CIC 36 Community Interest Statements

A Guide to the Dataset

March 2026

1. Overview

This dataset contains structured text extracted from 62,088 CIC 36 forms filed with Companies House between 2005 and 2025. 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.

Each record in the dataset captures what a Community Interest Company (CIC) pledged to do for its community at the point of incorporation. The CIC 36 form — the "Community Interest Statement" — requires applicants to describe who will benefit from the company's activities, what activities it will undertake, how those activities will benefit the community, and how any surplus will be used for community purposes.

The data was extracted using large language models (OpenAI gpt-4.1-mini) from scanned PDF forms, enabling researchers to analyse CIC community commitments at scale for the first time. Prior to this project, CIC 36 forms were only available as individual PDF documents on the Companies House website, making large-scale analysis impractical.

2. What are CIC 36 Forms?

Community Interest Companies (CICs) are a legal form created in 2005 specifically for social enterprises — businesses that trade for community benefit rather than private profit. Unlike traditional limited companies, CICs are subject to a "community interest test" and an "asset lock" that ensures their assets and profits are used for the benefit of the community they serve.

At incorporation, every CIC must file a Form CIC 36 "Community Interest Statement" with Companies House. This form requires the applicant to set out the company's community purpose in structured detail across three sections:

- **Section A: Beneficiaries** — a description of who the company's activities will benefit (e.g., "residents of the local community", "young people aged 16–25", "people with disabilities").
- **Section B: Activities and community benefits** — one or more activity/benefit pairs, where the applicant describes each activity the company will carry out and explains how that activity will benefit the community.
- **Section B also includes** a statement on how any surplus will be used for community benefit (e.g., "reinvested in community programmes", "distributed to other CICs").

An example CIC 36 form (https://assets.publishing.service.gov.uk/media/689db7a11fedc616bb1339a1/example-form-cic36.odt) is available from the CIC Regulator. These forms are publicly available through the Companies House filing system, but until now they have only been accessible as individual PDF documents. Many of these PDFs — particularly those filed before 2015 — are

scanned images rather than digitally generated documents. This dataset makes the content of these forms available as structured, machine-readable data for the first time.

3. Dataset Contents

62,088
CIC Incorporations

2005–2025
Registration Years

99.8%
Beneficiaries Extracted

85.2%
Surplus Use Present

Field Descriptions

Field	Description	Type
`uid`	Organisation identifier (format: GB-COH-{company_number})	Text
`company_number`	Companies House registration number	Text
`regy`	Year of incorporation	Numeric
`remy`	Year of dissolution	Numeric
`beneficiaries`	Section A: Who the company's activities will benefit	Text
`surplus_use`	Section B: How any surplus will be used for community benefit	Text
`activity_1..10`	Section B: Description of each activity (up to 10)	Text
`community_benefit_1..10`	Section B: How each activity benefits the community	Text

Sample Data

Company No.	Reg. Year	Beneficiaries	Activity 1	Community Benefit 1
06027161	2006	The general public within England and Wales	To organise multi-lingual health awareness conferences	Increased awareness of health services available to them
08077438	2012	Local people in the market town of Witham, Essex	To revitalise Witham town centre for local people	The community will benefit by an increased range of social and community activities within the town...
10993492	2017	Elderly people aged 65 and above, vulnerable people and marginalised groups living in the Great Dartmoor area	Supplying affordable and nutritious hot meals	The above community will benefit by the promotion of health and well-being, in particular through healthy eating
13342076	2021	Vulnerable adults affected by mental health, drugs, alcohol, and any kind of addiction or disposition	To provide education, training, and employment	By reducing the risk of harm, offence, addiction, and bringing balance to people's mental health and well-being
15791927	2024	Children and young people	Provision of healthy meals	Providing children and young people with healthy and nutritious meals within education settings

