The role of education and training offer data in skills intelligence

Presenter: Alice Bertoletti | European Commission – JRC

Coauthors: Cosgrove, J., López Cobo, M., European Commission – JRC

Cedefop Research Conference · Harnessing Web Data for Next-Generation Skills Intelligence

28 May 2026 · Thessaloniki

European Commission

Motivation and Background

- Technological change **continuously reshapes the skill content of occupations** - labour demand seems to adjust faster than curricula (Acemoglu & Restrepo 2018; Deming & Kahn 2018)
- **Higher education** is the primary channel for acquiring skilled-employment competencies, yet universities adapt slowly (Acemoglu & Autor 2011; Goldin & Katz 2008)
- Employers report persistent difficulty filling vacancies - **evidence of structural, not cyclical, misalignment** (McGuinness et al. 2018)
- Mismatch **measures available** in the literature (e.g., self-reports, vacancy-to-employment ratios) are **constructed ex post**, after sorting and occupational selection have operated. Curriculum-level sources of misalignment remain unidentified (Pellizzari & Fichen 2017; Flisi et al. 2017)

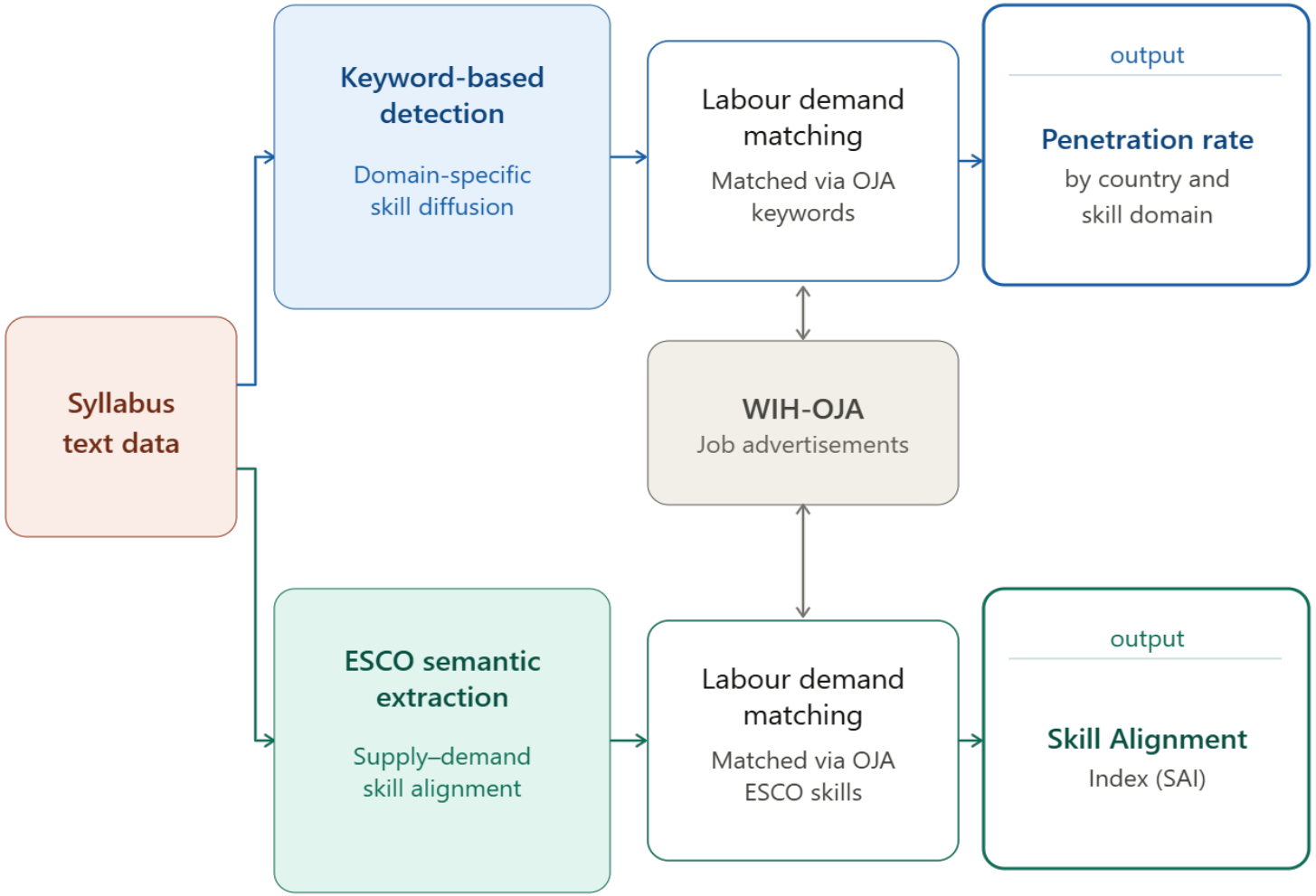
Gap in Europe: no large-scale, systematic evidence how higher education is responding to the skills employers' demand

Motivation and Background

JRC contribution

Curriculum analytics	>	EU-wide **text-based analysis of skills curricula**.
Supply-demand linkage	>	**Comparison** between academic skills and labour-market demand.
Complementary methodologies	>	**Keyword-based** detection **& Semantic ESCO** skill extraction.
Policy diagnostics	>	**Indicators** to support policy and curriculum alignment.
Domain applications	>	**Use cases** across skills and technology domains: AI, Virtual Reality, Digital and Green skills.

