Nonprofit Financial Records

A Guide to the Dataset

February 2026

Part 1: Substantive Insights

1. Overview

This dataset provides longitudinal financial data — balance sheet items, profit and loss aggregates, employee counts, and CIC34 community-impact narratives — for nonprofit companies registered at Companies House. It was produced by the UK Third and Civil Society Sector Database (https://uk-third-sector-database.github.io/) project, which collects, processes, and links public administrative data on civil society organisations across the United Kingdom.

The data is drawn from the annual accounts that nonprofit companies file at Companies House. Two complementary sources are used: structured XBRL filings (machine-readable accounts) and PDF accounts processed through a structured-output LLM extraction pipeline. The merged dataset gives broader coverage than either source alone, particularly for smaller filings such as Community Interest Company (CIC) abridged accounts.

The dataset spans financial years from 0 to 2109 and contains 1,299,302 financial-year records for 212,732 nonprofit companies. Each row represents one company's accounts for a single financial year.

1,299,302

Financial Records

212,732

Nonprofit Companies

3

Source Streams

0–2109

Financial Years

2. What are Nonprofit Financial Records?

Companies registered at Companies House are required to file annual accounts. The level of detail filed varies by company size — micro-entities and small companies file abridged or filleted accounts, while medium and large companies file full statutory accounts. CICs additionally file a CIC34 community-impact report alongside their accounts.

XBRL Accounts (Companies House monthly extracts)

Companies House publishes monthly bulk extracts of accounts filed in machine-readable XBRL format. The extracts contain structured tagged values for balance sheet items, profit and loss aggregates, and employee counts. These are downloaded from the Companies House accounts download site (https://download.companieshouse.gov.uk/en_monthlyaccountsdata.html).

PDF Accounts (LLM-extracted)

For older filings and CIC abridged accounts that pre-date or were filed outside the XBRL stream, the project extracts financial line items from the original PDF accounts using a structured-output LLM pipeline. The extracted values are mapped onto the same column schema used for XBRL, so the merged dataset has a uniform shape regardless of source.

CIC34 Community Interest Reports

Community Interest Companies file a CIC34 alongside their accounts. The CIC34 contains four narrative sections: the company's activities and impact, stakeholder consultation, directors' remuneration, and any asset transfers. These narratives are extracted from the same PDF filings.

3. Dataset Contents

The dataset is organised into seven column groups. Each row represents one company-year record. Where a value is missing, the cell is left blank.

Identifiers and Time Period

Field	Description	Type
`uid`	Spine identifier (e.g. `GB-COH-08411754` , `GB-CHC-1000019`)	Text
`coyno`	Companies House registered number (8-digit, zero-padded)	Text
`entity_current_legal_name`	Company name as filed	Text
`fy`	Financial year (calendar year of FYE)	Numeric
`fys` / `fye`	Financial year start / end dates	Date

Aggregate Financials

Field	Description
`turnover_gross_operating_revenue`	Total turnover / operating income
`operating_profit_loss`	Operating profit or loss
`profit_loss_for_period`	Profit or loss for the period

Balance Sheet

Balance-sheet line items including `tangible_fixed_assets` , `debtors` , `cash_bank_in_hand` , `current_assets` , `creditors_due_within_one_year` , `creditors_due_after_one_year` , `net_current_assets_liabilities` , `total_assets_less_current_liabilities` , `net_assets_liabilities_including_pension_asset_liability` , `called_up_share_capital` , `profit_loss_account_reserve` , and `shareholder_funds` .

Profit & Loss Detail

Detailed P&L items including `cost_sales` , `gross_profit_loss` , `administrative_expenses` , `raw_materials_consumables` , `staff_costs` , `depreciation_other_amounts_written_off_tangible_intangible_fixed_assets` , `other_operating_charges_format2` , `profit_loss_on_ordinary_activities_before_tax` , `tax_on_profit_or_loss_on_ordinary_activities` , `wages_salaries` , `dividends_paid` , and `government_grant_income` .

Staff

`average_number_employees_during_period` : average number of persons employed during the financial year.

CIC34 Narratives (CICs only)

Four free-text fields extracted from the CIC34 form: `cic34_activities_impact` , `cic34_stakeholder_consultation` , `cic34_directors_remuneration` , `cic34_asset_transfer` .

