This report defines a *canonical global salary index*, combining data from international agencies, national statistics and private sources into a coherent, reproducible measure of wages. We inventory major data sources (ILO, OECD, World Bank, IMF, Eurostat, national offices like BLS/ONS, and commercial platforms like Glassdoor/Payscale/ADP) and compare their data fields, access methods, licensing, coverage and frequency. We then outline rigorous normalization: mapping occupations and industries to international classifications (ISCO, ISIC/NAICS), harmonizing regions, converting currencies and adjusting for inflation and purchasing power parity (PPP). We discuss adjusting gross wages for taxes, social contributions, benefits and self-employment, weighting by employment shares, and handling missing or biased data. Statistical methods (log-transforming pay, trimming or winsorizing outliers, calculating medians and weighted means) ensure robust aggregation. We propose constructing the index by aggregating country-level real wages into a base-year index (e.g. 100 in a reference year) and tracking it over time. Validation against official statistics (e.g. ILO wage reports, national surveys) and sensitivity tests (varying PPP or weights) are emphasized. Deliverables include the index specification, data schema, ETL pipeline design (with a mermaid flow diagram), reproducible code outline, a RESTful API specification, and governance notes on privacy/compliance. Throughout, we prioritize official data, with links to sources, and provide tables comparing sources and diagrams of the ETL and index workflows.

1. Data Sources Inventory

We classify data sources into **international/official**, **national**, and **commercial**. Each source's data fields, access, license, coverage and update frequency is summarized below (see comparison table).

- **ILO (ILOSTAT)**: Global labour statistics including wages/earnings and labour income. ILOSTAT provides free, open SDMX-JSON API access to tens of millions of data points on average earnings (hourly, monthly, annual), median wages, low-pay rates, etc. Coverage spans ~190 countries over recent decades, though country/year availability varies. Data are generally annual; definitions aim at gross earnings before deductions. License: open (CC-BY or similar). Reliability: official (ILO compiles from national sources).
- **OECD (OECD.Stat)**: Provides detailed wage data for OECD (and some partner) countries. Key datasets include average annual wages per full-time-equivalent (FTE) employee, hourly wages, and compensation (including or excluding bonuses). Access via OECD's SDMX-JSON API; data are updated annually (latest available up to ~2021). OECD data cover ~38 countries with standardized methodology. License: free to use with citation (OECD Open Data).
- **World Bank (WDI)**: World Development Indicators include related data: *inflation (CPI)*, *PPP conversion factors*, GDP per capita, and some labour metrics (e.g. GDP share of wages, or wage & salary employment rates). Access via World Bank API (data.worldbank.org) under CC BY-4.0. PPP data (e.g. "PPP conversion factor, GDP, local currency per international $") allow converting wages into constant international dollars. Coverage: ~150+ countries annually. World Bank also houses IMF's GFS data (e.g. "Compensation of employees (LCU)").
- **IMF (IFS/WEO)**: The IMF's International Financial Statistics (IFS) and World Economic Outlook (WEO) publish macro data including national CPI series (for inflation adjustment), exchange rates, and wage indicators for government or total economy. Access via IMF API (imf.org/data) under Creative Commons. CPI indices can be used to deflate nominal wages to real terms (see inflation adjustment). Coverage: most countries, frequent updates.

