Use of Big Data for different analyses of occupational and skills dynamics in the labour market.

Combination of data from different sources

Speaker: Mauro Pelucchi

29/11/2022
Data production system

Tunisia (04/2020 to 07/2022)
680,191 OJVs - > 175,203 deduplicated

Ukraine general (04/2020 to 07/2022)
2,571,655 OJVs - > 1,304,262 deduplicated

Georgia (04/2021-07/2022)
129,271 OJVs - 84,817 deduplicated

Egypt (new)
1,307,678 OJVs – 391,701 deduplicated

Kenya (new)
(collection started in september 2022)
Collecting and decoding labor market data

Real-time job market data offer up-to-date insights not possible through traditional sources

Capture job market data
Tagging and structuring
A common language
Drawing conclusions

Data ontology allows for comparisons
Insight from in-demand skills and real-life career patterns
Overall Data Flow

Ingestion

Pre-Processing

Information Extraction

ETL

Presentation Area
Data Ingestion

- **Goal:**
  - Collect massive amount of heterogeneous data from several unpredictable number of sources

- **Challenges:**
  - Guarantee data completeness and consistency

- **Approach**
  - Develop a multi-technique framework (crawling, scraping, API) to fit different website characteristics: volume of vacancies, different languages, technology structure, non-invasive approach and policy agreements
  - Prevent data sources losses, via redundancy policies
  - Detect and collect metadata to improve information value

- **Features:**
  - To grant process governance in data Ingestion phase, we developed a monitoring system with a scheduling tool and an alerting module

---

ETF
European Training Foundation
Data Pre-Processing

• **Goal**
  • Feed information extraction phase with proper data

• **Challenges**
  • Measure, monitor and increase Data Quality, to maximize completeness, consistency, complexity, timeliness and periodicity

• **Approach**
  • Develop a multi-phase pipeline, focused on:
    • Vacancy Detection: analyze website page to select only content referred to vacancies
    • Deduplication: detect duplicated vacancy posts to obtain a single vacancy entity
    • Date detection: identify release and expire dates through vacancy description analysis
    • Vacancy duration: method to define expire date, when not explicitly available

• **Features**
  • Guarantee Data Quality during all processing phases
Data Classification

- **Goal:**
  - Extract and structure information from data, to be provided to the presentation layer

- **Challenges:**
  - Handle massive amount of heterogeneous data written in different languages

- **Approach:**
  - Develop an adaptable framework, language dependent, tailored on different information features. Some relevant challenges:
    - Occupation feature classification: combined methods such as Machine Learning, Topic Modeling and Unsupervised Learning
    - Skill feature classification: another different combined methods, such as Text Analysis with corpus based or Knowledge based similarity

- **Features:**
  - Guarantee Explainable information extraction, logging classification methods and relevant features.
Presentation

- **Goal:**
  - Provide data in an Analytic environment, to support users across their different analysis needs

- **Challenges:**
  - Personalized data access to different types of stakeholders

- **Approach**
  - Develop different navigation paths, for different users:
    - **Data scientists/analysts:** access data via a Discovery Lab environment, to ensure free data analysis needs
    - **Decision makers:** access information through storytelling and dynamic dashboards
Dashboard with point and click capabilities

Labour Market Portal
(from 2017 Big Data Hackathon – Poland Team)

Map (from 2017 Big Data Hackathon – French Team)

Dashboard (from 2017 Big Data Hackathon – Norwegian Team)
Lab session
Revealed comparative advantage
RCA

Given a set of occupations $\tilde{O} = \{o_k, k = 1, \ldots, m\}$, a set of skills $\tilde{S} = \{s_j, j = 1, \ldots, p\}$, and a matrix $M_{m \times p}$ that contains a value of $sf$ for each pair of occupations $o_k \in \tilde{O}$ and skills $s_j \in \tilde{S}$, $rca$ for $o_i$ and $s_l$ is defined as:

$$rca(o_i, s_l) = \frac{sf(o_i, s_l)/\sum_{j=1}^{p} sf(o_i, s_j)}{\sum_{k=1}^{m} sf(o_k, s_l)/\sum_{k=1}^{m} \sum_{j=1}^{p} sf(o_k, s_j)}$$
Occupation similarity

From the notebook:

- We will select a country
- For the country we will calculate the RCA
- And we will see our create bridge between occupations
RCA

In this notebook we will see the computation of RCA and now use it to calculate a KPI about the distance from occupations.

