SPECIFICATION: SERVICES

Contract Name: Research Web Scraping ECI Labour Market

Contract Reference: P2503-1

**Version 1.1**

Contents

[1. Introduction 3](#_Toc201071185)

[1.1 About the ECITB 3](#_Toc201071186)

[2. Background 3](#_Toc201071187)

[3. Scope of the Contract 4](#_Toc201071188)

[4. Requirements 4](#_Toc201071189)

[4.4 Mandatory Requirements 5](#_Toc201071190)

[Web Scraping and Data Handling Requirements 5](#_Toc201071191)

[The supplier must: 5](#_Toc201071192)

[Taxonomy Development Requirements 6](#_Toc201071193)

[5. Service Levels and Key Performance Indicators (KPIs) 6](#_Toc201071194)

[6. Contract Management and Review 7](#_Toc201071195)

[7. Insurance 8](#_Toc201071196)

[8. Environment and Sustainability 8](#_Toc201071197)

[9. Health and Safety 8](#_Toc201071198)

[10. Equality, Diversity and Inclusion 8](#_Toc201071199)

[11. Security 8](#_Toc201071200)

[*12.* General Data Protection Regulation (GDPR) 9](#_Toc201071201)

[*13.* Contract Type and Duration 9](#_Toc201071202)

# Introduction

## 1.1 About the ECITB

The Engineering Construction Industry Training Board (ECITB) is the employer-led skills, standards and qualifications body for the development of the engineering construction workforce of Great Britain. An arms-length body of the UK Government, the ECITB reports to the Department for Education. Our vision is a trailblazing Engineering Construction Industry where critical infrastructure, energy security and net zero ambitions are achieved.

The ECITB works with employers and training providers to attract, develop and qualify the engineering construction workforce in a wide range of craft, technical and professional disciplines. We invest about £28 million each year to support skills development within the industry.

More information about the ECITB can be found here: [**What is the ECITB? - ECITB**](https://www.ecitb.org.uk/)

# Background

Labour market taxonomies are essential because they inform public policy, standardise job classifications, and ensure all parties (employees, employers, and policymakers) have equal access to information. They help job seekers who want to change industries by informing them about the opportunities and requirements of vacancies or upskilling their training skills. They also guide training providers in aligning curricula with industry needs (Ospino Hernandez, 2018).

However, in niche sectors such as the Engineering Construction Industry (ECI), the Standard Occupational Classification (SOC) taxonomy lacks the detail necessary to capture the emerging skills and occupations within the industry. Additionally, the available workforce data based on the SOC and collected from traditional sources such as surveys, administrative data and data from industry bodies has limitations[[1]](#footnote-1) and lacks the granularity that employers and policymakers require to respond effectively.

As many organisations have pointed out, such as the Warwick Institute for Employment Research (IER), the World Economic Forum (WEF), and LMI For All, a skill taxonomy has a myriad of benefits in terms of labour market intelligence for policymakers, education providers, and employers. Integrating web scrapping methodologies in developing a taxonomy can overcome many of the limitations of traditional data sources. In this approach, online job advertisements can be scraped from the web using artificial intelligence; then, the data is cleaned and analysed by algorithms. Its key advantage is that the data can be provided in real-time, including skill requirements, and, if available, the data can be focussed at the regional level.

By developing this research, the ECITB will complement its current labour market sources, such as the LFT and Census, with skills intelligence data that will allow it to identify skills gaps, complement its forecasts on employment trends regarding skills (green skills, digital skills, etc.), and support other products like Connected Competence.

# Scope of the Contract

This project aims to develop comprehensive labour market intelligence for the ECI by using web scraping techniques to analyse job advertisements from online platforms such as Indeed, Glassdoor, LinkedIn, and others. This project seeks to:

* Understand the evolving skills landscape within the ECI sector.
* Map the geographical and sectoral distribution of skills demand.
* Provide evidence-based insights to support workforce development strategies.

The following objectives will guide the research:

1. Identify which vacancy platforms ECI employers use.
2. Develop and implement web scraping algorithms to extract job posting data from Indeed, Glassdoor, LinkedIn and other platforms.
3. Create and enhance an occupational and skills taxonomy for the ECI built upon SOC codes with additional ECI-specific classifications.
4. Analyse the extracted data to identify skills gaps, emerging skills requirements and high-demand occupations across regions and sectors.
5. Quantify occupational shortages and estimate occupational demands across ECI sectors and regions.
6. Produce periodic reports and dashboards to inform ECITB’s stakeholders.
7. Look at the educational pathways of the five most in-demand vacancies.

