CANARIA

Canaria Job Data: Data Collection and Enrichment Methodology

1. Executive Overview

The **Canaria Job Intelligence Platform** is a large-scale, multi-source labor market data system designed to support **research-grade analysis of employer labor demand**. The platform continuously collects, standardizes, enriches, and deduplicates online job postings to produce a longitudinal dataset that is analytically robust, transparent, and suitable for economic research, workforce planning, and market intelligence applications.

Unlike traditional job board scrapers that surface fragmented or duplicate posting records, Canaria is architected around the concept of a **canonical job entity**. Multiple observations of the same underlying hiring intent across job boards, employer career sites, applicant tracking systems, and reposting cycles are resolved into a single analytical unit. This enables accurate measurement of labor demand, posting persistence, and firm-level hiring behavior over time.

Temporal Coverage	2022 to present (historical archive; volume varies by source and time period)
Geographic Scope	United States (primary); global expansion initiated in 2024
Update Cadence	Daily to hourly, depending on source and delivery mode
Primary Sources	Job aggregators (e.g., Indeed, LinkedIn Jobs) and direct employer career sites via ATS platforms

CANARIA

Scale	Billions of raw URLs processed, hundreds of millions of distinct postings identified after deduplication

Data Collection and Enrichment Philosophy

Canaria treats job postings as observable signals of employer labor demand, not as direct measures of hiring outcomes. The platform focuses on making what is observable reliable, standardized, and analytically usable, while explicitly preserving uncertainty and avoiding inference beyond the data's scope.

Raw job postings are ingested through a distributed scraping infrastructure and processed through a multi-stage pipeline combining deterministic parsing, machine learning–based enrichment, and job-level entity resolution. Detailed methodology is described in Sections 2 and 3

All raw, parsed, and enriched fields are retained for transparency and auditability.

Key Differentiator: Canonical Job Entity Resolution

A central challenge in online job postings data is severe duplication, driven by reposting behavior, aggregator syndication, and minor text changes over time. Without correction, this duplication inflates job counts and biases trend analysis.

Canaria addresses this through a configurable, multi-signal deduplication framework that resolves repeated observations of the same hiring intent into a single canonical job entity. The result is a canonical job entity that supports:

- Accurate labor demand measurement
- First-seen and last-seen posting visibility analysis
- Firm-level hiring behavior research
- Longitudinal trend analysis without double counting

CANARIA

Intended Use

The Canaria dataset is designed for research, analytics, and decision-support use cases, including labor economics, workforce analytics, compensation research, skill demand analysis, and policy evaluation. The system prioritizes methodological rigor, reproducibility, and clarity over opaque or black-box data products.

This document describes the platform architecture and methodology as of **January 2026.** Ongoing improvements to data collection, enrichment models, and coverage may result in future enhancements.

2. Data Collection Architecture

2.1 Primary Job Posting Sources

Canaria operates a distributed, continuously-running web scraping system that targets two source categories:

Category	Examples	Description
Job Aggregators	Indeed, LinkedIn Jobs	Daily crawls of search results and individual postings
Applicant Tracking Systems	Greenhouse, Lever, Workday	Direct employer career portals, 200k+ companies

2.2 Enrichment Data Sources

CANARIA

To address missing data and enhance analytical depth, Canaria integrates three categories of external data sources:

Glassdoor Salaries	User-reported compensation by title, company, location, and experience level
LinkedIn & Indeed Company Profiles	Company size, industry, founding year (when available), headquarters, employee count
Google Maps Business Data	Physical locations, ratings, hours, phone numbers, business categories

These sources are matched to job postings via fuzzy company name matching combined with deterministic identifiers. Enrichment occurs post-scraping as part of the data processing pipeline.

Company enrichment is maintained as a complementary dataset and can be joined to job postings when needed; job postings remain the primary unit of analysis to avoid enrichment-driven bias in labor market measures

3. Machine Learning Enrichment Layer

3.1 The Model Garden Architecture

The Model Garden is a microservice-based NLP pipeline that transforms raw job text into structured, analytics-ready fields. It operates independently of the scraping system, allowing separate scaling and model iteration. Job text is processed through multiple specialized models and returned as structured data for storage. This could be offered as a **standalone API** for end users to enrich other internal texts.

