

mInimize coSts and mAximize Benefits in the transitionary European regional Labour markets

Deliverable No 4.5
Title: Initial ISABEL web crawlers

WP4: JCD and skills analysis to manage effectively and fairly the JCD process

Funded by the European Union

DOCUMENT ID	D4.5
DUE DATE	30/06/2026
DELIVERY DATE	30/06/2026
WORK PACKAGE	WP4
TASK(S)	T4.4
DISSEMINATION LEVEL	PU
VERSION	V2
AUTHOR/S	CERTH, IER
FILENAME	ISABEL_D4.5_web_crawlers_V2.docx

Abstract
This deliverable reports the first operational version of the web crawler suite developed under ISABEL WP 4. Five crawler modules are described: patent data, online job postings and skills, academic and grey literature publications, official European statistics, and policy document semantic scanning. For each module the document covers functional design, technical specifications, data source coverage, and maturity status. Two modules are fully operational (UK job-posting scrapers and Greek policy semantic scanner); the remaining three are at design or early-implementation stage. The document also describes integration with the ISABEL Medallion data lake (Bronze/Silver/Gold tiers on CERTH on-premises MinIO storage), the quality-assurance framework, and deployment via GitLab CI/CD pipelines. Planned enhancements include extension to all six ISABEL target countries and a multilingual embedding model for the policy scanner.

Views and opinions expressed in this document are those of the author(s) and do not necessarily reflect those of the European Union or REA. Neither the European Commission nor the granting authority can be held responsible for them.

Contributors

NAME	ORGANISATION
Michalis Angelou	CERTH
Nicholas Vretos	CERTH
Symeon Samaras	CERTH
Maria kaklidou	CERTH
Vasileios Rekkas	CERTH
Sangwoo Lee	IER
Rosie Day	IER
Luke Bosworth	IER

Peer Reviewers

NAME	ORGANISATION
Fotini Vlachaki	PROG
Lukasz Trembaczowski	US

Versions

Version	Date	Author / Organisation	Description
ToC	15/05/2026	CERTH	Table of Contents
V1	15/06/2026	CERTH	Initial draft
V2	26/06/2026	CERTH	Integrated review comments

Funded by
the European Union

Executive summary

This deliverable reports the first version of the ISABEL web crawlers, developed under WP4. The suite provides an automated data ingestion layer that feeds the ISABEL Medallion data lake, supplying the platform with high-frequency, heterogeneous signals from five complementary source categories: patent databases, online job postings, academic and grey literature, official European statistics, and policy document archives.

In this first version, two modules are fully operational: the UK job-posting scrapers (Guardian Jobs and Reed.co.uk) and the Greek green-transition policy semantic scanner. Three modules — patent data, academic publications, and official statistics — are at design or early-implementation stage and will be brought to production in the next release cycle. The operational scrapers are pending full integration with the platform Bronze layer, which is the primary tracked task for version 2.

The document covers the shared architectural principles governing all crawlers, the technical specifications and maturity status of each module, the Bronze-layer data schema, quality-assurance procedures, and the deployment approach on the CERTH on-premises infrastructure. Planned enhancements include extending job-posting coverage to all six ISABEL target countries (Denmark, Greece, Poland, Spain, France, and the United Kingdom), completing the ESCO-based skills standardisation pipeline, adopting a multilingual sentence-transformer model for the policy scanner, and full Bronze-layer integration.