JRC work: Two complementary approaches



European Commission

Data sources: Academic offer text

- Both approaches **exploit large-scale web data:** curriculum text from university websites.

Keyword approach

StudyPortals

Unit: programme *(title · summary · long description · course list)*

224 K	programmes / year
World (EU-27)	coverage
2020–25	time period
100%	English syllabi
Mean 198	total words / programme
ISCED level	master, bachelor and short courses
~ 1,300	Universities in EU

ESCO extraction

OpenSyllabus

Unit: course *(title · description · topic outline · learning outcomes)*

360 K	Courses/year
World (EU27 + UK)	coverage
2018–22	time period
90%	English syllabi
Mean 208	total words / courses
ISCED level	master, bachelor
1,104	Universities in EU

Keyword-Based Approach

- **Objective:** Measure diffusion of a **specific emerging technology**
- **Best for:** Emerging tech with **well-defined vocabulary**
- **Data input:** Programme titles + short/long descriptions (**StudyPortals**)
- **Taxonomy:** **Semi-automated keyword** lists from literature (Samoili et al. 2021)
- **Matching:** Exact/**phrase matching**

Table 2. Most relevant keywords within each AI domain

AI domain	AI subdomain	Keyword	
Reasoning	Knowledge representation; Automated reasoning; Common sense reasoning	case-based reasoning	inductive programming
		causal inference	information theory
		causal models	knowledge representation & reasoning
		common-sense reasoning	latent variable models
		expert system	semantic web
		fuzzy logic	uncertainty in artificial intelligence
		graphical models	
Planning	Planning and Scheduling; Searching; Optimisation	bayesian optimisation	hierarchical task network
		constraint satisfaction	metaheuristic optimisation
		evolutionary algorithm	planning graph
		genetic algorithm	stochastic optimisation
		gradient descent	

Source (Samoili et al. 2021)

Potential Improvement:
NLP assisted keyword generation to reduce manual updating and extend coverage

Keyword-Based Approach: Workflow

1 Scope

Skills Domain · educational level·

2 Taxonomy

Identification of reference taxonomy: literature

3 Keyword

Inclusion of keywords in the query builder

4 Pre-process

Language filter · lower-case · harmonisation to improve matching accuracy

5 Matching

Identification of programmes containing the keyword

6 Validate

Manual spot-check

7 Classify

Specialised vs. broad (based on the intensity of their focus)

8 Indicators

Penetration rate · subdomain shares · field/country breakdown

Keyword-Based Approach: Penetration Rate

$$PR_{d,c,l} = {}^{N_{relevant}}/_{N_{total}}$$

Notation

N relevant	Count of programmes containing at least one domain keyword; weighted by keyword density when a programme matches multiple subdomains
N total	Total number of programmes in the reference population
d	Technology domain (e.g. AI, Virtual World)
c	Country
l	Educational level (Bachelor, Master, short cycle…)

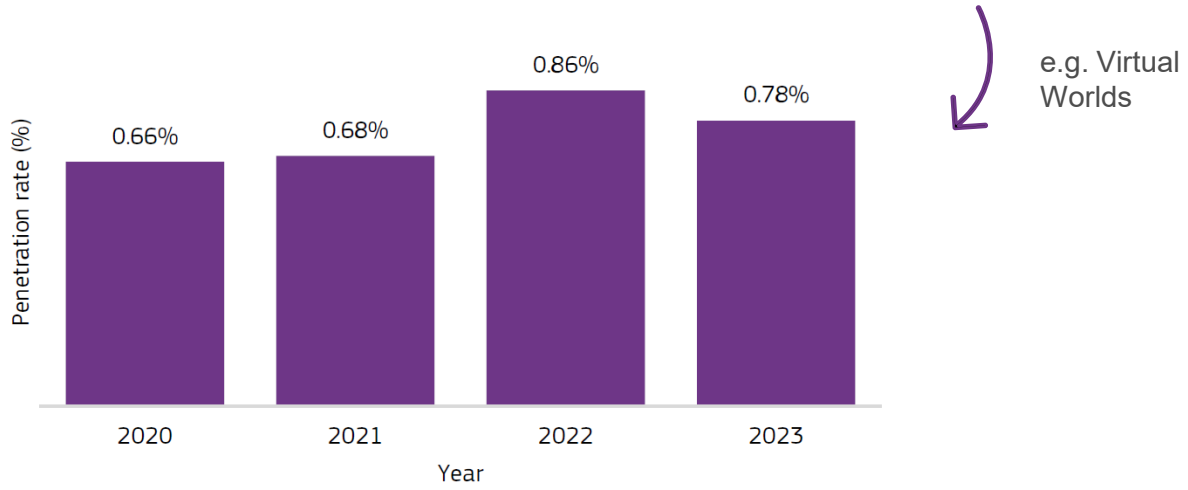
DECOMPOSITION DIMENSIONS

By country
Cross-country heterogeneity in curriculum exposure; benchmarks national HE systems against EU average

By field of study
Identifies how deeply has AI penetrated a specific ISCED-F

By technology domain
Computes PR per domain (AI, VR…) and subdomain to map diffusion depth across the curriculum

Across year
Tracks longitudinal changes in PR to capture speed of curriculum adaptation to labour demand



Source (Herrero et al., 2026)

European Commission

Keyword-Based Approach: Indicators by field of study

PENETRATION RATE

$$PR_f = N^w_{rel,f} / N^w_{total,f}$$

Share of AI-relevant programmes *within* field f.