- **Eurostat**: The EU's statistical office provides detailed **earnings** data via its API (free, updated twice daily). Notable datasets: the *Structure of Earnings Survey (SES)* (EU-wide, every ~4 years) giving median/hourly wages by industry/occupation/gender, and regular labour cost/earnings indices. SES data cover employees (≥10 firms, excluding agriculture/public sector) and report gross hourly earnings (including overtime/bonuses). The API yields SDMX-JSON or JSON-stat data (no auth required). License: open (Eurostat open data). Coverage: 27 EU + 3 EFTA countries.
- **National statistical offices**: Each country typically publishes earnings data, often through labour force or business surveys. Examples:
  - *USA – BLS OES*: The Occupational Employment Statistics (OES) survey provides employment and mean/median wages for ~830 occupations, by state/metro and by industry. Updated annually (May); data accessed via BLS Public Data API (JSON) or downloads.
  - *USA – BLS NCS*: The National Compensation Survey covers wages/benefits by occupation/industry, but less publicly accessible.
  - *UK – ONS ASHE*: Annual Survey of Hours and Earnings gives average and median full-time earnings by occupation, industry, region. ONS offers an open API (no key required) and downloads; updates yearly (latest for 2025, OGL license).
  - *Others*: Many countries (e.g. Canada's LFS, Australia's ABS, India's MOSPI, Brazil's IBGE) have similar surveys. Often data by broad occupation (ISCO or national) and industry (ISIC or NAICS). Licenses vary (often open). National coverage: each country's workforce. Frequency: mostly annual or quarterly.
- **Commercial/job-related data**: Several private sources compile salaries from job postings or user surveys. These include:
  - *Glassdoor*: User-submitted salaries and employer reports. **Note**: Glassdoor's public API was closed in 2021 (now only partners can access it). Data fields include salary ranges, distributions by title/location, but they are proprietary and biased toward tech/US sectors. Access would require scraping or partnerships. License: proprietary (no public license).
  - *PayScale*: Payscale's database (~60M profiles globally) offers salary reports by job, skill, region. They provide a REST API (Jobalyzer) for customized compensation reports (inputs: title, location, skills; outputs: base/bonus/total pay). Access requires subscription. Data fields include base pay, bonuses, percentiles, and factors (experience, education). Coverage: mainly US (global reach growing).
  - *LinkedIn Salary/Indeed*: Aggregated salary estimates based on user inputs or job ads. No public API; some scraping tools exist.
  - *ADP, Paychex, Mercer, others*: Payroll processors and consulting firms hold large pay databases (often by industry). They release occasional reports (e.g. ADP jobs reports, Mercer salary guides) but not open data. Inclusion would require licensing.
  - *Job boards (Indeed, Monster)*: Can be scraped for offered salaries but often missing ranges or obscured; not considered reliable for rigorous index without major filtering and bias adjustments.

Table 1: Comparison of major salary data sources (fields, access, etc):

| **Source / API** | **Data Fields** | **Access / License** | **Coverage / Frequency / Notes** |

|---|---|---|---|
ILOSTAT (ILO)	Earnings (average, median) by industry, occupation, sex; labour income shares	Free SDMX-JSON API (no key); open license (CC-BY-like)	Global (~190 countries); annual data; time series often from 2000s onward. Official and comprehensive.
OECD.Stat	Avg. annual wages (FTE), median wages, compensation of employees	OECD SDMX-JSON API; free with citation	OECD+ countries (~40); annual, up to ~2021; high-quality official data.
World Bank (WDI)	GDP, CPI, PPP conversion factors, wage share indicators	World Bank API (CSV/JSON); CC BY-4.0	Global (~200); annual (most data); includes PPP (ICP) and CPI. Useful for conversion factors.
IMF (IFS/WEO)	CPI indexes, national accounts (compensation share), exchange rates	IMF API (SDMX) or data.imf.org; free	Global; CPI monthly/annual; WEO inflation; coverage varies.
Eurostat	Structure of Earnings Survey: median/mean hourly wages, low-pay, etc	Eurostat REST API (JSON-stat/SDMX); open data (ODC BY)	EU27+EFTA; SES (every 4yrs, latest 2022), also monthly labour cost indexes. Very detailed by industry/occupation.
US BLS (OES)	Mean/median wages, employment by occupation, industry, area	BLS Public Data API (JSON); open (no auth required)	USA; annual survey (May); ~830 occupations; breakdown by state/metro and industry.
US BLS (NCS)	Total compensation by occupational category (industry)	Published tables; no bulk API (needs special access)	USA; less frequent; includes benefits.
UK ONS (ASHE)	Average and median gross weekly earnings by occupation, industry, region	ONS API (open, no key); OGL license (UK open)	UK; annual (latest 2025); full-time employees in firms >=10.
Canada StatsCan	Wage/salary data by occupation/industry (NOC/NACE)	CANSIM/StatCan tables; some API	Canada; annual/quarterly surveys (LFS, SSE).
Others (Eurostat etc)	*Example:* Germany Destatis has Arbeitsentgelt (earnings) by sector	Varies (web download, API)	Many national offices (France INSEE, India MOSPI, etc.). Usually official.
Glassdoor	User-submitted salaries, percentiles, company ratings	No public API (closed 2021); data via web/SaaS	Crowd-sourced; biased; coverage mainly US/tech jobs; usage requires license.
PayScale (Jobalyzer)	Salary distribution by job title, location, skills, etc.	REST API (paid); proprietary	~60M profiles (global, US-heavy); updated continuously.
Indeed/LinkedIn	Aggregated salary by job/location (user/contributor data)	Limited APIs; mostly scrape; proprietary	Large coverage but unverifiable accuracy; more useful for trends than absolute values.
ADP/Paychex/Mercer	Payroll/compensation surveys by industry/size	Private reports (no public API)	High quality but commercial; could provide calibration data if licensed.