Funded by
the European Union

Table of Contents

Executive summary 4
1. Introduction 7
2. Overview of Web Crawlers 8
2.1 Architecture and Design Principles 8
2.2 Data Sources and Coverage 8
3. Patent Data Crawler 9
3.1 Functionality 9
3.1.1 Data Extraction 9
3.1.2 Metadata Processing 9
3.2 Technical Specifications 9
3.2.1 Data Sources and APIs 9
3.2.2 Performance and Scalability 9
4. UK Platforms Data Crawler 10
4.1 Functionality 10
4.1.1 Job Posting Extraction 10
4.1.2 Skills Standardisation 10
4.2 Technical Specifications 10
4.2.1 Data Collection Infrastructure 10
4.2.2 NLP Processing Pipeline 11
5. Publications Crawler 12
5.1 Functionality 12
5.1.1 Academic Database Harvesting 12
5.1.2 Topic and Trend Analysis 12
5.2 Technical Specifications 12
5.2.1 API Integration and Harvesting 12
6. Official Statistics Crawler 13
6.1 Functionality 13
6.1.1 Eurostat Data Retrieval 13
6.1.2 Complementary European Data Authorities 13
6.2 Technical Specifications 13
6.2.1 API Integration and Authentication 13
6.2.2 Data Format Normalisation 14
6.2.3 Scheduling and Incremental Updates 14
7. Policy Document Semantic Scanner 15
7.1 Functionality 15
7.1.1 Semantic Text Extraction 15
7.1.2 Relevance Scoring and Filtering 15
7.2 Technical Specifications 15
7.2.1 Model and Infrastructure 15
7.2.2 Target Topics Configuration 15
8. Data Integration and Storage 16
8.1 Platform Data Lake Integration 16
8.2 Data Deduplication and Cleaning 16
8.3 Bronze-Layer Canonical Record Schema 16
9. Quality Assurance and Validation 17
9.1 Data Quality Metrics 17
9.2 Validation Procedures 17

Funded by the European Union

10. Deployment and Configuration 18
10.1 Installation and Setup 18
10.2 Scheduling and Monitoring 18
11. Future Enhancements 19
11.1 Planned Capabilities 19
11.2 Technology Improvements 19
Bibliography 20

List of tables

Table 1: List of crawlers 8
Table 2: Bronze-layer canonical record schema 16

Funded by
the European Union

1. Introduction

This deliverable reports the first operational version of the web crawlers developed as part of ISABEL WP4 – "JCD and skills analysis to manage effectively and fairy the JCD process". The ISABEL platform is designed to provide European regional policymakers with timely, granular intelligence on Job Creation and Destruction (JCD) dynamics and emerging skills shortages across six target countries: Denmark, Greece, Poland, Spain, France, and the United Kingdom. Achieving this objective requires the collection of heterogeneous, high-frequency data from a wide range of digital sources that supplement, and in many respects surpass, the timeliness and topical specificity of traditional official statistics.

Web crawlers are automated software agents that systematically retrieve and extract content from online sources according to pre-defined rules. Within ISABEL they contribute to the data ingestion layer feeding the Bronze tier of the platform's Medallion data lake, operating upstream of the Docling-based document processing and DuckDB analytics layer. The crawlers documented here collect data from patent databases, online job boards, academic and grey literature repositories, European statistical authorities, and policy document archives.

It is important to note the first-version scope. Of the five crawler modules documented here, two are fully operational: the UK job-posting scrapers and the Greek policy semantic scanner. The remaining three (patent data, publications, and official statistics) are at design or early-implementation stage. Extending job-posting and policy-scanner coverage to all six target countries is the primary capability objective for the next version. The operational scrapers also write to local and network destinations at present; alignment with the shared Bronze-layer ingestion interface is a tracked near-term integration task, as discussed in Sections 8 and 11.

This document is structured as follows. Section 2 provides a cross-cutting overview of the architecture and design principles common to all crawlers and their relationship to the wider ISABEL platform. Sections 3 to 7 describe each crawler module in detail. Section 8 covers integration with the platform data lake and storage layer. Section 9 addresses quality assurance and validation. Section 10 describes deployment and operational configuration. Section 11 outlines planned future enhancements, and Section 12 concludes.