Thanks to Simone Perigo for the Notebook

declare @sql text = N'

-- SQL to calculate the KPI

SELECT idcountry, idesco_level_4, esco_level_4, idescoskill_level_3, escoskill_level_3, COUNT(DISTINCT general_id) AS num_qry
FROM hr_skill_analysis
WHERE idesco_level_4 IN ('1330', '23521', '23512', '23513', '23514', '23519', '23521', '23523', '23528', '23511', '23512', '23533', '23514', '23521', '23522')
AND year_grab_date = 2020
AND idcountry IN ('IT', 'DE', 'FR', 'ES', 'EE', 'CZ')
AND idescoskill_level_3 IS NOT NULL
AND (idskilled) = 0
AND job_occupation = 0

GROUP BY idcountry, idesco_level_4, esco_level_4, idescoskill_level_3, escoskill_level_3

-- SQL to fetch data from default.data_rca.csv

SELECT * FROM default.data_rca.csv LIMIT 10
'

-- Execute SQL

EXEC sp_executesql @sql

-- Display the results

SELECT * FROM default.data_rca.csv LIMIT 10

-- Spark Jobs

<table>
<thead>
<tr>
<th>idcountry</th>
<th>idesco_level_4</th>
<th>esco_level_4</th>
<th>idescoskill_level_3</th>
<th>escoskill_level_3</th>
<th>num_qry</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>1330</td>
<td>Information and communications technology service managers</td>
<td>STACK_157</td>
<td>elasticsearch</td>
<td>112</td>
</tr>
<tr>
<td>CZ</td>
<td>23511</td>
<td>Systems analysts</td>
<td>ESCOv1.10485</td>
<td>use digital device operating systems</td>
<td>116</td>
</tr>
<tr>
<td>ES</td>
<td>23512</td>
<td>Information and communications technology user support technicians</td>
<td>ESCOv1.3727</td>
<td>manage system security</td>
<td>272</td>
</tr>
<tr>
<td>DE</td>
<td>23513</td>
<td>Computer network and systems technicians</td>
<td>ESCOv1.529</td>
<td>cloud technologies</td>
<td>6469</td>
</tr>
<tr>
<td>FR</td>
<td>1330</td>
<td>Information and communications technology service managers</td>
<td>STACK_40</td>
<td>eclipse</td>
<td>65</td>
</tr>
<tr>
<td>DE</td>
<td>23513</td>
<td>Computer network professionals</td>
<td>ESCOv1.4158</td>
<td>quality standards</td>
<td>1080</td>
</tr>
</tbody>
</table>

Showing all 10 rows.
rawdata.io

https://app.rawgraphs.io/

<table>
<thead>
<tr>
<th>Load your data</th>
<th>annunci_aggregati.csv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paste</td>
<td></td>
</tr>
<tr>
<td>Upload a file</td>
<td></td>
</tr>
<tr>
<td>From URL</td>
<td></td>
</tr>
<tr>
<td>Try our samples</td>
<td></td>
</tr>
</tbody>
</table>

| 1. "Professionals", "Internship", "5236" |
| 2. "Service and sales workers", "Temporary", "8921" |
| 3. "Skilled agricultural, forestry and fishery workers", "Internship", "8" |
| 4. "Managers", "Temporary", "4578" |
| 5. "Clerical support workers", "Temporary", "18736" |
| 6. "Armed forces occupations", "Temporary", "2" |
| 7. "Skilled agricultural, forestry and fishery workers", "Self Employment", "22" |
| 8. "Elementary occupations", "Temporary", "13820" |
| 10. "Clerical support workers", "Self Employment", "2053" |
| 11. "Clerical support workers", "Permanent", "4802" |
| 12. "Managers", "Permanent", "9173" |
| 13. "Technicians and associate professionals", "Internship", "8061" |

39 records in your data have been successfully parsed!

Your data seems ready to go. But if you want to stack it anyway, click here.
A sunburst is similar to the treemap, except it uses a radial layout. The root node of the tree is at the center, with leaves on the circumference. The area (or angle, depending on implementation) of each arc corresponds to its value. Based on http://bl.ocks.org/mbostock/4063423

https://app.rawgraphs.io/
Strong role for professionals

Source: ETF Dataset - 2020-2022 Sept
Detailed occupations

UA - ETF Dataset - Unique Job postings 2020-2022

- Business services agents not elsewhere classified: 117,073
- Accounting associate professionals: 88,901
- Stock clerks: 67,969
- Commercial sales representatives: 43,165
- Shop sales assistants: 40,112
- Sales and marketing managers: 39,638
- Software developers: 38,330
- Systems analysts: 34,896
- Special needs teachers: 34,820
- Travel attendants and travel stewards: 27,760

TN - ETF Dataset - Unique Job postings 2020-2022

- Software developers: 10,917
- Engineering professionals not elsewhere classified: 6,746
- Administrative and executive secretaries: 6,324
- Contact centre salespersons: 6,027
- Personnel clerks: 5,736
- Business services agents not elsewhere classified: 5,617
- Clerical support workers not elsewhere classified: 4,842
- Systems analysts: 4,587
- Accounting associate professionals: 4,462
Skills and ESCO

ESCO taxonomy allows to group skills into categories. Major distinction between hard and soft skills.

