Thinking ahead for Europe

Using web data for labour market research

Workshop "Big-Data-Analysen in Wissenschaft und Praxis"
IW Köln, Cologne, 6 March 2017

Karolien Lenaerts
Researcher, CEPS

Outline of the presentation

- Background and motivation
- Available web data sources, their current use and their advantages and limitations
- Our work based on vacancy data and other data from online job boards: vacancies and tags
- Conclusions

Background & Motivation

Why do we consider web-based data sources for labour market research?

Our focus within the InGRID project

- **Unemployment in EU was very high after crisis**
- **Socio-ecological transition affects labour markets:**
  - Not new phenomenon
  - Longstanding academic interest in these dynamics
- **New occupations and skills have emerged, others have become obsolete or are greatly transformed**
- Our work for the InGRID project aimed at a better understanding of these dynamics

Our focus within the InGRID project

- **What are new occupations and skills and how are they identified?**
  - **Conceptualisation**: not always very clear (Crosby, 2002): new, emerging and evolving occupations
  - **Identification**: surveys, employer interviews, monitoring, trade publications, job listings, occupational classification
  - But: often complex, lagged, imprecise, subjective, narrow in focus, irregular updates, data-intensive, ...
- **What about web data?**

Background for our work

- **Advancement of Internet affects labour markets:**
  - Search, matching, selection, recruitment, services, etc.
  - Internet has developed into a research topic
  - Autor (2001), Carnevale et al. (2014), Kuhn (2014)
- **The Internet is increasingly used as both a research platform and data source**
  - Many forms: surveys, experiments, interviews and focus groups, ethnographies (Hooley et al., 2012)
  - For labour research: advocated by Kuhn and Skuterud (2004); Askitas and Zimmermann (2009, 2015)

Available web data sources and their advantages and limitations

What is interesting about web-based data sources?

Potential web data sources and use

- **Many potential data sources:**
  - Online job portals (Vacancies, CVs, other information)
  - Online intermediaries (Platforms)
  - Online surveys, Google Trends, social media data
- **Most studies use job portals and surveys**
  - Other sources have not been overlooked, but their use is more limited, though on the rise
- **Can also be combined (with traditional data)**

Advantages of web data

- **Web data sources bring important advantages when compared with traditional sources:**
  - Data collection, processing and analysis are fast, flexible, easy, cheap and allow for large, diverse samples
  - Data available in real time, and therefore allow to capture current trends
  - Internet enables researchers to fill gaps and to study difficult-to-capture phenomena (e.g. self-employment, on-the-job search)

Web data can be used to overcome issues related to traditional sources and fill gaps where traditional sources are weak or absent

Limitations of web data

- **Yet, there are important limitations:**
  - Issues related to completeness, selection bias and **data representativeness**:
    - Can results be generalised?
    - Biased towards specific sectors, regions, applicants ... ?
    - Very important issue, only little researched so far
  - Ethical, technical and other issues: privacy, anonymity, computer literacy, data quality, data accuracy

Examples of vacancy data

- **Advantages:** Detailed, easy to collect, derive other information from the job portal

- **Limitations:**
  - Incomplete: not all available jobs advertised online, only small part of demand → representative?
  - Data collection/processing: data are volatile, duplicates, not standardised, semantic analysis is complicated
  - Selection issues: not all job seekers use Internet in job search, digital divide, attract only specific users?

Our work based on vacancy data and other information extracted from job portals

Methodological issues and pilots

Collecting data from online job boards

- **Access existing database:**
  - High N
  - Clean, coded data
  - Good coverage if sufficient sources are covered

- **Web crawling (spidering techniques)**
  - Necessary in some cases
  - Possibility to include more variables/metadata
  - Relatively low barrier to entry
  - But data processing can be complicated: parsing data, semantic analysis

What to analyse: vacancies or tags?