2. Data Normalization

Before aggregation, we must align all data to common classifications, currencies, and definitions:

- **Occupation & Industry Mapping**: Convert all occupation and industry codes to international standards. We recommend using ILO's *ISCO-08* for occupations (free and widely used) and UN's *ISIC Rev.4* (or regional variants NACE/NAICS) for industries. Many national sources use their own codes (e.g. US SOC/NOC, NAICS) so crosswalks are needed. For example, U.S. SOC codes can be mapped to ISCO via published concordances; NAICS to ISIC can use UN/EU concordances. Once mapped, we can compare wages for the "same" occupation globally. (Labor economists routinely map occupations to analyze earnings distributions.)
- **Regional Harmonization**: All country or sub-national data should use a consistent geography reference (ISO country codes). For subnational indices, group into larger regions (e.g. EU regions, US states) if needed, but the core index will use country-level aggregates.
- **Currency Conversion and PPP**: To compare wages across countries, convert local currencies to a common unit (e.g. USD) adjusting for purchasing power. We use *nominal exchange rates* for simple currency conversion, and *PPP conversion factors* (from World Bank/ICP) to control for price level differences. Specifically, first express wages in local currency units (LCU). Then:
  1. Adjust for local inflation to get real LCU (base year).
  2. Convert real LCU to international dollars via the PPP factor.

The World Bank defines PPP as a "currency conversion factor... [that] eliminates the differences in price levels between countries, allowing comparisons". Use the official PPP series (e.g. ICP 2017/2021, extrapolated annually). As a check, also compute an index using market exchange rates and report sensitivity.

- **Inflation Adjustment**: Convert nominal wages to constant prices using a consumer price index (CPI) or wage-specific deflator. For each country, select an appropriate price index (e.g. national CPI from IMF WEO/IFS or World Bank). For each year *t*:

  $$\text{Real wage}_{t} = \frac{\text{Nominal wage}_{t}}{\text{CPI}_{t}} \times 100.$$

The U.S. BLS uses CPI to compute real wages, a standard approach. We use the CPI series with a common base year (e.g. 2010=100). This yields real wages in constant local currency. (Alternatively, use a GDP deflator or specific wage deflator if available.) After inflation deflation, convert to PPP-adjusted international dollars as above.

- **Taxes and Social Contributions**: Decide whether the index should reflect *gross* wages or *net* (take-home pay). We propose indexing **gross wages**, but document tax effects separately. To optionally adjust, use OECD's *Taxing Wages* data or ILO/National tax statistics. For example, OECD reports the labor tax wedge (taxes+SSC minus cash transfers as % of labor cost) for average earners. One could compute **net wage = gross wage × (1 – tax wedge)** for comparability. Alternatively report a *tax-adjusted* index by subtracting average tax rates. All cash benefits (bonuses, severance, etc.) are included in gross wage where data permit; fringe benefits (health insurance, etc.) are often excluded and should be noted.
- **Benefits and Self-Employment**: Decide on inclusion of non-wage benefits (pensions, health contributions) and self-employed income. For simplicity, focus on *wage & salary earnings* of employees (the bulk of income statistics). Self-employed income is typically excluded from wage surveys; one could add a separate component using labor share of national income or GNI data, but this is complex and may be omitted initially. Clearly document these scope choices.

- **Weighting and Sample Adjustments**: When aggregating microdata, use survey sample weights (if available) or weight occupations by employment share. For aggregated data, weight country contributions by population or economically active population (or total wage bill). A common choice is *workforce share weighting* (so larger economies count more). For example, a country's index weight could be proportional to its total number of wage earners or total GDP.
- **Missing or Biased Data**: Many sources have gaps (e.g. SES every 4-yr) or biases (job sites oversample certain sectors). We will impute missing country-year values using interpolation or by borrowing from nearby countries (econometric models using GDP/capita as predictors). We will also flag outliers: for skewed wage distributions, apply log-transform and z-score thresholds. For instance, drop observations >3 standard deviations above the mean on log scale or use percentile trimming (e.g. cap at 99th percentile). Inspect each source for systematic bias (e.g. Glassdoor might overstate tech salaries) and adjust or exclude as needed. Uncertainty from such imputation should be tracked.