Metadata

Provenance and lineage columns including `source_file` , `file_type` , `taxonomy` , `balance_sheet_date` , `from_prior_year` (whether the row came from a prior-year companion filing), `source_dataset` (`xbrl` or `pdf_extraction`), `is_cic` , and `csotype` (`CIC` / `Charity` / `Co-operative / Mutual` / `Other`).

---

4. Coverage & Completeness

The combined dataset draws from XBRL and PDF sources. Coverage varies by financial year, by company size, and by individual line item. The charts and tables below summarise where the data is fullest and where it is sparsest.

Records by Source and Year

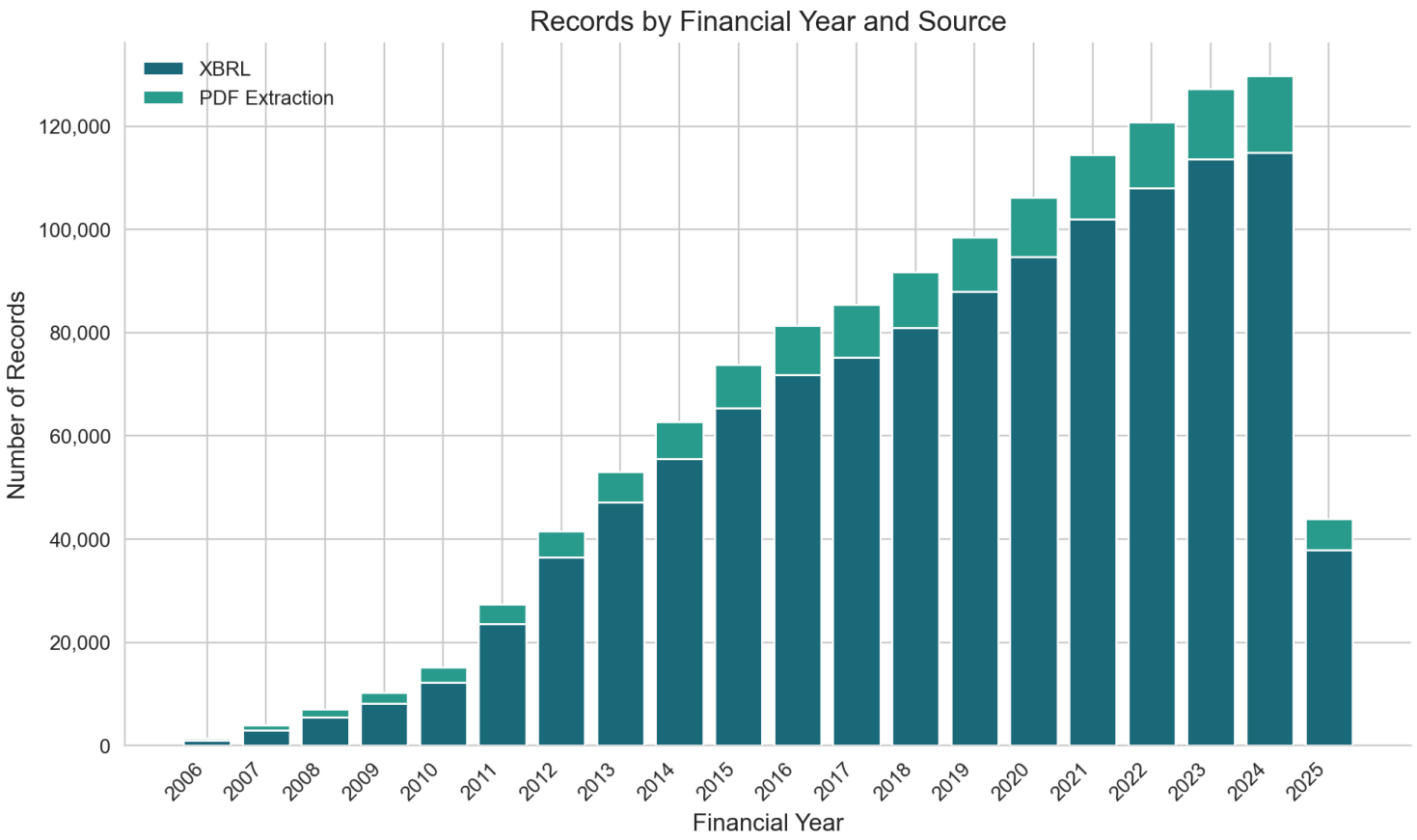

Figure: Records by source and year. Earlier years are XBRL-dominated; recent CIC filings come predominantly via the PDF extraction stream.