Funded by
the European Union

2. Overview of Web Crawlers

2.1 Architecture and Design Principles

The ISABEL web crawler suite is built on a modular architecture in which each data source is addressed by an independent, self-contained crawler component. This design enables parallel development and deployment, limits failure blast radius, and facilitates the addition of new data sources without architectural refactoring. The crawlers run as scheduled GitLab CI/CD jobs on the CERTH on-premises node, consistent with the platform architecture. All crawlers share a common set of design principles:

- Responsible crawling: request rates are throttled, robots.txt directives are respected, and crawlers identify themselves via appropriate User-Agent strings.
- Fault tolerance: all crawlers implement retry logic with exponential backoff, checkpoint-based resumption, and graceful error handling to ensure robustness over long-running collection sessions.
- Reproducibility: output is date- and time-stamped, and raw payloads are preserved in the Bronze layer alongside cleaned extracts to enable reprocessing.
- Configurability: all operational parameters (rate limits, pagination depth, target endpoints, output paths) will externalised as configuration, avoiding hard-coded values in processing logic, which is now the case, in the current state of the development.
- Containerisation: all crawlers are packaged as Docker images, version-controlled in the project GitLab, and built and deployed through the platform GitLab CI/CD pipeline.

2.2 Data Sources and Coverage

The initial crawler suite covers five principal data source categories: (1) patent databases, providing innovation trend signals; (2) online job postings, providing real-time skill demand data; (3) academic and grey literature, providing forward-looking research signals; (4) official European statistics, providing authoritative quantitative benchmarks; and (5) policy document corpora, providing thematically relevant regulatory and strategic context.

The current implementation prioritises depth over breadth. Table 1 provides a consolidated status overview.

Table 1: List of crawlers

Crawler	Data Sources	Bronze Integration	Notes
Patent Data	USPTO, EPO OPS, WIPO PatentScope	Planned	No working extraction; specification complete.
Skills / Job Postings	Guardian Jobs, Reed.co.uk (UK)	In progress	Live scrapers; Bronze integration tracked for v2.
Publications	arXiv, PubMed, SSRN, OAI-PMH repos	Planned	API integrations partially coded.
Official Statistics	Eurostat, Cedefop, 6 NSIs	Planned	Manifest-based pull design complete; parsers in progress.
Policy Scanner	Greek green-transition PDFs	Planned	Working on Greek corpus; multilingual model planned.

3. Patent Data Crawler

Status: Design and scaffolding stage — not yet operational. The specification below describes the intended first-version behaviour.

3.1 Functionality

The patent data crawler is designed to systematically retrieve patent bibliographic data from major international patent offices. Patents are a leading indicator of technological innovation and are widely used in the labour economics literature as a proxy for the emergence of new skill requirements. By tracking patent activity across technological domains, the ISABEL platform can identify which technical competencies are likely to grow in demand in the near-to-medium term. Due to restrictions, the crawlers uses the EPO API.

3.1.1 Data Extraction

The crawler retrieves patent documents from designated databases with a focus on recent filings and grants. For each patent, the following structured fields are extracted: unique patent identifier, title, abstract, full claim text where available, filing date, grant date, priority date, International Patent Classification (IPC) codes, and Cooperative Patent Classification (CPC) codes. The crawler is configured to prioritise patents filed within a defined lookback window of initially five years, while retaining the ability to conduct historical backfills as required.

3.1.2 Metadata Processing

Beyond core bibliographic fields, the crawler normalises inventor and assignee metadata to enable entity resolution across multiple filings. Assignee names are standardised against a reference list of known corporate entities, research institutions, and government bodies. IPC codes are mapped to a human-readable taxonomy cross-referenced with ISCO occupational codes and ESCO skill definitions, facilitating downstream linkage between patent-derived innovation signals and labour market skill demand.

3.2 Technical Specifications

3.2.1 Data Sources and APIs

The crawler is being built to integrate with the USPTO Open Data Portal (REST API, structured JSON), the European Patent Office OPS API (Open Patent Services, bibliographic XML), and WIPO PatentScope (HTML parsing where structured API access is unavailable). A fallback HTML parsing mode using lxml XPath selectors is provided for sources that do not expose structured endpoints.