> RQ: *"How deeply has AI penetrated a specific field of study?"*

CONCENTRATION - SHARE

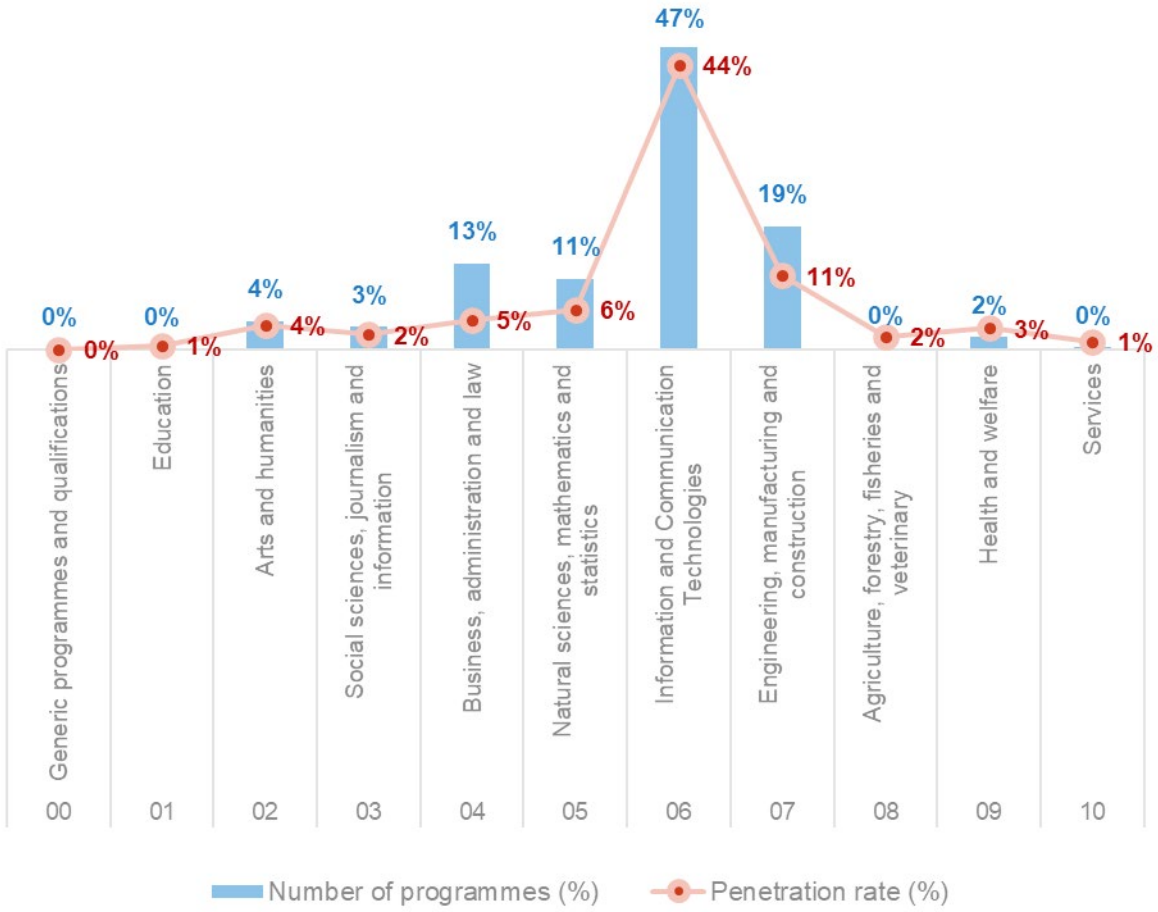
$$S_f = N^w_{rel,f} / \Sigma_f N^w_{rel,f}$$

Distribution of AI programmes *across* ISCED fields.