CANARIA

3.2 In House AI Models

Title Normalization	Canonicalizes noisy titles (abbreviations, typos) to standardized forms. Example: SW Eng II → Software Engineer Confidence scores allow researchers to filter or weight normalized titles depending on precision requirements.
SOC Classification	Assigns 6-digit Standard Occupational Classification codes using text classification on title + description. Enables occupational analysis aligned with the BLS taxonomy.
Employment Type	Classifies postings as Full-time, Part-time, Contract, or Temporary based on description text patterns and explicit mentions.
Remote Work Status	Determines if position is Remote, Hybrid, or On-site through keyword detection and location analysis.
Seniority Level	Infers career level (Entry, Mid, Senior, Lead, Executive) from title patterns and responsibility descriptions. Always returns a classification (100% complete).
Salary Prediction*	Regression model trained on 50M+ Glassdoor/Indeed observations. Requires a valid State, ZipCode, and SOC code. Returns -1 (null) when prerequisites are missing. Observed MAPE under 15% on validation data.

NAICS Prediction*	Assigns North American Industry Classification System (NAICS) codes by matching company industry labels and job text to standardized industry categories. Enables industry-level analysis aligned with official economic taxonomies.

**in progress, to be deployed*

3.3 Named Entity Recognition (NER)

A fast, dictionary-based keyword processor runs in-pipeline to extract structured entities from job descriptions:

- **Technical Skills:** Programming languages, tools, frameworks (empty array if no matches)
- **Soft Skills:** Leadership, communication, problem-solving
- **Certifications:** Professional licenses and credentials
- **Qualifications:** Education requirements, years of experience
- **Benefits:** Health insurance, 401(k), PTO mentions
- **Contact Signals:** emails and phone numbers extracted from posting text (returned as arrays; empty when not present)
- **Work Requirements:** visa sponsorship indicators, citizenship or residency requirements, security clearance requirements and level mentions
- **Work Conditions:** travel requirements and travel frequency, shift work and shift type mentions, language requirements
- **Role Characteristics:** urgent hiring cues, manager or lead indicators, number of openings, team size mentions, start date mentions

Extracted skills are filtered through a **title-skill relevance model** to remove spurious matches (e.g. Java mentioned in a Barista posting). The Skills field returns an empty array when no relevant keywords are detected.

Multi-valued attributes are represented as arrays; when no signal is present, fields are returned as empty arrays rather than null

3.4 Job Entity Resolution and Semantic Deduplication

In addition to field-level enrichment, Canaria performs job-level entity resolution to produce a canonical job entity that serves as the primary unit of analysis for downstream research and analytics. This process identifies and unifies near-duplicate postings that represent the same underlying hiring intent across different sources, URLs, and time periods:

- **Vector similarity (descriptions)**: Captures semantic similarity even when text is slightly altered.
- **MinHash/Jaccard (company names)**: Handles variations like “Macy’s,” “Macys Inc,” or “Macy’s LLC.”
- **Title similarity modeling**: Normalizes title variants such as “Junior Mechanical Engineer,” “Mechanical Eng I,” etc.
- **Geo-location clustering**: Groups postings within an adjustable radius (e.g. 10–50 miles).
- **Configurable posting window**: Defines whether two postings within a given time range (e.g. 1–6 months) should be considered versions of the same job.
- **Graph-based processing**: Used to enhance coverage and provide client-specific policy controls
    - Captures **transitive similarity** (A ≈ B, B ≈ C → unify A, B, C).
    - Supports custom deduplication policies:
        - Keep **latest** or **first** version of a job
        - Adjust similarity thresholds (e.g., 0.9 vs 0.95)
        - Change radius definitions per region or program
        - Different logic per job family or market

3.5 Field Resolution and Final Values

Job postings frequently contain conflicting or incomplete information across raw fields, parsed fields, and ML enrichment. During final dataset assembly, Canaria applies deterministic precedence logic to produce consistent final values while retaining raw and intermediate fields for transparency. For example, location may be derived from the raw posting string, normalized via parsing, and then resolved into a final standardized geography field.