3.2.2 Performance and Scalability

Target throughput under standard network conditions is approximately 500 to 1,000 patent records per hour. The implementation will include an in-memory caching layer to avoid redundant API calls within a session and a checkpoint mechanism that persists crawl state at intervals, allowing interrupted runs to resume from the last checkpoint. The crawler will run as a scheduled GitLab CI/CD pipeline job within the platform.

4. UK Platforms Data Crawler

Status: Operational (UK platforms). Extension to all six target countries is planned for the next version. Bronze-layer integration is in progress.

4.1 Functionality

The skills data crawler aggregates job postings from major employment platforms. Job postings are among the richest and most timely sources of information about employer skill demand, capturing requirements not yet reflected in official occupational classifications or training curricula. The first version implements two fully operational scrapers targeting UK-based platforms, Guardian Jobs and Reed.co.uk, establishing the collection pattern that will be replicated for additional European platforms in later versions (see Section 11).
The UK-only scope of the current version is deliberate: establishing a robust, validated scraping and storage pattern on a well-understood market before replicating across five further national contexts, upon users request, wherever there are relevant and required information sources reported. The six ISABEL target countries are Denmark, Greece, Poland, Spain, France, and the United Kingdom.

4.1.1 Job Posting Extraction

For each posting the crawler extracts, where present: job title, employer name, full description text, geographic location, advertised salary, posting date, application closing date, contract type, working hours, seniority level and sector. Collection follows a two-phase pattern: a link harvesting phase collects posting URLs from paginated search results, and an individual scraping phase visits each URL to extract structured metadata. Because postings are user-generated, missing values in non-essential fields are expected and handled gracefully without interrupting the run.
Both scrapers implement checkpoint persistence using Python pickle files, so an interrupted run resumes from the last successfully scraped record rather than re-harvesting all links. This matters particularly for Guardian Jobs, which may list several thousand active postings at any time.

4.1.2 Skills Standardisation

Skills standardisation is not active in this version. The current scrapers capture raw skill-tag strings only: the Reed scraper preserves extracted tags by appending them to the description field, making them available for downstream processing. Normalisation against the ESCO framework is planned as a Silver-layer transformation and is described under the future enhancements in Section 11. It will be implemented in a later version.

4.2 Technical Specifications

4.2.1 Data Collection Infrastructure

The Guardian Jobs scraper targets jobs.theguardian.com. It first reads the total vacancy count from the search header to compute pagination, then iterates across result pages to collect listing URLs. Individual pages are fetched sequentially using a flexible XPath-based field extractor (Guardian) and BeautifulSoup with CSS data-attribute selectors (Reed) that performs case-insensitive matching against definition-list elements. User-Agent strings are rotated across three standard browser identifiers, and a randomised delay (mean three seconds) is inserted between page requests. Output is currently written as CSV and pickle.
The Reed scraper targets www.reed.co.uk and scrapes a default of twelve search-result pages per run, yielding roughly 200 to 300 unique links. Given Reed's stricter rate limiting, 403 and 429 responses trigger a 60-second backoff before retry, and request delays are randomised between 2 and 5 seconds. A master dataset file is maintained as a running archive; each run deduplicates against it using the job_id extracted from the posting URL.

Note: both scrapers currently write to divergent local and network destinations. Aligning both to the platform Bronze layer via the shared ingestion interface described in Section 8 is a tracked near-term task scheduled for version 2.

4.2.2 NLP Processing Pipeline

Once raw posting text has accumulated in the Bronze layer, a Silver-layer NLP transformation is planned to extract structured skill and occupational signals from the description field. The design of this pipeline will be documented in the next version of this deliverable, alongside the ESCO-based standardisation stage described in Section 4.1.2, once the integration approach is confirmed with the analytical workstream.