- **Vacancies:** assemble and process advertisements and then perform a text analysis
  - pros: highly detailed, real-time information
  - cons: data- and time-intensive

- **Tags:** used to structure information on portal, keep track of tags and number of matching vacancies
  - pros: easy and fast, less data- and time-intensive
  - cons: less details, not possible on every portal

Analysing Vacancies

Two examples

Skill requirements in US

- **Aim**: to map requirements of US employers for 30 most-frequently advertised occupations of different complexities:
  - formal education and specialised training
  - cognitive skills and non-cognitive skills
  - experience
  - other

- **Methodology & data:** 2 million vacancies published on Burning Glass → keywords in vacancies

Skill requirements in US

- Sum of % of ads listing education and skills
- Sum of % of ads listing all requirements

ISCO	Average sum of skills	Average sum of all requirements
1	4.3 (431%)	4.9 (488%)
2	3.9 (390%)	4.6 (464%)
3	3.8 (380%)	4.7 (466%)
4	3.8 (384%)	4.5 (446%)
5	3.8 (380%)	4.5 (447%)
7	3.4 (337%)	4.3 (426%)
8	3.3 (333%)	4.2 (420%)
9	2.4 (240%)	3.1 (310%)

US employers are rather demanding in their job vacancies:

- Positive relation with complexity
- Also for low- / medium-skilled occpations
- A lot of variation
- Top 5: security guards, tellers, event planners, managers and first-line office supervisors
- Education, service skills, experience

IT skills requirements in the US

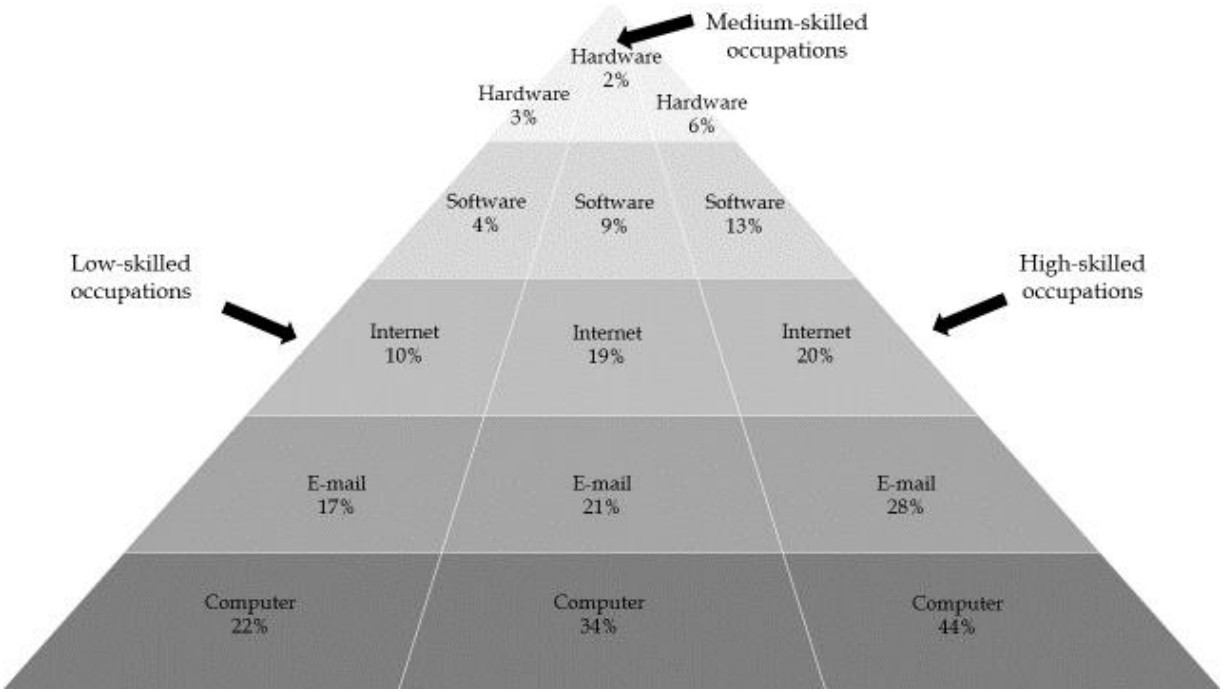
- **Aim**: to investigate importance of digital skills on US labour markets:
  - Three levels: basic and general digital skills, intermediate digital skills (productivity software), advanced digital skills
  - Low-skilled, medium-skilled, high-skilled occupations
  - High unemployment yet many unfilled vacancies: skill gaps and mismatches?