Source Overlap

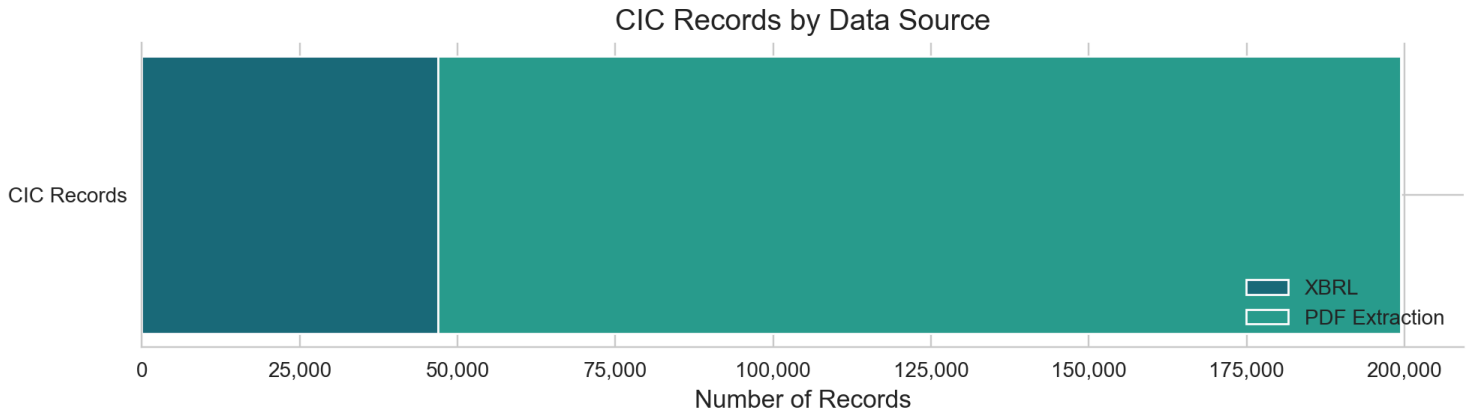

Figure: Most records are uniquely sourced; the XBRL stream is preferred where both are available.