#

# Requirements

**Performance Requirements**

**Expected outcomes**

The expected outcomes for this project are the following:

1. A catalogue of the industry’s needs regarding skills by sector.
2. A list of high-demand occupations by region and sector, identifying trends and shortages in the labour market.
3. A customised skill taxonomy for the ECI.
4. Policy recommendations for workforce planning, reskilling, and qualification development.
5. Internal recommendations relevant to course development and grant spend.
6. Document containing the methodology of the taxonomy and a manual to explain how to operate, update, and interpret outputs from the scrapping and classification algorithms.

**Overall measures**

To ensure quality and accountability, the following performance standards must be met:

* The algorithm should be designed to scale during periods of high data volume (e.g., season spikes).
* The accuracy in classifying job roles and skills must be higher than 90%, based on random sample testing.
* Duplicate detection and removal should be above 90%.
* The taxonomy must be built using SOC 2020 as a base, with ECI-specific sub-codes as classification.
* NLP models must be trained using UK-based job descriptions and updated quarterly.
* The solution must comply with UK GDPR.
* Open source or interoperable formats should be used to ensure compatibility with ECITB current systems.

## 4.4 Mandatory Requirements

**Supplier profile**

The ideal supplier should be a data and technology consultancy or research organisation with proven experience in:

* Web scraping on LinkedIn, Glassdoor and other large scale job platforms.
* Natural Language Processing (NLP) for job description analysis.
* Analysing Labour Market data and identifying skills trends.
* Taxonomy development aligned with SOC frameworks.
* Skills, intelligence, or workforce planning.
* Familiarity with UK data protection and ethical standards.

Additionally, the supplier must demonstrate a clear understanding of the potential risks involved in the delivery of this project by indicating how they will identify, monitor, and mitigate the following categories of risks throughout the contract lifecycle:

* Data reliability
	+ As previous ECITB surveys have demonstrated, some employers within the industry do not post vacancies online, which can lead to an underrepresentation of regions, sectors and occupations.
	+ Not all industry roles may be accurately represented, as online websites tend to be skewed to managerial and technical jobs, excluding low-paid jobs.
	+ Potential biases towards specific industries.
	+ Vacancies may be duplicated, which can be solved during the cleaning process through a factor analysis.
* Data classification
	+ Transversal knowledge, skills, and competencies are applicable to a wide variety of occupations and sectors.
	+ Need to validate the accuracy of the taxonomy.
* Data privacy and compliance
	+ Web scraping is subject to legal and ethical constraints, which can limit the scope of the websites available for the research.
* Computational requirements
	+ Computational power
	+ Storage needs
	+ Website changes

### Web Scraping and Data Handling Requirements

### The supplier must:

* Implemented automated data collection from job platforms such as Indeed, Glassdoor, and LinkedIn platforms.
* Apply Natural Language Processing (NLP) to extract and categorise skill-related keywords from job descriptions.
* Use machine Learning algorithms to cluster job roles and identify patterns and trends in skill demand.

The methodology must include:

1. The most frequently used job platforms by ECI employers.
2. A system for regularly collecting job posts, capturing at minimum the following fields and with the option to extend the fields as needed:
	1. Company posting the vacancy
	2. Name of the recruiter
	3. Sector
	4. Type of job
	5. Location
	6. Pay rate
	7. Skill
	8. Qualification/s
	9. Project (if specified)
	10. Start date
3. A quality control system to prevent duplication and track where and when posts were found.
4. A process to homogenise job titles.
5. A method to differentiate between general job skills and technical or coding skills.
6. The NLP and classification model must be developed using open-source libraries and be fully accessible and modifiable by ECITB after delivery.
7. Final code and models must be delivered in a reusable format, with approvable documentation and version control (GitHub or other to be agreed).

### Taxonomy Development Requirements

* Use international best practices to add sub-codes to the SOC classification, enabling more granular skill and role descriptions aligned to ECI needs.
* Ensure the taxonomy integrates with UK labour market data frameworks.
* Review the skills and occupation groups with industry experts.
* Get feedback from people who work in the roles.