> RQ: *"Where are AI courses concentrated?"*

By field of study

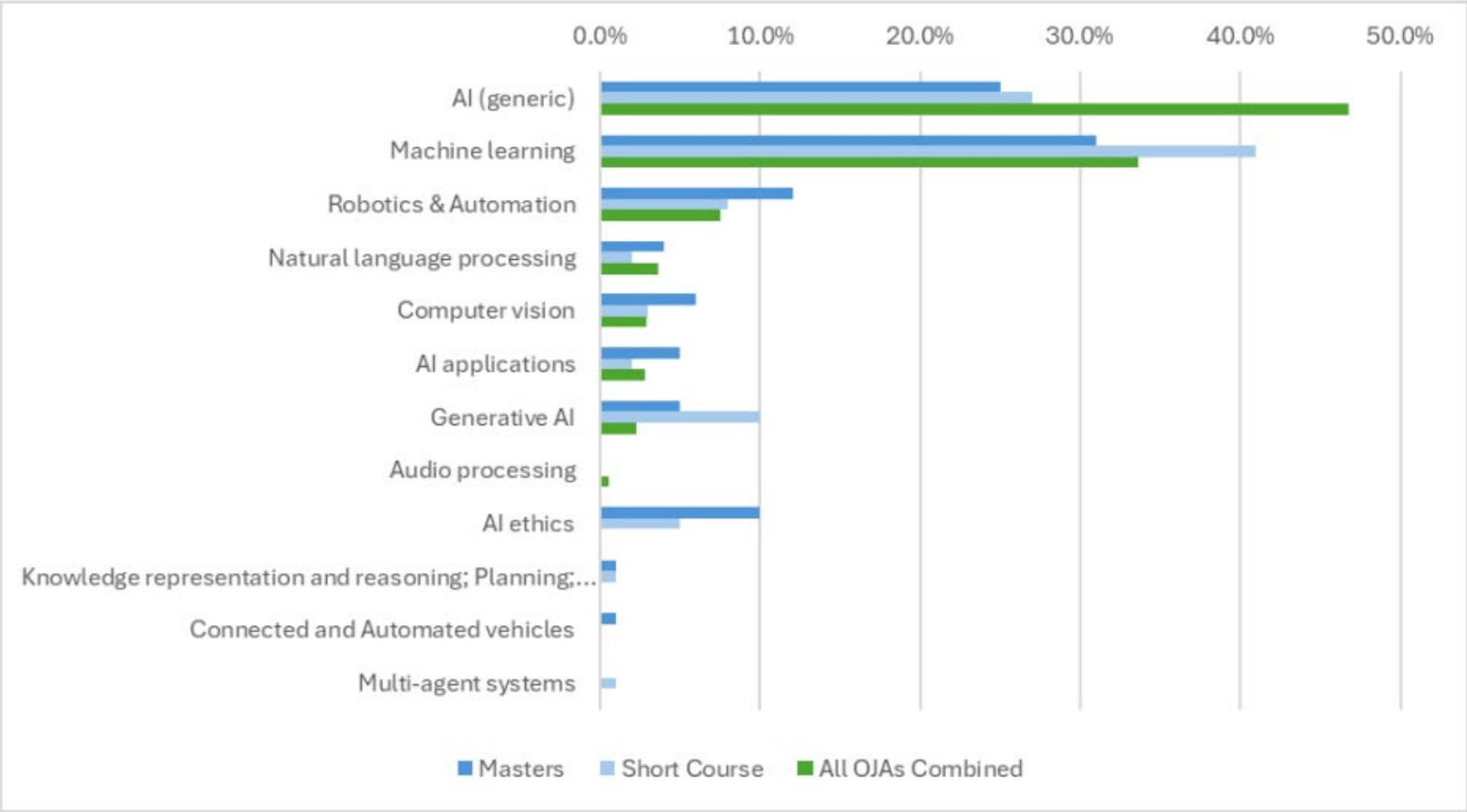
EU 2024-25



(Bertoletti, Cosgrove & López-Cobo, 2025)

Keyword-Based Approach: Combined analysis with OJA

Distribution of AI-related education and training offer (2024-2025) and AI-related ICT specialist OJAs (2020-23) by AI subdomain (%), EU



(Bertoletti, Cosgrove, López-Cobo, 2025)

- **Common keyword** taxonomy applied to both supply and demand
- **OJA text** → Analysis was carried out with the support of Cedefop
- The keywords are clustered using a taxonomy of **AI subdomains (**Samoili et al. 2021)
- ⚠️ **Temporal gap:** Supply data 2024–25 vs. demand data 2020–23.

European Commission

ESCO semantic Skill Extraction

- **Objective:** Broad structural comparison of **ALL skill domains** across **supply and demand**
- **Best for:** **Systematic cross-skill**, geographical benchmarking
- **Data input:** Course syllabi full text (**OpenSyllabus**) + OJA skill records (**WIH**)
- **Taxonomy:** Full **ESCO v1.2** → 750 semantic clusters via k-means
- **Matching:** Dense retrieval and **cosine similarity**

Skill Alignment Index (SAI) as pre-market measure:

SAI captures structural curriculum–demand fit before graduate allocation operates, identifying misalignment that post-market measures conflate with occupational mobility and selection.

ESCO Semantic Skill Extraction: Workflow

1 Embedding

Syllabus texts and ESCO skill definitions encoded into a shared **semantic vector space** using the (Qwen3-Embedding-0.6B bi-encoder)

2 Skill assignment

Cosine similarity computed for every syllabus × ESCO-skill pair. Top-ranked candidates retained if they pass the **calibrated threshold** (μ + 1σ)

3 Extraction and Validation

Retained output is a set of **syllabus pairs with scores and ESCO** IDs. The results are manually **validated** with feedback to improve calibration

ESCO skills: NO transversal skills, NO teaching skills

4 WIH - OJA data

- 2018-22 aggregated OJA
- **Only HE** OJA
- **ESCO skills** already available
- **No transversal** skills (26%)
- **928.7 million skill** mentions from 97.5million OJAs

5 Clustering

13,939 ESCO skills reduced to **750 semantic clusters** via k-means. Interpretable labels generated by an LLM (Claude 4.5 Sonnet). Label quality validated through expert review.