- **Methodology & data:** 2 million vacancies published on Burning Glass → keywords in vacancies

IT skills requirements in the US

Basic and general digital skills: all occupations

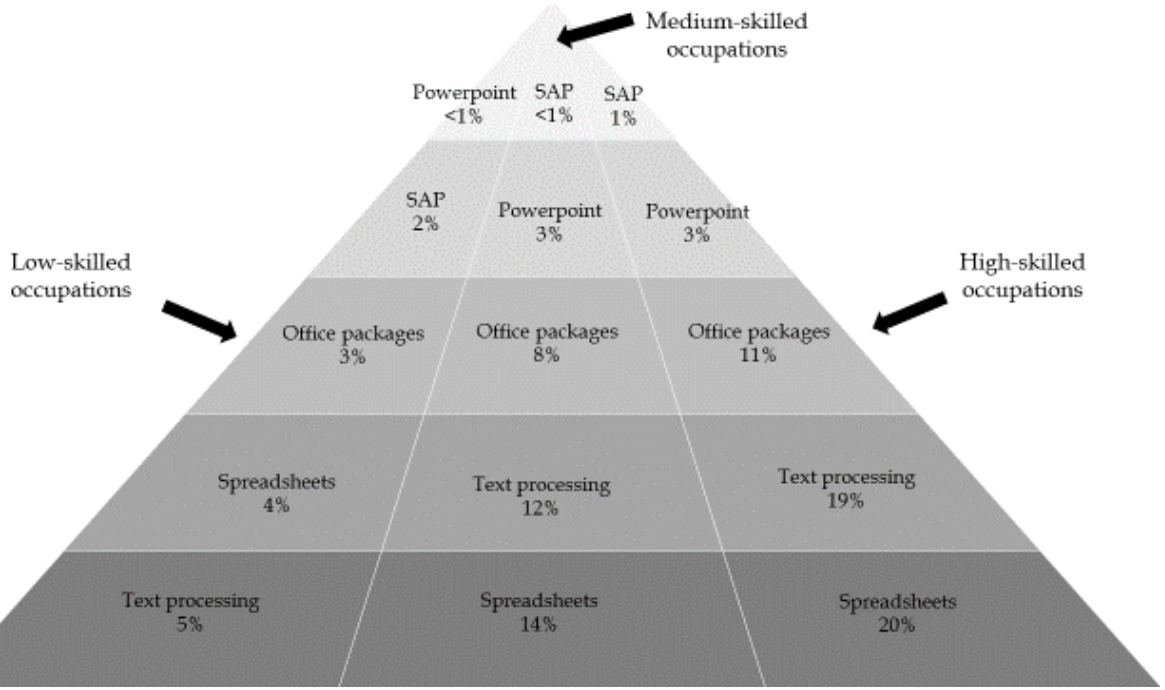
- Computer: 35%
- Software: 9%
- Hardware: 3%
- Internet: 19%
- E-mail: 22%



IT skills requirements in the US

Intermediate digital skills: all occupations

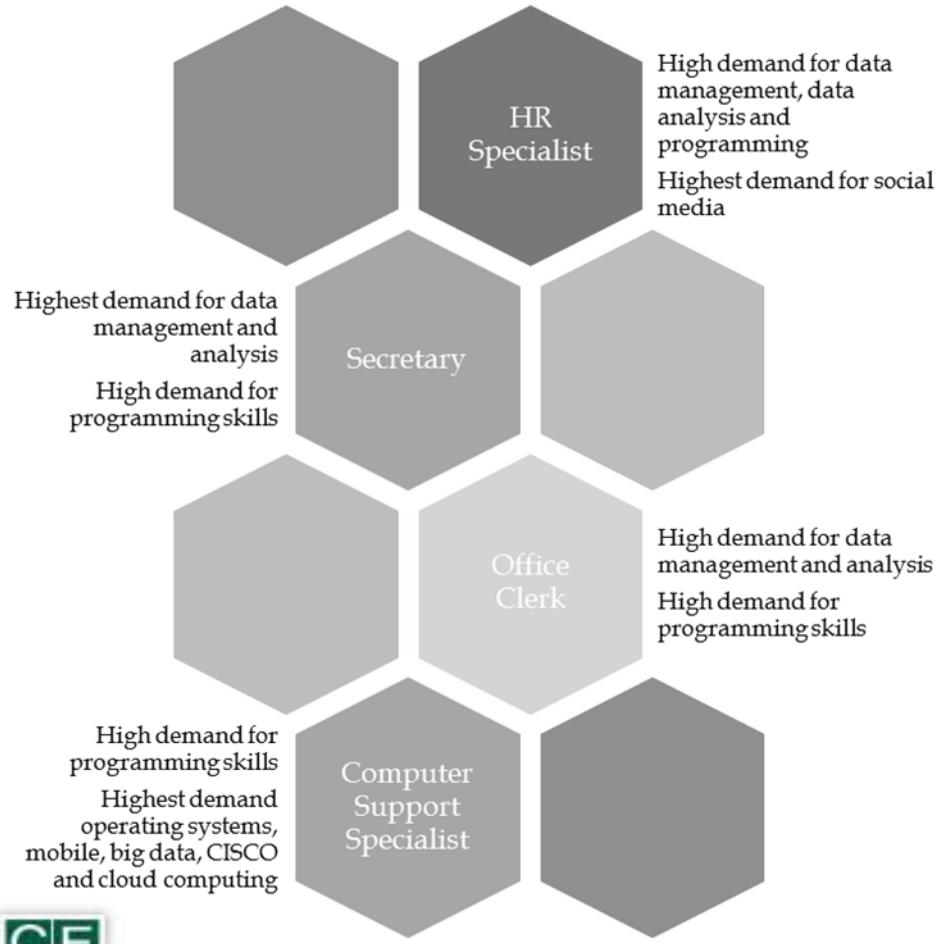
- Word / text processing / MS Word: 13%
- Spreadsheet: 14%
- PowerPoint: 3%
- Office: 9%
- SAP: 1%



