharing of private sector data requires the ex-ante proof of a "public emergency", and the current modes of automated data retrieval, such as web-scraping, could be prohibited by the Data Act if they were to be interpreted as coercive or deceptive according to Article 11. In light of this well-intended but potentially contradictory proposal to current EU data legislation, we need to find an agreement that gives public bodies acting in the interest of the public the right to access data (including via modes of web-scraping).

For the case of using platform data in the interest of the public, the bar of data access should be lowered. Like in the case of OLPs or online job vacancy sites, actionable data is already publicly available and can easily be retrieved via automated data retrieval, e.g., web-scraping. Here, the proposed Data Act is contradictive: Data retrieval in the interest of the public after Article 15 could be prohibited by Article 11 when means of retrieval are perceived as coercive or deceptive. Practically speaking, how should researchers proceed when the purpose of their analysis is in the interest of the public but platforms deny cooperation and portray web-scraping as a coercive means of data access in their terms of service? Does Article 15 trump Article 11?

Automated data retrieval should be included in the Data Act as a valuable option if platform providers are too slow, unwilling or technically not capable of sharing data with public sector recipients acting in public interest according to Article 15 of the Data Act. Before scraping data, public sector bodies should need to directly contact both EU authorities and the platform provider explicitly stating which data they plan to retrieve, what their plans of investigation are, and to what extent these align with the public interest. At the same time, they are required to assure that neither the business interest of the platform nor privacy of the respective platform users are at risk. In case that platform providers do not reply to this request or indicate that they are unable or not willing to share the data, public sector bodies should be allowed to start retrieving the data via modes such as web-scraping, if possible.

Ideally, partnerships between public sector bodies and platform providers, like they are already practised in some cases, should become the gold standard in the EC, as they enable targeted access to the data in need, prevent unnecessary legal costs, and have unique opportunities in outreach. Public sector data recipients should be able to negotiate independently with platforms, as they best know which data they need and this

practice would reduce unnecessary administrative burden on all sides to a minimum. The EC could provide guidelines and forms for data recipients when requesting a data sharing framework with platforms, which allows them to describe how they balance Article 15 with other potentially conflicting legislations, such as a violation of platform terms of reference or General Data Protection Regulations. The establishment of independent data ethic boards within the public sector body that govern and document this process is advisable. In addition, the EC could support smaller platforms that are limited in financial and technical resources to provide the data requested in the public interest.

CONCLUSION

In summary, the growing (digital) skill gap is only one of many societal challenges that could be described and potentially eased with the use of online generated platform data. Similar data-savvy investigations around the issues of gentrification (Jain et al., 2021, using Airbnb data) and social media polarisation (Darius & Stephany, 2021, using Twitter data) have highlighted the matchless value of online generated platform data for social data science analysis. Early investigations on the complex ecosystem of skill formation show that online data can indeed be a valuable tool for designing sustainable reskilling policies. To leverage the full potential of this resource the EC's Data Act needs to be amended in order to make public interest its focal point, allowing data access via web-scraping, while enabling strategic public–private partnerships. Only if platform providers can be integrated in this renewal of the social contract, will the true value of online generated data be released for the benefit of society.

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.
- Allen, I. E., & Seaman, J. (2015). Grade Level: Tracking Online Education in the United States. Babson Survey Research Group. Babson College, 231 Forest Street, Babson Park, MA 02457.
- Anderson, K. A. (2017). Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences*, *114*(48), 12720-12724.
- Autor, D. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives* 29, no. 3, 3-30.
- Autor, D. (2019). Work of the past, work of the future. National Bureau of Economic Research (2019)Working Paper # 25588.

- Baldwin, R., & Forslid, R. (2020). *Globotics and development: When manufacturing is jobless and services are tradable* (No. w26731). National Bureau of Economic Research.
- Bastian, M., Hayes, M., Vaughan, W., Shah, S., Skomoroch, P., Kim, H., & Lloyd, C. (2014). Linkedin skills: large-scale topic extraction and inference. In Proceedings of the 8th ACM Conference on Recommender systems (pp. 1-8).
- Bates, T. (2019). What's right and what's wrong about Coursera-style MOOCs. EdTech in the Wild.
- Brunello, G., & Rocco, L. (2017). The effects of vocational education on adult skills, employment and wages: What can we learn from PIAAC?. SERIEs, 8(4), 315-343.
- Brunello, G., & Wruuck, P. (2019). EIB Working Papers 2019/05-Skill shortages and skill mismatch in Europe: A review of the literature (Volume 2019/5).
- Calanca, F., Sayfullina, L., Minkus, L., Wagner, C., & Malmi, E. (2019). Responsible team players wanted: an analysis of soft skill requirements in job advertisements. EPJ Data Science, 8(1), 1-20.
- Cappelli, P. (2014). Skill gaps, skill shortages and skill mismatches: Evidence for the US (No. w20382). National Bureau of Economic Research.
- Card, D., & DiNardo, J. E. (2002). Skill-biased technological change and rising wage inequality: Some problems and puzzles. Journal of labor economics, 20(4), 733-783.
- Collins, A., & Halverson, R. (2018). Rethinking education in the age of technology: The digital revolution and schooling in America. Teachers College Press.
- Corporaal, G.F., & Lehdonvirta, V. (2017) Platform Sourcing: How Fortune 500 Firms are Adopting Online Freelancing Platforms. Oxford: Oxford Internet Institute.
- Darius, P., & Stephany, F. (2021). How the Far-Right Polarises Twitter:'Hashjacking'as a Disinformation Strategy in Times of COVID-19. In *International Conference on Complex Networks and Their Applications* (pp. 100-111). Springer, Cham.
- De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing & Management,54(5), 807-817.
- European Union: European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act), 23 February 2022, COM(2022) 68 final, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0068 [accessed 21 April 2022]
- Escobari, M., Seyal, I., & Meaney, M. (2019). Realism about Reskilling: Upgrading the Career Prospects of America's Low-Wage Workers. Workforce of the Future Initiative. Center for Universal Education at The Brookings Institution.
- EIB, (2018), Retooling Europe's Economy, Luxembourg.

- Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). Spotify teardown: Inside the black box of streaming music. Mit Press.
- Horton, J. J. (2010). Online Labor Markets. In A. Saberi (Ed.), Internet and Network Economics (pp. 515–522). Springer: Berlin Heidelberg.
- Gorwa, R. (2019). What is platform governance?. *Information, Communication & Society*, *22*(6), 854-871.
- Illanes, P., Lund, S., Mourshed, M., Rutherford, S., & Tyreman, M. (2018). Retraining and reskilling workers in the age of automation. McKinsey Global Institute.
- Jain, S., Proserpio, D., Quattrone, G., & Quercia, D. (2021). Nowcasting gentrification using Airbnb data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-21.
- Kahn, S., & Ginther, D. (2017). Women and STEM (No. w23525). National Bureau of Economic Research.
- Kässi, O., & Lehdonvirta, V. (2018). Online labour index: Measuring the online gig economy for policy and research. *Technological forecasting and social change*, 137, 241-248.
- Kässi O, Lehdonvirta V. & Stephany F. How many online workers are there in the world? A data-driven assessment. Open Res Europe 2021, 1:53 https://doi.org/10.12688/openreseurope.13639.4
- Kester, L., Lehnen, C., Van Gerven, P. W., & Kirschner, P. A. (2006). Just-in-time, schematic supportive information presentation during cognitive skill acquisition. *Computers in Human Behavior*, 22(1), 93-112.
- Krueger, D., & Kumar, K. B. (2004). Skill-specific rather than general education: A reason for US–Europe growth differences?. Journal of economic growth, 9(2), 167-207.
- Lawrence, J. A., & Ehle, K. (2019). Combatting Unauthorized Webscraping-The remaining options in the United States for owners of public websites despite the recent hiQ Labs v. LinkedIn decision. *Computer Law Review International*, *20*(6), 171-174.
- Lehdonvirta, V., Margaryan, A., & Davies, H. U. W. (2019). Skills formation and skills matching in online platform work: policies and practices for promoting crowdworkers' continuous learning (CrowdLearn).
- Nedelkoska, L. and G. Quintini (2018), "Automation, skills use and training", OECD Social, Employment and Migration Working Papers, No. 202, OECD Publishing, Paris, https://doi.org/10.1787/2e2f4eea-en.
- Organisation for Economic Co-operation and Development. (2021). *Designing Active Labour Market Policies for the Recovery*. OECD Publishing.

- Pouliakas, K. (2021). Understanding Technological Change and Skill Needs: Technology and Skills Foresight. Cedefop Practical Guide 3. Cedefop-European Centre for the Development of Vocational Training.
- Schmidpeter, B., & Winter-Ebmer, R. (2018). How do automation and offshorability influence unemployment duration and subsequent job quality?
- Stephany, F. & R. Luckin (2022) 'Is the workforce ready for the jobs of the future? Data-informed skills and training foresight', Working Paper 07/2022, Bruegel.
- Stephany, F., Neuhäuser, L., Stoehr, N., Darius, P., Teutloff, O., & Braesemann, F. (2022). The CoRisk-Index: a data-mining approach to identify industry-specific risk perceptions related to Covid-19. *Humanities and Social Sciences Communications*, *9*(1), 1-15.
- Stephany, F. (2021). One size does not fit all: Constructing complementary digital reskilling strategies using online labour market data. Big Data & Society, 8(1). https://doi.org/10.1177/20539517211003120.
- Stephany, F., Dunn, M., Sawyer, S., & Lehdonvirta, V. (2020). Distancing Bonus Or Downscaling Loss? The Changing Livelihood of Us Online Workers in Times of COVID-19. Tijdschrift voor economische en sociale geografie, 111(3), 561-573.
- Stephany, F., Kässi, O., Rani, U., & Lehdonvirta, V. (2021). Online Labour Index 2020: New ways to measure the world's remote freelancing market. Big Data & Society. https://doi.org/10.1177/20539517211043240.
- Tamm, M. (2018). Training and changes in job Tasks. Economics of Education Review, 67, 137-147.
- Wulf, J., Blohm, I., Leimeister, J. M., & Brenner, W. (2014). Massive open online courses. *Business & Information Systems Engineering*, 6(2), 111-114.