3. Index Construction

With normalized wage data, we build the index in steps:

1. **Aggregate to Country-Level**: For each country-year, compute a summary wage measure (e.g. median or mean wage). Preferably use *median real wage* to reduce outlier influence (supported by log-normal wage theory). If only mean wages are available (e.g. GDP/GNI per employed), note that distinction.
2. **Apply Weighting**: Combine countries into a global index. Two strategies: (a) **Global Average**: Weighted average of country wages, with weights = share of world employed population (or share of world wage bill). (b) **Fixed Base Index**: Choose a base year (e.g. 2020) and base =100. For each country-year, compute index = (CountryRealWage / CountryRealWage_base)×100. Then compute world index = weighted average of country indices (weights as above). This yields a time series of a "world wage level" indexed to base.
   Mathematically:
   $$ \text{Index}_t = 100 \times \frac{\sum_i w_i \cdot W_{i,t}}{\sum_i w_i \cdot W_{i,0}},$$
   where $W_{i,t}$ is country *i*'s real wage in year *t*, and w_i is the country's weight (e.g. workforce). Optionally compute sub-indices by occupation or industry similarly.
3. **Time-Series Consistency**: Use chain-linking if definitions change. If sources update base or classification (e.g. ISCO version), apply concordance. Fill short gaps with linear interpolation or last observation carried forward only for minor gaps; major gaps may force model-based imputation (e.g. ARIMA or growth-correlation with GDP).
4. **Outlier Detection**: Identify anomalies in the series. For example, if a country's index jumps >20% in one year (beyond inflation), investigate data source changes. Possible actions: smooth sudden changes unless they reflect real policy shifts. The log transformation method helps spot extreme tails. Document any manual corrections (e.g. splice series where data is deemed unreliable).
5. **Uncertainty Quantification**: Estimate confidence bounds. For series based on survey data, use reported standard errors if available (e.g. from OECD). Absent that, perform *bootstrapping*: resample microdata or use Monte Carlo simulations of imputation error to derive a range. At minimum, present a sensitivity range by comparing index with/without certain sources (e.g. private data excluded).

4. Validation and Sensitivity

Validate the index at multiple levels:
- **Against Official Statistics**: Compare aggregate index trends to benchmarks like ILO's Global Wage Report or national average wages. For example, check if the world wage index growth aligns with global GDP per capita growth. Compare country indices to independent series (OECD's average wages, World Bank's GDP per worker, etc). Large discrepancies signal issues.
- **Cross-Checking**: For overlapping periods, ensure consistency between sources (e.g. EU countries in ILO vs. Eurostat data). Use sector breakdowns where possible to check plausibility (e.g. manufacturing vs services wages).
- **Sensitivity Analyses**: Systematically vary key assumptions:
  - Use market exchange rates instead of PPP and compare indices.
  - Alter weighting scheme (e.g. equal-weight countries vs population-weight) and measure index changes.
  - Exclude one major data source (e.g. Glassdoor) and assess impact.
  - Try alternative inflation series.

  Document which choices most affect results. Ideally, the core index should be robust (small changes).

5. Implementation Design

5.1 Canonical Index Specification

The *Global Salary Index (GSI)* is defined as follows: let $w_{i,t}$ = real median wage in country *i* at time *t* (in PPP-adjusted international dollars, base year 100). The GSI at time *t* is

$$GSI_t \;=\; \frac{\sum_i E_i \, w_{i,t}}{\sum_i E_i \, w_{i,0}} \times 100,$$

where E_i = employment in country *i* (or worker population) in base year *0*. Thus $GSI_0 = 100$. We provide both GSI (average) and **sub-indices** by occupation (e.g. GSI managers, GSI manufacturing) using the same formula on subsets. The data schema includes fields: *Country, Year, Occupation (ISCO code), Industry (ISIC code), Mean/Median Wage (local), CPI deflator, PPP, Adjusted Wage, Weight*, etc.

5.2 Data Schema

A relational schema or flat table for the final data might include columns:

Column	Type/Description
Country ISO code	String (e.g. "USA")
Year	Integer (e.g. 2024)
Occupation ISCO code	String (e.g. "2512" for software developers)
Industry ISIC code	String (optional; could leave as "TOTAL" for all jobs)
Wage (local currency)	Float (e.g. 50000) – nominal average/median as reported
CPI index (base 100)	Float – for inflation adjustment
PPP factor (LCU to Int$)	Float – to convert to international dollars (base year)
Real Wage (Int$ base year)	Float – inflation & PPP-adjusted wage
Employment weight	Float – country's worker population or share

We also track metadata: data source identifier, confidence flags, etc. The schema can be in JSON or CSV form for the API.