6 SAI construction

TF-IGF vectors for supply (syllabi) and demand (OJAs) per country/region. **Cosine similarity** normalised to the cross-regional median (= 100%)

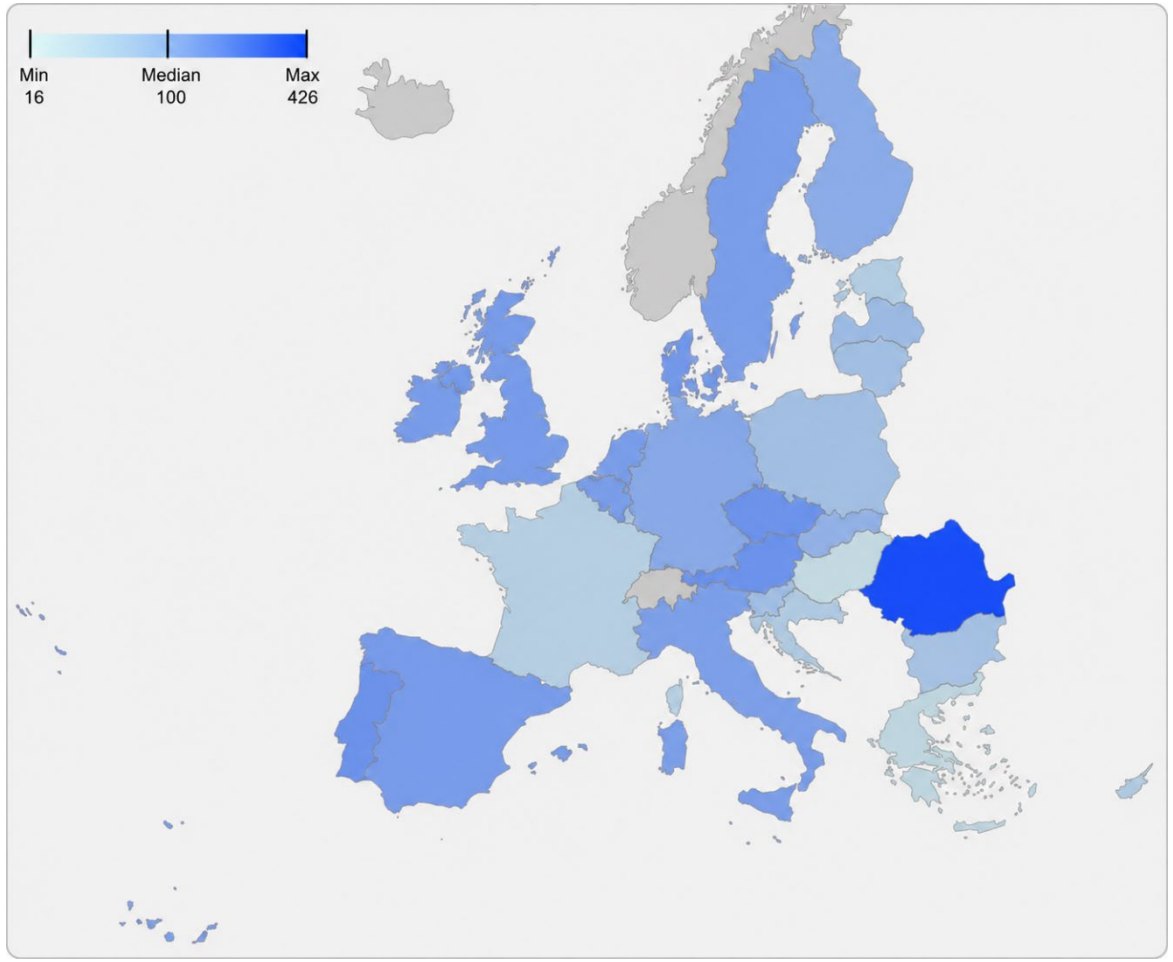
ESCO Semantic Skill Extraction: Skill Alignment Index

SAI Index Definition

$$SAI = \frac{\rho(SCP^{TF-IGF}, SCV^{TF-IGF})}{P_{50}(\rho)} \times 100\%$$

- **SAI = 150%:** Alignment is 50% higher than the median country
- **SAI = 40%:** Alignment is 60% lower than the median country

- **ρ:** *Cosine Similarity*
- **SCP:** *Skill Cluster vector for teaching Programs*
- **SCV:** *Skill Cluster vector for Vacancies*
- **TF-IGF:** *frequency-inverse geographic frequency transformation, weighting by their relative importance within each country and their geographic specificity*