Variable Availability

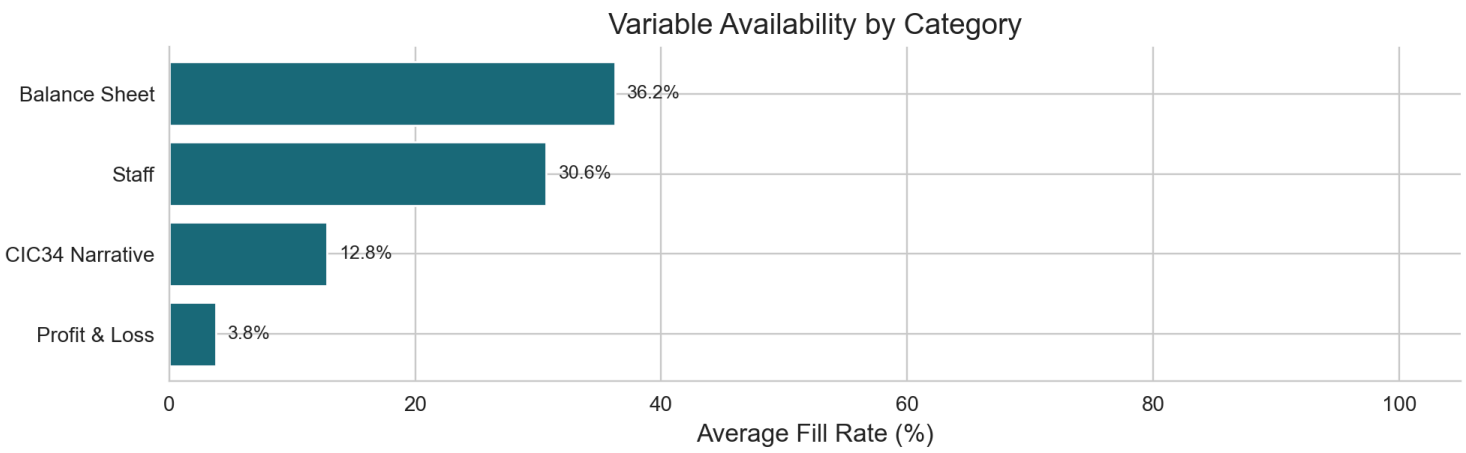

Figure: Fill rates for the financial columns across all records.

Headline Coverage

Dataset Summary

Combined accounts dataset headline statistics

Metric	Value
Total Records	1,299,302
Unique Organisations	212,732
Unique CICs	42,345
Variables	50
Earliest Financial Year	0
Latest Financial Year	2,109
XBRL Records	1,146,801
PDF Extraction Records	152,501

Year-by-Year

Coverage by Financial Year

Unique organisations with accounts data per year

Financial Year	Unique Orgs	Total Active Orgs	Coverage Rate
2,010	15,165	295,446	5.1%
2,011	27,403	301,070	9.1%
2,012	41,528	290,345	14.3%
2,013	53,042	297,220	17.8%
2,014	62,667	302,837	20.7%

Coverage by Financial Year

Unique organisations with accounts data per year

Financial Year	Unique Orgs	Total Active Orgs	Coverage Rate
2,015	73,747	310,745	23.7%
2,016	81,347	318,114	25.6%
2,017	85,492	321,242	26.6%
2,018	91,697	322,376	28.4%
2,019	98,495	328,432	30.0%
2,020	106,124	341,339	31.1%
2,021	114,503	348,225	32.9%
2,022	120,749	354,050	34.1%
2,023	127,268	361,433	35.2%
2,024	129,703	369,367	35.1%
2,025	43,900	377,876	11.6%

5. What Can You Learn?

The dataset supports questions about the size, financial health, and activity of the UK nonprofit company sector. A non-exhaustive list of uses:

Sector Growth Trends

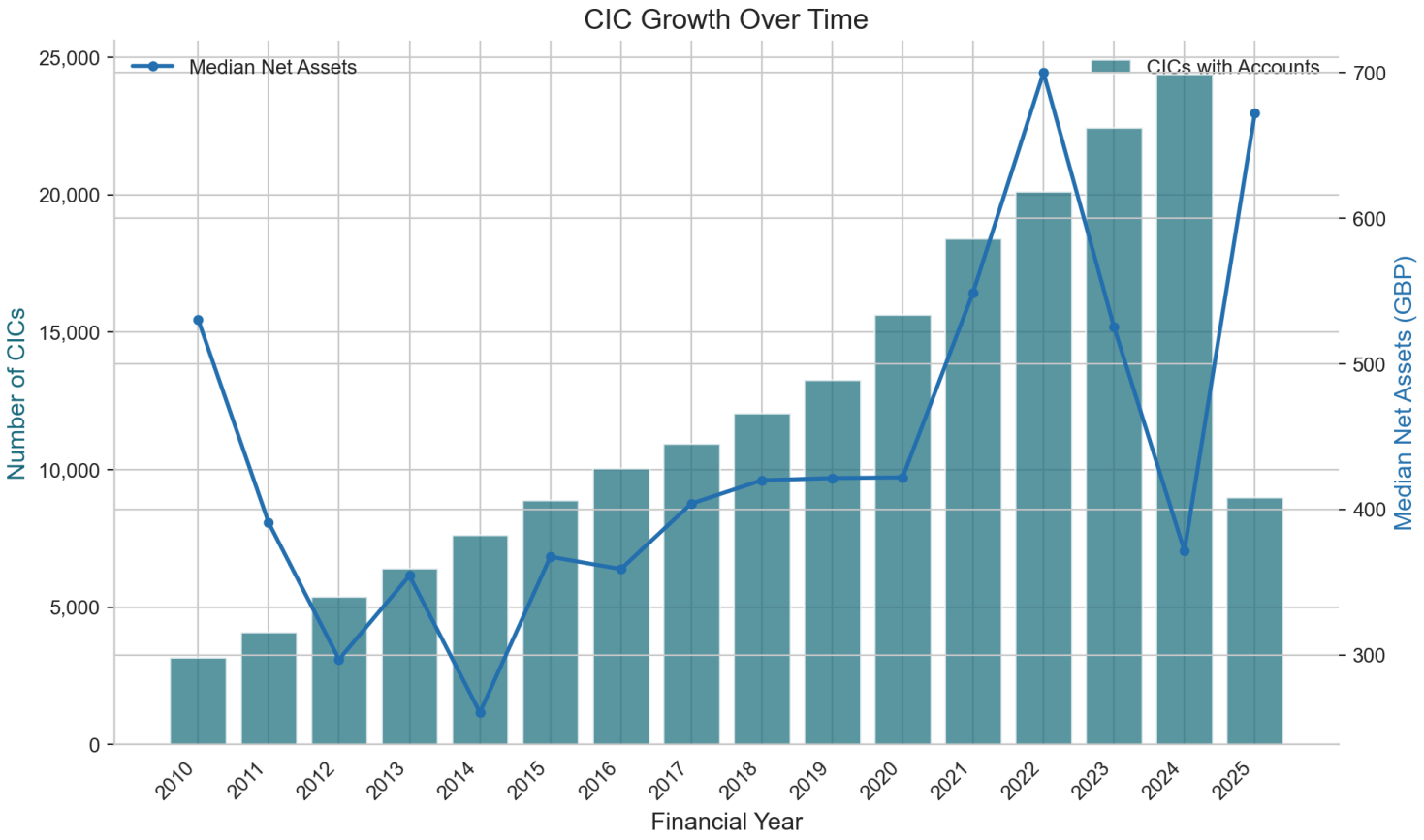

Figure: Cumulative growth of Community Interest Companies with reported accounts over time.