5. Publications Crawler

Status: Early implementation stage. API integrations are partially coded; full operational release is planned for the next version.

5.1 Functionality

The publications crawler indexes academic papers, working papers, research reports, and technical documentation from academic repositories and grey-literature archives. This literature provides forward-looking evidence about emerging technologies and industrial transformations and their anticipated labour market consequences, complementing the real-time signal from job postings with a longer-horizon perspective. These procedures are integrated in the Researcher companion (D3.3 and D3.5)

5.1.1 Academic Database Harvesting

Target sources include arXiv, PubMed, and institutional repositories exposed. For each publication the crawler extracts title, author list, publication or submission date, venue, DOI, abstract, and subject or keyword classifications where provided. Harvested records will be stored in the ISABEL Medallion data lake and indexed for full-text search.

5.1.2 Topic and Trend Analysis

Publication abstracts will be processed using Latent Dirichlet Allocation (LDA) topic modelling to identify dominant research themes, applied on a rolling basis to track shifts in thematic emphasis over time. These trend signals help anticipate medium-term skill demand shifts driven by technological or industrial change and contextualise the patent and job-posting streams.

5.2 Technical Specifications

5.2.1 API Integration and Harvesting

Integration combines REST API calls, OAI-PMH harvesting, and structured HTML parsing. OAI-PMH enables incremental harvesting based on the datestamp field, retrieving only records added or modified since the previous run. arXiv is accessed through its REST API with keyword- and subject-class filters; SSRN, lacking a public API, is accessed by HTML parsing.

Citation-count enrichment from external bibliometric APIs and keyword extraction for cross-source linking are under evaluation and will be specified in the next version once the core harvesting pipeline is operational.

Funded by
the European Union

6. Official Statistics Crawler

Status: Early implementation stage. The manifest-based pull design is complete; source-specific parsers are in progress.

6.1 Functionality

The official statistics crawler retrieves structured quantitative datasets from Eurostat and complementary European public data authorities. Unlike the other modules, which handle unstructured or semi-structured text, this crawler ingests authoritative time-series and cross-sectional labour market statistics that form the quantitative backbone of the ISABEL platform and provide the reference framework against which signals from patents, job postings, and literature are calibrated.

The crawler uses a pull-based architecture: rather than discovering content by link traversal, it operates against a versioned manifest of known dataset identifiers and API endpoints, checking each configured dataset on a schedule and retrieving only versions that postdate the most recent locally stored copy.

6.1.1 Eurostat Data Retrieval

Eurostat is the primary source of harmonised official statistics. The module integrates with the Eurostat JSON-stat / SDMX API, which provides machine-readable access to the full Eurostat database. Key dataset families targeted include:

- Labour Force Survey (LFS) aggregates: employment and unemployment rates by NUTS2 region, occupation (ISCO-08), sector (NACE Rev.2), age group, and education level (ISCED 2011).
- Job Vacancy Statistics (JVS): quarterly vacancy rates by country and economic activity, an official benchmark for the job-posting volumes captured by the skills crawler.
- Earnings and labour costs: gross earnings, labour cost indices, and wage growth disaggregated by sector and region.
- Education and training statistics: participation in formal and non-formal education, vocational training enrolments, and qualification attainment by NUTS region.

JSON-stat responses are parsed and normalised into a flat, tidy tabular structure with consistent column naming before storage. Dimensional metadata (labels, codes, hierarchies) is retrieved from the metadata endpoints and stored alongside the data to support downstream filtering and aggregation.

6.1.2 Complementary European Data Authorities

Beyond Eurostat, there are interesting data from the following providers:

- Cedefop (European Centre for the Development of Vocational Training): skills intelligence data including the European Skills and Jobs Survey, occupational skill profiles, and outputs of the European Skills Forecasting Model.
- EURES (European Job Mobility Portal): aggregate cross-border job-posting volumes and occupational demand across EEA member states, via the EURES open data exports.
- National Statistical Institutes (NSIs): datasets are retrieved from NSI open data portals for all six ISABEL target countries. Indicatively: Statistics Denmark (Danmarks Statistik), the Hellenic Statistical Authority (ELSTAT) and the Hellenic Mechanisn of Labour Market Diagnosis, Statistics Poland (GUS), the National Statistics Institute of Spain (INE), the French National Institute of Statistics and Economic Studies (INSEE), and the Office for National Statistics (ONS).