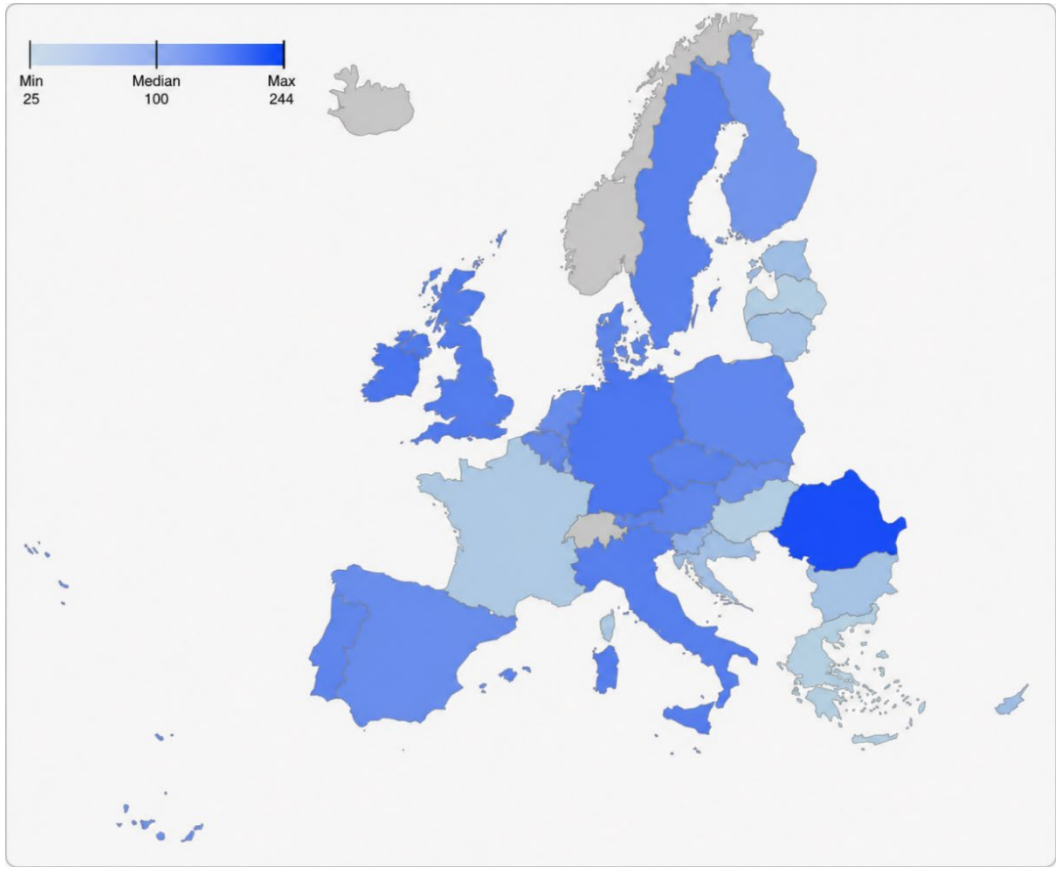


(Bertoletti et al., forthcoming 2026)

European Commission

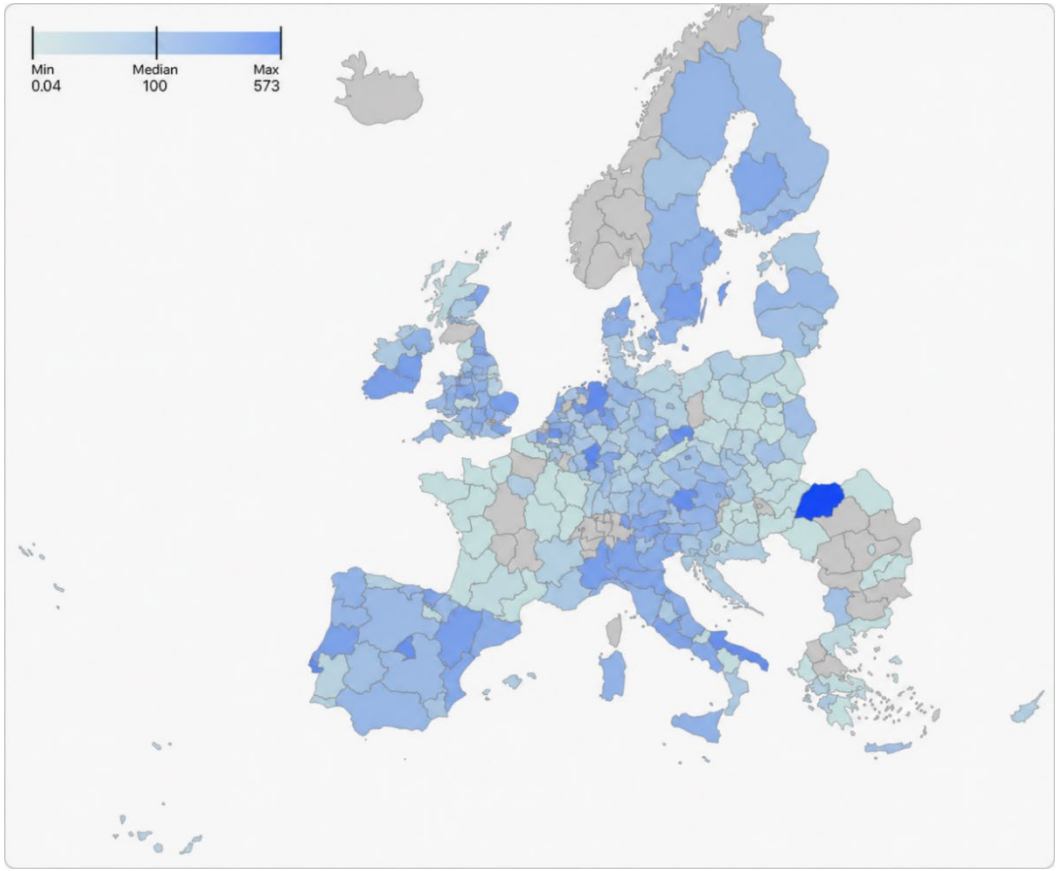
ESCO Semantic Skill Extraction: Skill Alignment Index