**Technical Documentation and Delivery Requirements**

* The supplier must deliver a clear, step-by-step user manual or guide explaining how to operate, update, and interpret outputs from the scraping and classification algorithms.
* The manual should be written for both technical and non-technical users and include instructions for troubleshooting common issues.
* The supplier must ensure that, by the end of the project, the ECITB has full access to the data, code, user guide, and necessary documentation to operate the system independently.

The ECITB will retain full ownership of any research outcome. Any changes or increases in contract scope will be agreed upon by both parties in writing prior to action in accordance with the agreed contractual terms.

# Service Levels and Key Performance Indicators (KPIs)

**Service Levels**

* All progress and review meetings regarding the project will be held online, unless otherwise agreed in writing.
* The supplier must provide a regular progress update (biweekly), including a short demonstration or walk-through where requested.

**Key Performance Indicators (KPIs)**

|  |  |
| --- | --- |
| KPI | Target |
| Job post collection volume |  |
| Classification accuracy | At least 90% for role and skill classification |
| Duplicate detection rate | Over 95% accuracy |
| Taxonomy validation | Industry experts review to confirm accuracy and relevance |
| Dashboard updates | Quarterly updates |

**Milestones and reporting**

The performance will be tracked against the following project milestones:

|  |  |
| --- | --- |
| Month | Deliverable |
| September | Scraping system prototype and NLP classification model |
| October | Data quality assurance system and initial collection begins |
| November | First working version of the skill taxonomy and the dashboard |
| December | Expert review of taxonomy and ECITB feedback |
| January | Final refinements to the dashboard and taxonomy |
| February | File reporting and delivery of all outputs |

#

# Contract Management and Review

**Meetings and Progress Review**

* An initial meeting will be held at the start of the contract to review timelines, roles, deliverables, and other topics relevant to the project.
* Biweekly check-in meetings will be held online during the development and implementation.
* Throughout the contract, extra meetings could be held with ECITB stakeholders and industry experts to monitor progress, discuss the classification of skills or job roles, or discuss any other issues.
* All meetings will be held remotely.

**Reporting Requirements**

* A report of the methodology used to classify the jobs and skills into the taxonomy for the ECI after this stage is completed. This should be written in a clear, concise, non-technical manner to support understanding among industry experts who will be involved in reviewing and validating the taxonomy.
* A final project report, including methodology, results, documentation, and recommendations.

# Insurance

The Supplier must hold the following insurances at the limitation values stated from commencement of the agreement and for the duration of agreement thereafter:

* Employer’s Liability Insurance - £5m
* Public Liability Insurance - £1m
* Public Indemnity Insurance - £1m

# Environment and Sustainability

The Supplier should adhere to any relevant ECITB Environmental and Sustainability policies as required. Copies can be provided on request.

# Health and Safety

As a minimum, suppliers should have a written Health and Safety Policy that complies with current legislative requirements (if applicable, i.e. your organisation has less than 5 employees this does not need to be provided).

Suppliers should not have been prosecuted for contravention of the Health & Safety at Work Act 1974 or equivalent national legislation or been the subject of a formal investigation by the Health and Safety Executive or similar body charged with improving health and safety standards.

# Equality, Diversity and Inclusion

As a minimum, the Supplier should have in place current Equality, Diversity and Inclusion Policy/ies (ED&I) that comply with current legislative requirements.

# Security

As a minimum, the Supplier should hold a current Cyber Essential certificate and ideally, Cyber Essentials plus (or equivalent in both cases). If equivalent cyber certification is held, we may request further details on this.

# General Data Protection Regulation (GDPR)

The supplier might receive a list of companies' names in scope to ensure that the vacancies are within the Engineering Construction and related industries, as the project needs accurate data on the skills and certifications in demand within the industry.

# Contract Type and Duration

The contract will be for an initial and total term of **6 months** with an optional extension of 3 months. Deliverables for the project including milestones and key timelines are as detailed within this specification.

The Supplier must agree to contract under the ECITB Service Agreement shared as part of the Procurement Pack.

1. Some of the limitations of these data sources are: frequency of data, sample size, collection, limited focus on skills and lack of comparability (LMI for all, 2025). [↑](#footnote-ref-1)