6.2 Technical Specifications

6.2.1 API Integration and Authentication

Eurostat and most NSI APIs are publicly accessible without authentication. The crawler supports optional API-key injection — held as environment variables injected at container runtime — for

sources that require registration. Requests are built programmatically from the manifest with parameters for geography, time period, and dimension selection. A conservative rate of no more than one request every three seconds is applied to Eurostat endpoints.

6.2.2 Data Format Normalisation

Official statistics arrive in varied formats: JSON-stat (Eurostat), SDMX-ML (some NSIs), XLSX (Cedefop, some NSI portals), and CSV (EURES). A format-aware ingestion layer applies source-specific parsers to normalise all inputs to the ISABEL canonical tabular format before storage, standardising geographic identifiers to NUTS codes, occupations to ISCO-08, sectors to NACE Rev.2, and time periods to ISO 8601 quarter notation. Every record is tagged with a data_source identifier and retrieval_timestamp for provenance and reproducibility.

6.2.3 Scheduling and Incremental Updates

The update of the data is expected to be on a schedule calibrated to each source's publication cadence: for quarterly for Eurostat LFS and JVS, monthly for earnings indices and NSI flash estimates, and annually for Cedefop forecast updates. A local metadata registry records the last-retrieved version and timestamp per dataset; on each run the crawler compares the latest available vintage against the registry and downloads only when a newer version exists.

7. Policy Document Semantic Scanner

Status: Operational on Greek corpus. Extension to remaining five target countries is planned; multilingual model switch is a priority enhancement.

7.1 Functionality

The policy document semantic scanner processes large corpora of policy documents — regulatory texts, national recovery and resilience plans, green-transition strategies, and Horizon Europe programme documents — that are typically distributed as PDFs and resist conventional keyword search due to their length and terminological diversity. Rather than full-text indexing, the scanner applies semantic similarity to identify and extract only passages thematically relevant to a configured set of ISABEL research topics. The current configuration targets the green-economy transition, with a working pipeline applied to Greek green-transition policy PDFs and planned extension to the remaining five target countries.

7.1.1 Semantic Text Extraction

Each PDF is processed with the PyMuPDF (fitz) library to extract page text, which is segmented into overlapping word-level chunks (default 150 words, 50-word overlap) to avoid splitting relevant passages across boundaries. Each chunk is encoded into a dense vector using the all-MiniLM-L6-v2 sentence-transformer model, a lightweight model that encodes efficiently on CPU hardware.

7.1.2 Relevance Scoring and Filtering

Cosine similarity between each chunk embedding and each target-topic embedding is computed, and chunks are retained only where similarity to at least one topic meets or exceeds the configured threshold (default 0.4). Each retained record stores the source document identifier, page number, matched topic, similarity score, and chunk text, sorted by descending similarity.

7.2 Technical Specifications

7.2.1 Model and Infrastructure

The all-MiniLM-L6-v2 model produces 384-dimensional embeddings and is served locally from the HuggingFace cache, requiring no external API calls at inference. On CPU-only hardware it encodes roughly 5,000 to 10,000 chunks per minute, allowing a 100 to 200 document corpus to be processed within a few hours.

Language limitation: this model is optimised primarily for English. For Greek-language documents, cosine similarity scores are typically 20–40% lower than for equivalent English text, which may reduce recall on non-English corpora. The planned switch to a multilingual model (e.g. paraphrase-multilingual-MiniLM-L12-v2) is a high-priority item for the next version (see Section 11).