Financial Profile

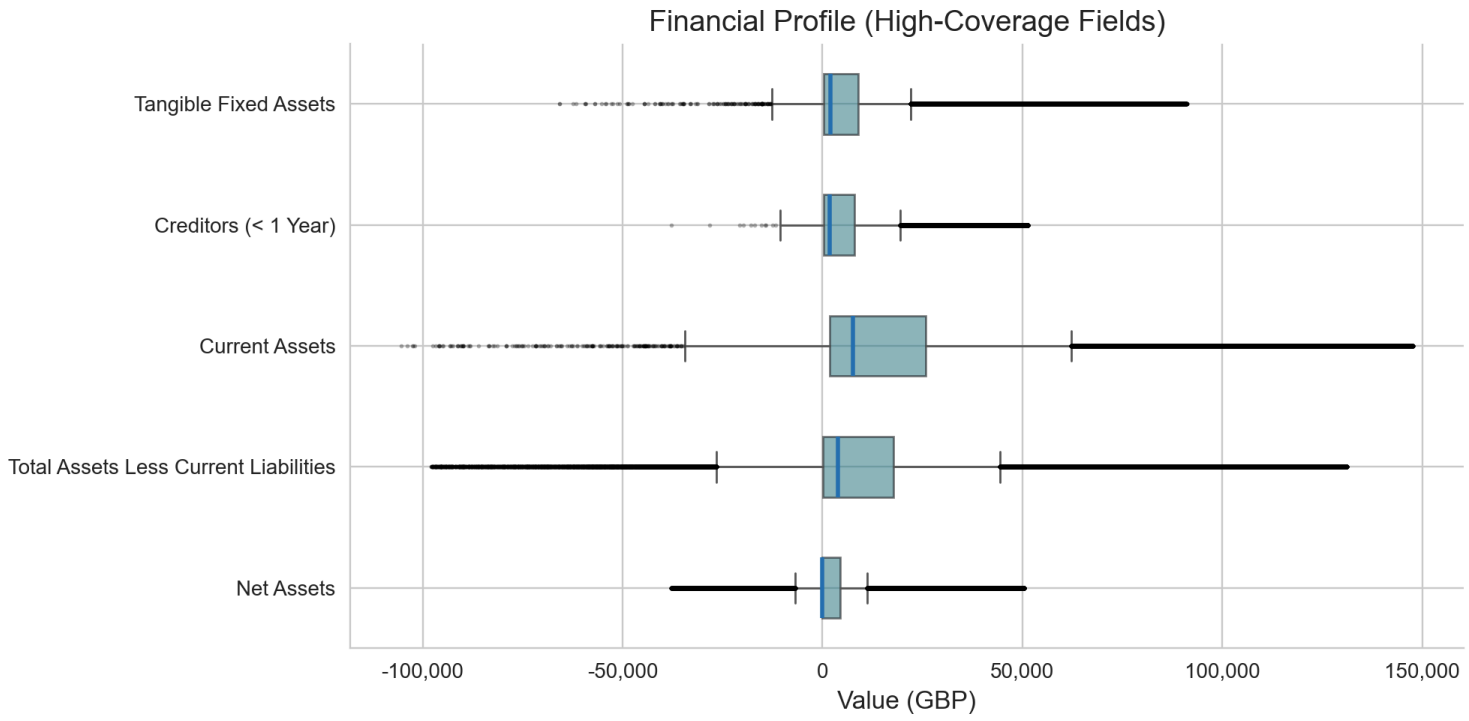

Figure: Distribution of net assets, turnover, and operating profit/loss across the sector.