D-SAI : Specialised SAI for Digital skill clusters



(Bertoletti et al., forthcoming 2026)

R-SAI: Specialised SAI at regional (NUTS-2) level



(Bertoletti et al., forthcoming 2026)

European Commission

Two Complementary Approaches to Skills Analysis

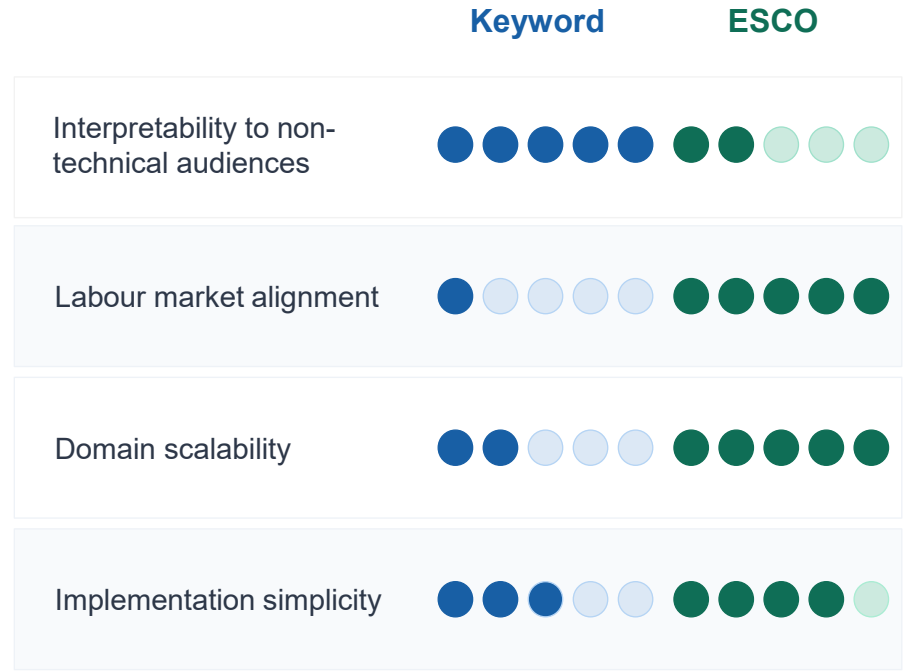
"What emerging technology skills are present in the academic offer, and how specialised is the focus? →**Keyword Based**

- Measures the penetration and **diffusion of a specific technology domain across curricula.**
- Transparent taxonomy, intuitive indicator, **directly communicable** to non-technical policy audiences.
- Keyword scalability is limited by the need to build a **separate taxonomy per domain**

"How well does the academic offer structurally match labour market skill demand?"→ **ESCO Semantic skills extraction**

- **Direct indicator of Alignment:** full academic offer against the ESCO taxonomy and benchmarks it against labour demand from OJA.
- **Scalable across skill clusters** and subdomains; diagnostic at regional level.

DIMENSION SCORECARD



● = high ○ = low *(qualitative, relative)*

European Commission

Limitations and contributions

LIMITATIONS STILL TO ADDRESS

Syllabi coverage and heterogeneity

Student coverage varies from 20% to 100% across EU countries (OpenSyllabus) and cross-country comparisons require caution where coverage is low (France, Bulgaria, Slovenia)

OJA representativeness

OJA overrepresent white-collar, formal-sector vacancies. Possible overstatement of transversal skills and Software skills while low representation of less codified skills (Sostero, & Fernández-Macías 2021).

English-language bias in syllabus data

Syllabi predominantly in English — non-Anglophone HE systems may have their skill content under-detected, generating measurement non-equivalence

GAPS ADDRESSED

Structural vs. cyclical mismatch

Conventional shortage indicators measure unfilled vacancies that are shaped by sorting, and cyclical demand. SAI is a pre-market measure, isolating the structural component of mismatch and enables early-warning diagnostics.

Actionable tool for skill bottleneck detection

SAI decomposes into bottleneck and excess-supply skill clusters — an actionable tool for university administrators and accreditation bodies