7.2.2 Target Topics Configuration

Target topic strings are maintained as a configurable list. The working set below represents the current provisional configuration, top-four of the topics are active and the rest subject to finalisation at a consortium's relevant topic-review discussion, the code support ten as it is:

- The transition to a green economy and its impact on the workforce.
- Creating new green jobs and sustainable employment opportunities.
- Ensuring a just transition so no regions or workers are left behind.
- Reskilling and upskilling programmes for workers in fossil fuel industries.

8. Data Integration and Storage

8.1 Platform Data Lake Integration

Crawler output is consolidated into the ISABEL Medallion data lake, hosted on MinIO — an S3-compatible object store running entirely on the CERTH on-premises node — rather than a conventional normalised data warehouse. Raw payloads land in the Bronze tier with minimal processing, preserving the source structure for reprocessing. GitLab CI/CD pipeline jobs, using Docling for document parsing and DuckDB for analytics, transform Bronze data into the Silver tier, applying standardisation, normalisation, and entity resolution. The Gold tier exposes analytically ready, cross-source datasets; these are surfaced to the Nuxt/Vue frontend through the ISABEL tool server (FastAPI) and OpenWebUI.

Each crawler writes to its designated Bronze path through a shared ingestion interface that enforces common metadata fields (source identifier, crawl timestamp, crawler version, and a record hash for deduplication) regardless of the heterogeneous structure of the source data. The canonical Bronze record schema is defined in Section 8.3.

Note: the existing UK job-posting scrapers currently write to local and network destinations rather than to this shared ingestion interface. Aligning both scrapers with the Bronze layer is a tracked near-term integration task scheduled for version 2.

8.2 Data Deduplication and Cleaning

Deduplication operates at two levels. Within a run, in-memory hash sets prevent processing the same URL or identifier twice. Across runs, the Silver tier compares a hash of the key identifying fields per record type (for example job_id for postings, patent number for patents, DOI for publications, dataset code plus vintage for official statistics) against a hash register before inserting new records. Validation detects corrupted or incomplete records, applies type conversions, and flags inconsistencies for review.

8.3 Bronze-Layer Canonical Record Schema

Every record written to the Bronze tier, regardless of source type, carries the following common metadata fields. Source-specific payload fields are nested within the raw_payload column and are not normalised at the Bronze tier.

Table 2: Bronze-layer canonical record schema

Field name	Type	Nullable	Description
source_id	VARCHAR	NOT NULL	Identifier of the originating crawler (e.g. 'reed_uk', 'eurostat_lfs').
crawl_timestamp	TIMESTAMP	NOT NULL	UTC timestamp of the crawl run that produced this record.
crawler_version	VARCHAR	NOT NULL	Semantic version string of the crawler image (e.g. '1.0.3').
record_hash	CHAR(64)	NOT NULL	SHA-256 hash of the raw payload; used for cross-run deduplication.
raw_payload	TEXT/JSON	NOT NULL	Full original response body preserved verbatim for reprocessing.
retrieval_url	TEXT	NULLABLE	Source URL or API endpoint from which the record was retrieved.
source_record_id	VARCHAR	NULLABLE	Native record identifier (job_id, patent number, DOI, dataset code).
geo_scope	VARCHAR	NULLABLE	ISO 3166-1 country code or NUTS code of geographic coverage.

Funded by the European Union

9. Quality Assurance and Validation

9.1 Data Quality Metrics

A standardised quality framework will be applied across all crawler outputs, covering four dimensions: completeness (proportion of non-null values per field, tracked per run), accuracy (cross-validation of sampled records against source pages or reference datasets), consistency (conformance to schema types, value ranges, and controlled vocabularies), and timeliness (lag between publication at source and availability in the data lake). Implementation of automated quality dashboards within the platform monitoring stack is scheduled alongside the Bronze-layer integration task in version 2; until that work is complete, quality checks are performed manually on each scraper run.