Employment Dynamics

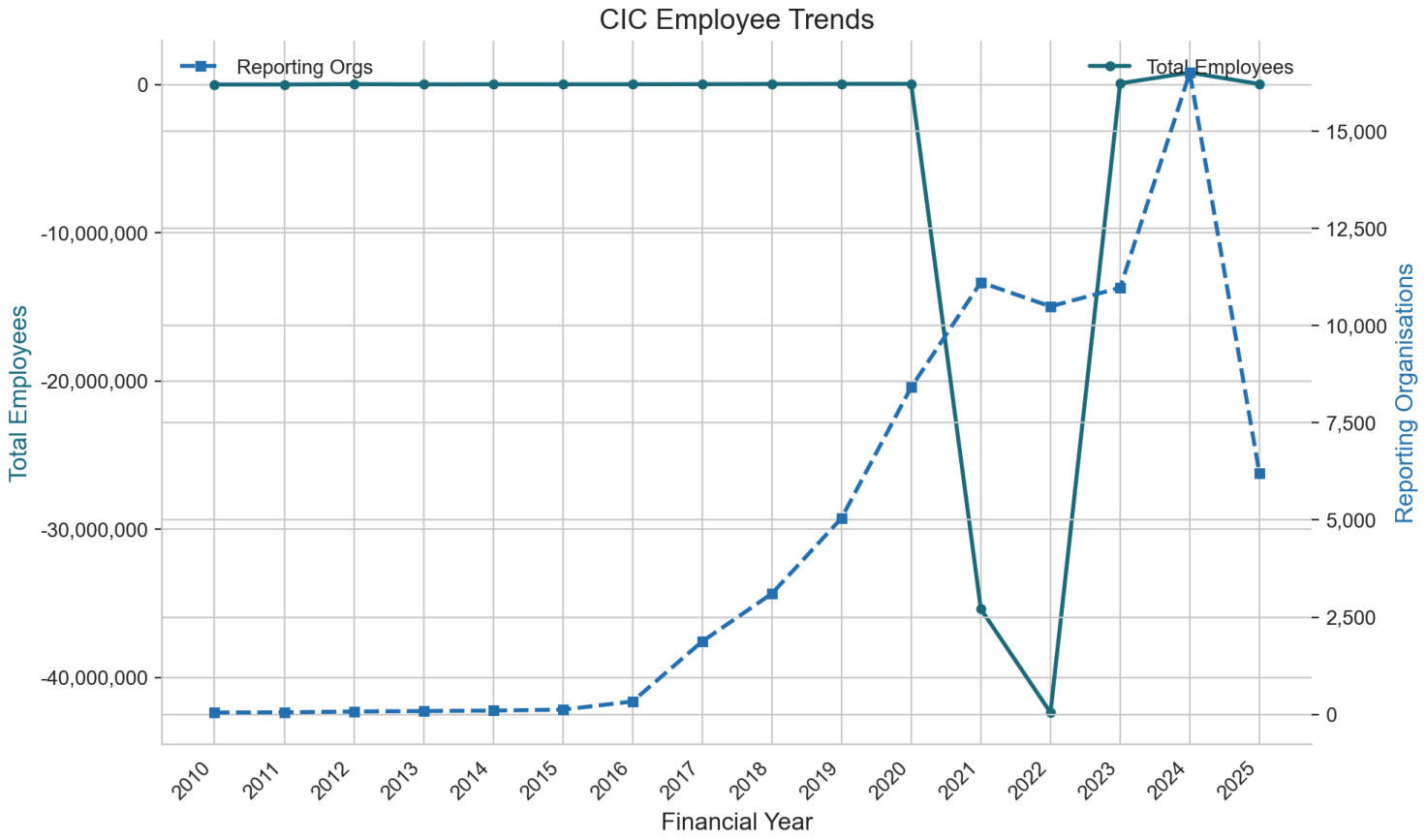

Figure: Average employee counts across years and company types.

CIC34 Narratives

The CIC34 narrative fields support qualitative analysis of community impact and stakeholder engagement. Coverage rates by field:

CIC34 Narrative Field Coverage

Coverage rates for CIC-specific narrative fields

Field	Non-Missing	Total CIC Records	Coverage Rate
Activities & Impact	170,145	199,409	85.3%
Stakeholder Consultation	167,801	199,409	84.1%
Directors Remuneration	166,099	199,409	83.3%
Asset Transfer	162,584	199,409	81.5%

Director Remuneration

Director remuneration patterns can be inferred from the CIC34 directors' remuneration narrative. Among CICs reporting remuneration: coverage rate 83%, median statement length 37 characters.

6. Limitations & Caveats

Reporting Lag

Accounts are typically filed 6–12 months after the financial year ends. The most recent financial years will therefore be incomplete.

Filing Thresholds

Many CICs file as micro-entities and are exempt from disclosing detailed P&L information. Coverage of fields such as `turnover_gross_operating_revenue` and `staff_costs` is materially lower than balance-sheet coverage as a result.

Variable Financial Year Periods

Financial years vary in length, particularly for newly incorporated or dormant companies. Compare values per-year cautiously when the FY length differs from twelve months.

PDF Extraction Quality

PDF-sourced records are extracted via a structured-output LLM pipeline with a label-mapping process to align extracted line items with the XBRL column schema. While extraction quality is high (~95% precision on cross-validated fields), residual errors are possible. The `source_dataset` column allows users to filter by source.

Spine Filtering

Only nonprofit companies present in the project Spine (the deduplicated register of UK third-sector and civil-society organisations) are included. Profit-making companies and other Companies House registrants are excluded.

What is NOT in the Data

The dataset does not include cash-flow statements, notes to the accounts in full, contingent liabilities, related-party transactions, or auditor qualifications. For these, consult the original PDFs at Companies House.

---

7. Citation & Licence

Suggested Citation

```
McDonnell, D. et al. (2026). Nonprofit Financial Records.
UK Third and Civil Society Sector Database.
Available at: https://uk-third-sector-database.github.io/data/
Licensed under CC BY 4.0.
```

8. Changelog

v1.1 — May 2026

- Incorporated ~42,000 newly extracted CIC PDF accounts.
- Added `csotype` column (CIC / Charity / Co-operative / Mutual / Other).
- Refreshed coverage statistics throughout the guidance.

v1.0 — February 2026

- Initial public release of the merged XBRL + PDF combined dataset.

Part 2: Technical Annex

This annex provides the full technical detail on data sources, processing steps, coverage statistics, and known limitations for the combined accounts dataset.

A1. Pipeline Architecture

The dataset is produced by a multi-stage pipeline that combines two source streams into a single deduplicated CSV.

Stream 1: XBRL Accounts (monthly bulk extracts)

Companies House publishes monthly ZIP archives of XBRL-tagged accounts. Each archive is downloaded, parsed using `stream-read-xbrl`, and appended to a per-month CSV with a uniform schema. The combined extract represents the 'XBRL' stream.