IT skills requirements in the US

Advanced digital skills:

- Start from an extensive list of keywords
- Only few vacancies refer to any of these skills (< 3%)
- Databases and data management: 12%
- Higher prevalence for medium- to high-skilled office jobs: secretaries, office clerks, accountants, …



Analysing Tags

Two examples

Language skills in Visegrad region

- **Aim**: to identify demand for foreign language skills in Czech Republic, Hungary, Poland, and Slovakia

- **Methodology & data:** analyse tags on 4 job boards (linked to 74,000 vacancies), then focus on a subset of occupations available in all four countries

Language skills in Visegrad region

Why the Visegrad region?

- Common roots, close collaboration
- Open to international trade and FDI
- EU Members since 2004, SK in EMU
- Recent migration flows

Demand side:

- English as international business language
- Strong economic and historical ties with Germany and Austria: German
- Shared border with Soviet Union: Russian
- Other languages: French, Spanish, Italian
- Languages of neighbouring countries

Supply side:

- Main national languages not commonly spoken in Europe, no bilingual countries

	English	German
EU27	**38%**	**11%**
Czech Republic	27%	15%
Hungary	20%	18%
Poland	33%	19%
Slovakia	26%	22%

Language skills in Visegrad region

Results across all occupations (74,000 vacancies):

	CZ	HU	PL	SK	Total
English	**28.19%**	**38.92%**	**63.99%**	**49.26%**	**51.89%**
German	**10.15%**	**10.86%**	**12.45%**	**14.59%**	**12.36%**
French	0.65%	1.25%	3.56%	1.50%	2.33%
Italian	0.19%	0.67%	1.65%	0.55%	1.05%
Spanish	0.15%	0.52%	2.13%	0.48%	1.23%
Russian	0.54%	0.21%	1.6%	0.48%	0.96%

Results for a subset of 59 occupations available in all four countries (66,000 vacancies):

- English: Positive relationship between demand and complexity of occupation and with median hourly wages
- No such relations for German

Occupations observatory

- **Concept:**
  - Traditional methods and data sources often fall short to (swiftly) capture new occupations
  - Official occupational classifications are generally not updated regularly (e.g. every 10 / 20 years)
  - Our pilot aims to propose a new methodology, based on real-time information