9.2 Validation Procedures

Automated validation — including schema-conformance testing against the Silver-tier schema, referential-integrity checks, and statistical outlier detection for field-distribution anomalies — is specified as a downstream GitLab CI/CD pipeline task and will be implemented as part of the Bronze-layer integration work in version 2. In the current version, validation is performed by manual inspection of scraper output files after each run, with particular attention to field-extraction failures that may indicate upstream HTML changes at the source site.

10. Deployment and Configuration

10.1 Installation and Setup

Crawlers are packaged as Docker images with pinned dependencies and version-controlled in the project GitLab. Images are built and deployed through the platform GitLab CI/CD pipeline onto the CERTH on-premises node, with per-crawler resource limits defined in the Docker Compose configuration. Configuration (target endpoints, rate limits, output paths, pagination depth) and credentials are injected at runtime from environment variables, so no credentials are baked into images. For local development each crawler can also run as a standalone Python script after installing its declared dependencies.

10.2 Scheduling and Monitoring

Execution is orchestrated as scheduled GitLab CI/CD jobs, with each crawler's schedule reflecting the publication cadence of its source. Downstream quality and transformation tasks are chained within the same pipelines so that quality assessment always runs on freshly ingested data before promotion to Silver. Operational metrics (job duration, record counts, error counts, data freshness) are exported to the platform monitoring workspace and surfaced on dashboards, with alerting on job failure, elevated error rates, or missed schedule windows.

11. Future Enhancements

11.1 Planned Capabilities

The next version will prioritise the following capability expansions:

- Complete and deploy the patent and publications crawlers to first production release, moving them from scaffolding to operational status.
- Extend the skills crawler beyond the UK to platforms covering all six target countries: StepStone / Infojobs (ES), Pracuj.pl (PL), Jobindex (DK), France Travail / HelloWork (FR), and additional UK platforms, to align data collection geography with the project's full six-country analytical scope.
- Implement the ESCO-based skills standardisation and NLP pipeline as Silver-tier transformations and connect aggregated skill-demand signals to the vacancy-based forecasting workstream.
- Integrate official-statistics outputs and the forecasting model results into a common Gold-tier schema so that observed statistics, scraped signals, and forecasts can be queried together.
- Adopt a multilingual sentence-transformer model (e.g. paraphrase-multilingual-MiniLM-L12-v2) for the policy scanner to enable reliable matching on Greek, Polish, Spanish, French, and Danish documents.

11.2 Technology Improvements

In parallel, the following technical improvements are planned:

- Align all crawlers on the shared Bronze-layer ingestion interface, removing the divergent local and network output destinations in the current UK scrapers.
- Distributed crawling using a task queue (for example Celery with Redis) to parallelise scraping across worker processes and raise throughput for high-volume sources.
- Machine-learning-based anomaly detection trained on historical crawler output to detect source HTML changes automatically, reducing reliance on manual spot-checks.
- Near-real-time ingestion for high-velocity sources such as job boards, reducing data latency from hours to minutes.
- Optimised scheduling for job vacancy sites, ensuring data latency is kept to a minimum (e.g., hours/daily batch ingestion) without triggering source rate-limits

Bibliography

Cedefop (2023). European Skills and Jobs Survey. Luxembourg: Publications Office of the European Union.
Eurostat (2024). Labour Force Survey and Job Vacancy Statistics. European Commission Statistical Office.
European Commission (2023). Horizon Europe Work Programme 2023–2024. Brussels: European Commission.
ESCO (2023). European Skills, Competences, Qualifications and Occupations — Version 1.1.0. European Commission.
Open Archives Initiative (2015). Protocol for Metadata Harvesting (OAI-PMH) Version 2.0. OAI.
USPTO (2024). Open Data Portal — Patent Full-Text Search API. United States Patent and Trademark Office.
European Patent Office (2024). Open Patent Services (OPS) API Documentation. EPO.

Funded by
the European Union