- **Hard skills:** job-specific skills and competences that are needed to perform a specific job or task (examples are knowledge of specific software or instruments, specific manual abilities etc.)
- **Soft skills:** transversal in nature and refer to the capacity of individuals to interact with others and the environment (examples are communication skills, problem solving etc.).
Transversal skills in ESCO

• **Thinking skills.** Ability to apply mental processes and reasoning to solve complex problems, to increase knowledge and to perform complex tasks.

• **Social interaction.** Ability to develop interact and engage with colleagues, clients and customers.

• **Application of knowledge.** General application of skills commonly used in the workplace and in learning; knowledge of the organization and the working environment.

• **Attitudes and values.** Person's work style, preferences and work-related beliefs that underpin behaviour. Examples are ability to adapt to change, to work independently, to meet commitments etc.

We add to transversal skills also Languages
Skills: transversals vs non

- **UA**: 25.96% Transversal, 74.04% Non transversal
- **TN**: 36.22% Transversal, 63.78% Non transversal
- **EG**: 25.40% Transversal, 74.60% Non transversal
- **GE**: 29.13% Transversal, 70.87% Non transversal
Most relevant transversal skills - UA

adapt to change
meet commitments
work independently
cope with pressure
manage quality

demonstrate good manners
demonstrate consideration
follow ethical code of conduct
deal with uncertainty
demonstrate willingness to learn

persist
work efficiently
manage frustration
attend to hygiene
demonstrate curiosity
make an effort

attend to detail
Most relevant non transversal skills - UA
Labour Market
Challenging factors
Labour Market – challenging factors

- Skills Evolution
- New Emerging Occupations
- Job Automatisation/Replacement
- .....
Compute Skills Rates and Relevance

**Skills Rate** estimate the incidence of Digital, Hard non-Digital and Soft skills in a single occupation;

**Skills Relevance** defines the weight of each skill for the considered occupation.

**Idea:** Exploit the informative power of the OJV for computing the Skill Rate and skill Relevance.

Skills Rate and Relevance methodology was defined and used by CRISP in 2017 and 2018 for the Italian Observatory of Digital Skills promoted by the Italian ICT associations.
Definition and computation method

\[ \text{skills Rate} = \frac{\text{frequency of skills (digital or hard or soft)}}{\text{frequency of digital + hard + soft skills}} \]

\[ \text{skills relevance} = \frac{n^\circ \text{ of vacancies of (occupation, skill)}}{n^\circ \text{ of vacancies of (skill)}} \times \frac{n^\circ \text{ of vacancies of (occupation, skill)}}{n^\circ \text{ of vacancies of (occupation)}} \]
Skills Rate for some occupations

- Mechanical engineers
- Software developers Systems
- Analysts
- Computer network professionals Advertising
- And marketing professionals Personnel and careers
- Professionals

- hard_skill_rate
- soft_skill_rate
- digital_skill_rate
Final remarks

• Knowledge become crucial to observe the challenging factors of LM:
  Skills Evolution
  New Emerging Occupations
  Job Automatisation/Replacement
  ...

• OJV are complementary to the other sources that we normally use to understand LM phenomena

• We are only at the beginning of exploiting the informative potential of OJV - Big Data
Not All Shortages are Gaps

Deeper Characterization of the Skills Gap

![Graph showing supply and demand ratios for different sectors.](chart.png)
# Identifying Future Skill Demands