Occupations observatory

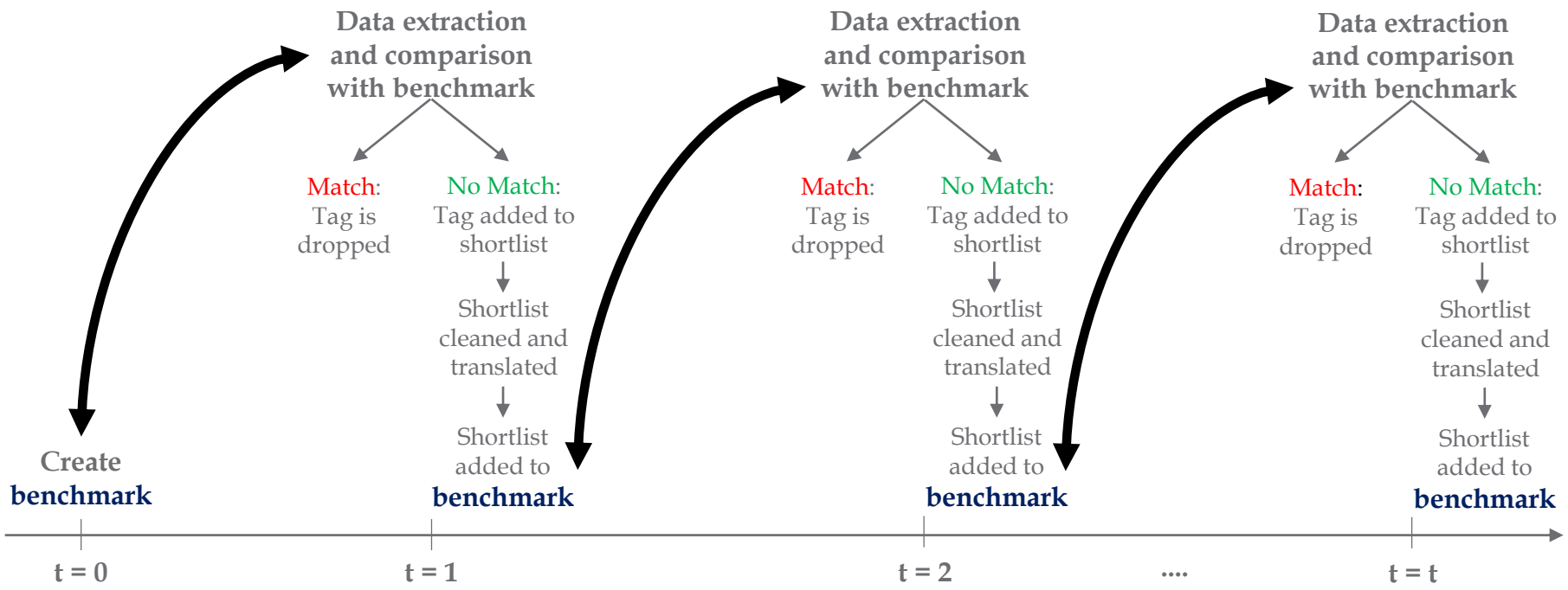
- **Methodology & data:**
  - **Underlying data come from online job boards:**
    - Meta-data of job boards instead of vacancies
    - BE, CZ, DE, DK, ES, FR, HU, IT, PL, SK, and UK
  - **Methodology: based on changes in the occupational classification of portal ('tags'):**
    - Tags are used to structure information on the portal
    - Keep track of tags (e.g. added) and matching vacancies
    - How: published online, query API, web crawling
    - Easy and fast, not so data- and time-intensive

Occupations observatory

- **Methodology & data: steps**

1. **Create the benchmark**: extract occupational classification from the 11 online job portals
2. **Extract occupational classification** again from each portal at the beginning of every month: what new tags were added?
3. **Translate new tags into English** using Google Translate, checked by native speakers and corrected if necessary
4. This generates a longlist of potentially new occupations by country, further cleaned and analysed to **create a shortlist**
5. In-depth assessment of tags on shortlist, summarised in an **occupation card** (responsibilities, requirements, presence)

Occupations observatory



Occupations observatory

- **Early results:**
  - **Successful proof-of-concept**:
    - Feasible to identify potentially new occupations on the basis of the occupational structure of job boards
    - Example of SK: 16 tags, e.g. drug safety specialist
  - **Avenues for further improvements:**
    - Monthly periodicity is too frequent
    - Combination with vacancies, other sources
    - Automation of the process

Conclusions

What did we learn from our work and other studies?

Conclusions

- **Internet data are increasingly being used as a data source**, also in the social sciences
- Several authors have **confirmed their potential in the field of labour economics**
- Most studies have focused on **data extracted from online job boards and surveys**, yet other sources also have a lot of potential (e.g. social networks, Google Trends)
- **The future of web-based labour market research is bright**

Thank you very much for your attention!

More details are available on www.ceps.eu and www.inclusivegrowth.be

karolien.lenaerts@ceps.eu