All raw, parsed, and enriched fields are retained in the final dataset to support auditing, robustness checks, and alternative modeling choices.

4. Data Storage and Access

4.1 Data Storage Strategy

Canaria employs multiple database technologies optimized for specific workloads:

- **Primary Storage:** Semi-structured job documents with a flexible schema supporting source variation. Indexed for fast queries on job identifiers, company, date, and location.
- **Transactional Storage:** Durable storage for scheduling metadata, run logs, and validation results. Ensures data consistency and audit trails.
- **Message Queuing:** High-throughput queues connecting pipeline stages, with crawler statistics and deduplication caching.
- **High-Performance Cache:** Existence checks to prevent redundant scraping of identical job postings (e.g. same URL or source identifier) across ingestion pipelines.

4.2 Historical Data Retention

All job postings are archived to a historical database via automated processes. First-seen and last-seen timestamps enable longitudinal analysis of posting durations, hiring

patterns, and market dynamics. The historical database contains full posting lifecycle data since 2022.

4.3 Access

Data delivery is available via flat files (CSV or Parquet) through secure channels such as SFTP, cloud storage delivery, or shared drives. High-frequency raw feeds (multiple times per day) are available when needed, with the trade-off that cross-day deduplication may be more limited than in curated historical extracts.

API access is not currently offered.

5. System Monitoring and Observability

To ensure consistent data quality, continuity of coverage, and robustness of the collection pipeline, Canaria maintains comprehensive monitoring and quality assurance systems across all ingestion and processing stages.

5.1 Operational Monitoring

The system collects operational data at regular intervals:

- Queue depths and processing backlogs
- Crawl rates (jobs/hour, success rates)
- HTTP status code distributions (detect source failures)
- Database record counts (growth tracking)
- Request statistics (requests sent, errors, retries)

Monitoring dashboards visualize trends and trigger alerts for anomalies (queue depth spikes, error rate thresholds, crawler failures).

5.2 Data Quality Monitoring

A pipeline component tracks field-level completeness for every scraped batch. Non-empty field counts are aggregated by source and time period, enabling detection of source degradation or scraping failures. Completeness metrics inform prioritization of enrichment efforts.

5.3 Source Health Tracking

Source URLs are monitored for consecutive failure streaks. After multiple consecutive errors, a source is flagged as potentially defunct and queued for manual review. This prevents wasted crawl attempts on dead endpoints.

6. Potential Research Applications

The Canaria dataset enables empirical analysis across multiple labor economics domains:

6.1 Wage Dynamics and Compensation

- Geographic wage differentials (metro vs rural, coastal vs inland)
- Salary transparency effects (pre-/post state disclosure laws)
- Compensation trends by occupation, industry, and company size

6.2 Labor Demand and Skill Requirements

- Skill demand evolution over time (e.g., AI/ML skill growth)
- Education-occupation mismatch (degree requirements vs SOC education norms)
- Occupational mobility pathways (skill overlap analysis)

6.3 Remote Work and Geographic Flexibility

- Remote work adoption rates by occupation and industry
- Remote work wage premia/discounts
- Geographic concentration trends and migration patterns

6.4 Firm-Level Hiring Behavior

- Posting volume and posting visibility trends (first-seen/last-seen observation patterns)
- Reposting and persistence analysis (re-observation windows by occupation, region, and firm)
- Company-level hiring patterns (expansion/contraction cycles)

7. Data Access for Researchers

7.1 Sample Dataset Specifications

The provided sample contains 1,000 job postings selected to represent:

- Geographic diversity (all US regions)
- Occupational coverage (major SOC groups)
- Salary range distribution (entry to executive)
- Company size variation (startups to Fortune 500)
- Temporal spread (2022 to 2025)

7.2 Full Dataset Availability

Researchers interested in accessing the complete historical dataset (1B+ postings) for academic research should contact Canaria directly. Custom data extracts can be provided based on research requirements (specific occupations, time periods, or geographic regions).

7.3 Citation Guidelines

If using Canaria data in published research, please cite as:

CANARIA

Canaria Job Intelligence Platform (2026). U.S. Job Postings Dataset (2022 to 2026). Multi-source labor market data with ML enrichment. Available from decanaria.com