**A Range of Lenses for Tracking Emerging Trends**

<table>
<thead>
<tr>
<th></th>
<th>Top IT Skills (Total postings)</th>
<th>Highest Paying IT Skills (Mean advertised salary)</th>
<th>Fastest Growing IT Skills (24 month projections)</th>
<th>Hardest to Fill IT Skills (Mean posting duration)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SQL</td>
<td>Zookeeper</td>
<td>TensorFlow</td>
<td>Public Cloud Security</td>
</tr>
<tr>
<td>2.</td>
<td>Java</td>
<td>TensorFlow</td>
<td>General Data Protection Regulation (GDPR)</td>
<td>Infrastructure as a Service (IaaS)</td>
</tr>
<tr>
<td>3.</td>
<td>JavaScript</td>
<td>Scala</td>
<td>Kubernetes</td>
<td>Cloud Technology Architecture</td>
</tr>
<tr>
<td>4.</td>
<td>Linux</td>
<td>AWS Redshift</td>
<td>Spring Boot</td>
<td>Cloud Infrastructure</td>
</tr>
<tr>
<td>5.</td>
<td>Python</td>
<td>AWS DynamoDB</td>
<td>Webpack</td>
<td>Ansible</td>
</tr>
<tr>
<td>6.</td>
<td>Data Analytics</td>
<td>Go Programming Language (Golang)</td>
<td>AWS Lambda</td>
<td>Apache Mesos</td>
</tr>
<tr>
<td>7.</td>
<td>Salesforce</td>
<td>Pig</td>
<td>Salesforce Lightning</td>
<td>Data Protection Planning</td>
</tr>
<tr>
<td>8.</td>
<td>C#</td>
<td>Apache Mesos</td>
<td>Redux</td>
<td>Work Breakdown Structure</td>
</tr>
<tr>
<td>9.</td>
<td>Scrum</td>
<td>AWS CloudFormation</td>
<td>Financial Microservices</td>
<td>Hadoop Cloudera</td>
</tr>
<tr>
<td>10.</td>
<td>C++</td>
<td>Deep Learning</td>
<td>Apache Kafka</td>
<td>OpenShift</td>
</tr>
</tbody>
</table>

*ETF* Working together Learning for life
The retail industry, which to a large extent represents the consumers of AR/VR, has shown increasingly fast growth in demand after 2015.
Identifying & Prioritizing Disruptive Jobs & Skills

A framework for identifying and planning ahead for the skills that are likely to challenge the market into the future.

- **Escalators**
  - Low cost to hire
  - Moderate need for new training programs
  - Moderate risk to future productivity

- **Disruptors**
  - High cost to hire
  - Strong need for new training programs
  - High risk to future productivity

- **Stabilizers**
  - Low cost to hire
  - Weak need for new training programs
  - Low risk to future productivity

- **Challengers**
  - High cost to hire
  - Moderate need for new training programs
  - Moderate risk to future productivity
Even in Emerging Fields like Data Science
Not All Skills Are Created Equal
Connecting multiple sources
Structured data vs Big data

**Structured data**

Purposefully collected and collated data which comes in neat, tidy structure. In Emsi’s case, this is typically data from government statistical surveys, designed to ask explicit questions of targeted samples of specific audiences.

**Big data**

Extremely large scale data captured from some transactions rather than as a specific data collection exercise. In Emsi’s case, this means harvesting job postings and worker profiles from different web-based sources.
Aggregate data vs Micro data

Aggregate data

Aggregate data comes to us in the form of summary statistics: the total number of X, the average level of Y, the percentage doing Z. A lot of structured data is available only in aggregate form, with limited access to the micro data.

Micro data

Micro data are more powerful, with every individual case’s data points available without aggregation; we can then cut it any way we like. Big data comes to us in micro data form, although we often prepare it for standard forms of aggregation.
Connecting multiple sources

- National Records
- Labour Force Survey (LFS)
- Business Register & Employment Survey (BRES)
- Census
- Office for National Statistics
- Workforce Job series (WJS)
- Annual Population Survey (APS)
- Mid-Year Population Estimates
- National Statistics
- Jobcentre data
- Surveys
- Other datasets
- Annual Survey of Hours & Earnings (ASHE)
- O*Net ISCO08 SOC
- National Records
- National Statistics
- Surveys
- Other datasets
Job Posting & Profile Analytics – document based LMI

• **JP & PA are Document-based LMI:** it’s a Big Data source where we have the entire process in our control, including the raw data

• **JPA is (near) real-time:** we scrape job postings every day and within a day or two we can use the intelligence they produce

• **JPA is rich:** we have millions job ads (growing every day) from the last 2 years

• **JPA is noisy:** job postings can move differently to underlying labour market demand and so should be handled in that knowledge
Live session
E.g. Occupations mapped to course area Accounting (JACS N400) in the City of Bristol, Bath and North East Somerset, North Somerset and South Gloucestershire April 2021 - May 2022
Multinational Skills Insights from Lightcast Global

E.g. Analysts & Data Scientists with Python programming skills

Identify concentrations of skills supply and demand across the world within occupations, and dig down into the data to glean insights from top skills for a particular population, top employers and population estimates.
Audit skills data from course documentation from Lightcast Skill Sync

E.g. MSc Advanced Mechanical Engineering, QMUL, & BA Games Design & Art, Soton