Stream 2: PDF Accounts (LLM-extracted)

For each PDF account filing, the pipeline (a) renders pages to JPEG images, (b) submits the image set to OpenAI's Batch API with a JSON-schema response format that prompts a structured extraction of line items, balance-sheet date, and CIC34 narratives, and (c) collects the JSON outputs to disk under `data/output/accounts-extractions/json-schema/{coyno}/{txn}.json`.

Stream Convergence: XBRL Format Conversion

The PDF JSON extractions are converted into the same column schema as the XBRL stream using an approved label mapping plus a Jaccard-similarity fallback for unmapped labels. Sign-aware columns (creditors, expenses) are normalised to positive magnitudes.

Merge and Provenance

The XBRL and PDF-converted CSVs are merged on `(company_number, financial_year)`. XBRL is preferred where both sources cover the same record. A `source_dataset` column records each row's provenance.

A2. Source Data

Companies House — XBRL Accounts

The official monthly bulk download of accounts in XBRL format is available at download.companieshouse.gov.uk/en_monthlyaccountsdata.html (https://download.companieshouse.gov.uk/en_monthlyaccountsdata.html). Historic archives are at historicmonthlyaccountsdata.html (https://download.companieshouse.gov.uk/historicmonthlyaccountsdata.html).

Companies House — PDF Accounts

For filings outside the XBRL stream (older accounts, CIC abridged filings, etc.), original PDFs are obtained via the Companies House document API and stored locally for extraction.

Companies House — CIC34 Forms

Community Interest Reports are filed as part of the same PDF accounts package for CICs and are extracted alongside the financial line items.

Spine

The project Spine (`data/input/spine/TSCS_spine.spine.csv`) is the deduplicated register of UK third-sector and civil-society organisations. It is the inclusion list for the dataset and the source of `uid` , `is_cic` , and `source_register` .

A3. Deduplication & Validation

Within-Source Deduplication

For each `(company_number, fy)` pair, the pipeline prefers current-year filings over prior-year companion rows. Same-source duplicate rows (e.g. balance-sheet vs P&L sections of the same filing) are coalesced by taking the first non-null value per column.

Cross-Source Deduplication

Where both XBRL and PDF sources cover the same `(company_number, fy)`, the XBRL row is preferred. The PDF row is dropped and not retained as a duplicate.

Date Validation

`fy` is normalised to integer-string. Where `fy` is blank but `balance_sheet_date` is populated, `fy` is derived (with a one-year offset for prior-year companion rows). Missing `fye` is imputed from the company's most common financial-year-end month-day.

Sign Normalisation

XBRL stores expenses and liabilities as positive magnitudes; PDF extractions sometimes preserve negative accounting signs. The pipeline applies `abs()` to nine sign-aware columns (`creditors_*`, `cost_sales`, `administrative_expenses`, etc.) so both sources use the same convention.

Spine Filtering

Only company numbers present in the Spine are retained. A left-join on `uid` populates `is_cic` and the derived `csotype`.

A4. Reproducibility

Code

The end-to-end pipeline lives in `code/companies-house/`:

- `pdf_accounts_extraction/openai-api/` — PDF-to-JSON extraction (rendering, batch submission, batch processing).
- `pdf_accounts_extraction/xbrl_format_conversion/` — JSON-to-XBRL column schema conversion, label mapping, dataset merge, and public-zip publishing.
- `reporting/guidance/` — this guidance document generator.

Dependencies

Managed via the uv (https://github.com/astral-sh/uv) package manager. Run `python code/companies-house/pdf_accounts_extraction/xbrl_format_conversion/dependencies.py` to install conversion-pipeline dependencies; analogous scripts cover the other modules.

Running the Pipeline

```
cd code/companies-house/pdf_accounts_extraction/xbrl_format_conversion
python run_pipeline.py stage3      # convert PDF JSON -> XBRL columns
python run_pipeline.py merge       # merge XBRL + PDF
python run_pipeline.py publish-tcss   # bundle + publish
```

Source Code

The release repository is at github.com/uk-third-sector-database/tso-database-builder (https://github.com/uk-third-sector-database/tso-database-builder).

Report generated: 2026-05-27 13:58:56

UK Third Sector Database — charitiesdata.org (https://charitiesdata.org)