5.3 ETL Pipeline Design

We recommend an automated ETL (Extract-Transform-Load) pipeline. A Mermaid flowchart summarizes it:

```mermaid
flowchart TD
  %% Extraction from sources
  subgraph Extract
    A1[ILOSTAT API] --> B[Raw Data Store]
    A2[OECD API] --> B
    A3[World Bank API] --> B
    A4[IMF API] --> B
    A5[Eurostat API] --> B
    A6[BLS API] --> B
    A7[ONS API] --> B
    A8[Glassdoor/Payscale (if any)] --> B
  end
  %% Transformation and Normalization
  subgraph Transform
    B --> C[Normalize fields: currency, CPI, PPP, taxonomy]
    C --> D[Map occupations to ISCO, industries to ISIC]
    D --> E[Adjust wages: inflation deflation, tax wedge]
    E --> F[Detect/Clean outliers and missing values]
    F --> G[Aggregate by country and weights]
  end
  %% Loading into index
  subgraph Load
    G --> H[Database: Global Wage Index]
    H --> I[API Layer & Analytics]
  end
```

This shows data from each source being pulled into a raw store, then sequentially normalized (currency, CPI, PPP), classified (ISCO/ISIC mapping), adjusted (inflation, taxes), cleaned, and aggregated into the final index database.

5.4 Reproducible Code Outline

A recommended stack: Python or R scripts scheduled via a workflow tool (e.g. Airflow, Apache NiFi). Each step is coded as a function/module:

1. **Extraction**: Functions use APIs (e.g. `requests` in Python) to download JSON/CSV from ILO, OECD, etc. Rate-limit and cache data. Log version/dates.
2. **Transform**: Use Pandas (or data.table in R).
   - Convert currencies: multiply by exchange rates.
   - Apply CPI: `real_wage = wage / CPI * 100`.
   - Apply PPP: `intl_wage = real_wage / PPP * 100`.

- Classification: use mapping tables (CSV lookup) to add ISCO/ISIC codes.
3. **Cleaning**: Apply filters (`value < upper_percentile`), impute missing (e.g. linear interpolation), annotate gaps.
4. **Aggregation**: Compute medians/means. Use population weights (from UN or World Bank).
5. **Indexing**: Chain-link to base year. Calculate confidence intervals (bootstrapping: resample country contributions).
6. **Export**: Save final table (e.g. in SQL or Parquet), and generate CSV/JSON for API.
Each step must be logged for reproducibility, and data snapshots stored (e.g. in Git or DVC).
5.5 API Design
We envisage a RESTful API for the index (e.g. built with Flask or FastAPI). Example endpoints:
- `GET /index?year=YYYY` – returns global index value and confidence bounds.
- `GET /country?country=USA&year=YYYY` – country wage index series.
- `GET /occupation?isco=2512` – index for occupation code (with breakdown by country).
- `GET /metadata` – data sources, update dates.
Responses in JSON. Rate limiting not critical if data is public. Use HTTPS and versioning.
5.6 Governance, Privacy, Compliance
- **Privacy**: We use aggregated data (no personal records). However, for any microdata (like Glassdoor) ensure compliance: remove any personal IDs, abide by terms of use, and aggregate enough to prevent re-identification.
- **Licensing**: Track licenses of all inputs. Use only open or permitted data for public release. Where proprietary (e.g. PayScale API), ensure usage rights or exclude from public dissemination.
- **Data Security**: Store raw data securely; only share aggregated index values.
- **Documentation**: Publish methodology and code (e.g. on GitHub) to enable auditing.
- **Update and Versioning**: Treat the index as a living dataset. Each update should be versioned with changelog.
Figure 2: Index Calculation Flow (illustrative process):
```mermaid
flowchart LR
    A[Country Wages (PPP, Real)] --> B[Apply weights (employment share)]
    B --> C[Compute weighted average]
    C --> D[Chain-link to base year]
    D --> E[Global Salary Index]
    E --> F[Sensitivity Analysis & CI]
```
This diagram shows how country-level normalized wages are combined (weighted average) and indexed to produce the Global Salary Index, then analyzed for sensitivity.
6. Tables and Diagrams
Table 1 (above) compares key sources. Additional tables can list classification standards (ISCO/ISIC) or tax rates by country. We have provided mermaid diagrams for the ETL pipeline and index flow. These visual aids clarify the process.
7. Conclusion
We have outlined a comprehensive methodology for a **Global Salary Index**. Key points: use official and high-quality data where possible (ILO, OECD, Eurostat, national surveys), augment

carefully with private data, perform rigorous normalization (CPI/PPP/tax adjustments), and apply robust statistical aggregation (median, weighted mean). The process is automated, transparent and reproducible, with validations against known benchmarks. The deliverables include detailed specifications, code and documentation to support the index's ongoing maintenance. This approach fills the need for a unified measure of wage trends worldwide, useful for economists and policymakers.
Sources: The above methodology builds on standards and publications from ILO, OECD, Eurostat and others.