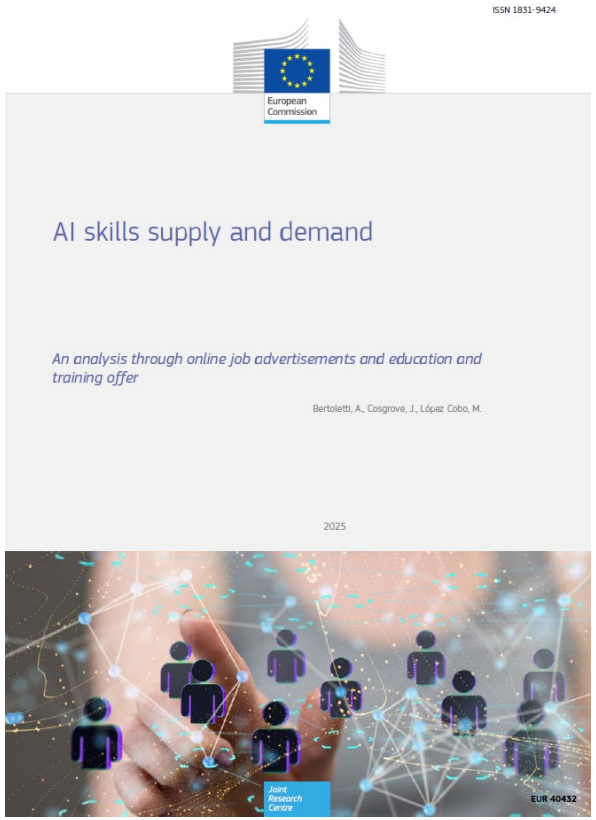
Domain-specific compositional analysis

Keyword approach provides transparent, replicable indicators when the policy question targets a specific domain — enabling detailed compositional analysis of syllabi content

Published outputs and forthcomings

Herrero, C., Bertoletti, A., Cosgrove, J., López Cobo, M. (2026)

Bertoletti, A., Cosgrove, J., López Cobo, M. (2025)



Bertoletti, A., Christmann, F., De Quinto, A., Lalanne, M., Torecillas-Jodar, J. (Forthcoming 2026)

Assessing the Alignment between University Curricula and Job Market Requirements in Europe: A New Skill-Based Index

Alice Bertoletti[*1], Federico Christmann[†2], Alicia de Quinto[‡2,3], Marie Lalanne[1], and Juan Torrecillas Jodar[§1]

[1]European Commission - Joint Research Centre
[2]Universidad Autónoma de Madrid
[3]Banco de España

May 14, 2026

Abstract

This paper develops a new measure of the alignment between university curricula and labor market skill demand in Europe. We combine 1.8 million university syllabi from Open Syllabus with skill requirements extracted from online job advertisements in the Web Intelligence Hub (WIH) for the EU-27 and the UK. Using the ESCO taxonomy, we extract skills from syllabus text with transformer-based semantic retrieval, aggregate granular skills into semantic clusters, and compare regional curriculum and vacancy skill profiles through a Skill Alignment Index (SAI). The results reveal substantial heterogeneity in curriculum–labor market alignment across countries and, especially, across NUTS 2 regions. Alignment is higher in many economically dynamic and knowledge-intensive regions. On field-specific skills, digital skills present larger bottlenecks, whereas green skill alignment varies sharply across countries.

JEL Classification: J23, J24, J63, O33.
Keywords: Online Job Advertisements, Skills, Skills Alignment, Syllabi, Text Analysis.

[*]alice.bertoletti@ec.europa.eu
[†]federico.christmann@estudiante.uam.es
[‡]alicia.quinto@bde.es
[§]juan.torrecillas-jodar@ec.europa.eu

European Commission

References

- Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics*, 4, 1043–1171.
- Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation, and work. In *The Economics of Artificial Intelligence* (pp. 197–236). University of Chicago Press.
- Bertoletti, A., Cosgrove, J., López Cobo, M. (2025) AI skills supply and demand - An analysis through online job advertisements and education and training offer, Publications Office of the European Union, Luxembourg, https://data.europa.eu/doi/10.2760/0059391, JRC143488.
- Bertoletti, A., Christmann, F., De Quinto, A., Lalanne, M., & Torrecillas Jodar, J. (forthcoming). Assessing the alignment between university curricula and job market requirements in Europe: A new skill-based index. *European Commission - Joint Research Centre*.
- Deming, D. J., & Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1), S337–S369.
- Flisi, S., et al. (2017). Measuring occupational mismatch: Overeducation and overskill in Europe — evidence from PIAAC. *Social Indicators Research*, 131(3), 1211–1249.
- Goldin, C. D., & Katz, L. F. (2008). *The Race Between Education and Technology*. Harvard University Press.
- Herrero, C., Bertoletti, A., Cosgrove, J., López Cobo, M. (2026) Observing Virtual Worlds: Multi-method Analysis of Skills, Challenges and Emerging Trends, Publications Office of the European Union, Luxembourg, https://data.europa.eu/doi/10.2760/4780686, JRC145833.
- McGuinness, S., Pouliakas, K., & Redmond, P. (2018). Skills mismatch: Concepts, measurement and policy approaches. *Journal of Economic Surveys*, 32(4), 985–1015.
- Pellizzari, M., & Fichen, A. (2017). A new measure of skill mismatch: Theory and evidence from PIAAC. *IZA Journal of Labor Economics*, 6(1).
- Samoili, S., et al. (2021). *AI Watch. Artificial Intelligence Taxonomy*. JRC Technical Report. Publications Office of the European Union.
- Sostero, M., & Fernández-Macías, E. (2021). The professional lens: What online job advertisements can say about occupational task profiles. *JRC Working Papers on Labour, Education and Technology*, No. 2021/13.

Thank you

Alice Bertoletti | alice.bertoletti@ec.europa.eu