4. Coverage & Completeness

Beneficiaries	Activity 1	Community Benefit 1
99.8%	**98.9%**	**98.4%**

Surplus Use
85.2%

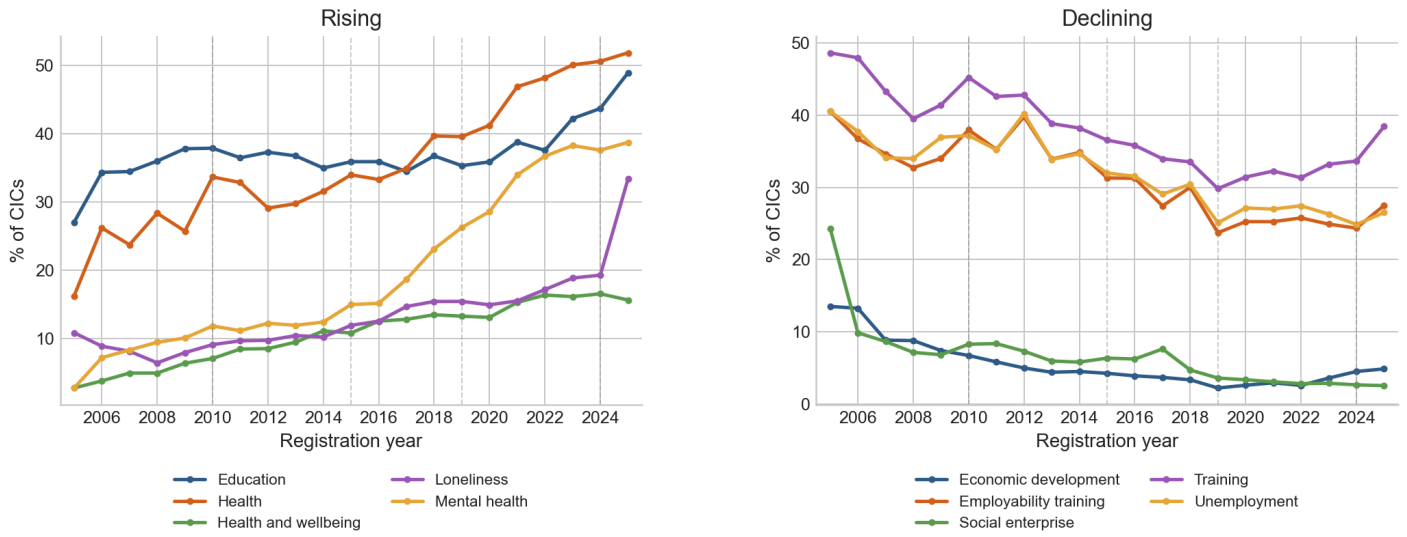
Most CICs report 2 activities (the median), with a mean of 2.3 activities per company and a maximum of 10. Nearly half of all CICs in the dataset report exactly 2 activity/benefit pairs. The number of activities tends to be slightly higher for CICs registered in more recent years.

> **Note:** 678 rows (1.1%) have no activities extracted. These are typically heavily degraded scans where the OCR could not recover sufficient text for the language model to identify any activity/benefit pairs. These records may still contain valid `beneficiaries` and `surplus_use` fields.

5. What Can You Learn?

The dataset enables large-scale analysis of what social enterprises in the UK set out to do at the point of incorporation. Using text classification techniques (such as topic modelling or keyword-based classification) on the activity descriptions, researchers can map how the CIC sector has evolved over two decades. The beneficiary descriptions allow analysis of who CICs intend to serve, while the surplus use statements reveal how social enterprises plan to reinvest their earnings.

Because each record is linked to a Companies House registration number, the dataset can be combined with other administrative data sources — including Companies House financial accounts, geographic data (registered office addresses), director records, and data from the CIC Regulator's annual reports — to build a comprehensive picture of the social enterprise sector.



Activity sectors with the largest increase (left) and decrease (right) in share of CIC registrations between 2005 and 2025. Dashed lines mark changes of government. Source: UK Third Sector Database, CIC 36 forms dataset.

The chart above shows how the activity composition of new CICs has shifted over time. Health-related activities have grown substantially as a share of new registrations, while education and training — once the dominant activity category — have declined. Researchers can investigate these patterns further by combining the dataset with Companies House geographic data and financial accounts.

Other possible analyses include: mapping beneficiary groups over time, examining geographic variation in CIC activity types, conducting survival analysis to understand which types of social enterprise are most durable, and comparing CIC community interest statements with the stated purposes of registered charities.

6. Limitations & Caveats

Incorporation only

CIC 36 forms are filed once at the point of incorporation. They capture what a company *planned* to do when it was established, not what it actually did or is currently doing. A company may have changed its activities significantly since incorporation, or may have ceased trading altogether. The dataset does not reflect any changes over time. However we have gathered this data through the annual CIC 34 form and made it publicly available: Nonprofit Financial Records (https://uk-third-sector-database.github.io/data/).

OCR quality

Many CIC 36 forms — particularly those filed before 2015 — are scanned images rather than digitally generated PDFs. The quality of optical character recognition (OCR) varies considerably across these documents. Some forms are heavily degraded, with poor print quality, skewed scans, or handwritten annotations that make text extraction difficult or unreliable.

Extraction accuracy

Data was extracted using large language models (OpenAI gpt-4.1-mini). While coverage is high (>98% for beneficiaries and activities), some extraction errors will exist. The language model may occasionally misattribute text between fields, merge or split activities incorrectly, or hallucinate content for heavily degraded documents. The `surplus_use` field has lower coverage (85.2%) primarily because this field is sometimes left blank on the original form, rather than due to extraction failures.

What is NOT in the data

The dataset does not include: financial information (turnover, assets, liabilities), director or officer details, current trading status, SIC codes, or whether the CIC is still active. These gaps can be addressed by using the wider data resources of the UK Third and Civil Society Sector Database (https://uk-third-sector-database.github.io/data/), which provides linked financial records, organisation registers, and procurement data for the same organisations.

7. Citation & Licence

Suggested Citation

McDonnell et al. (2026). *CIC 36 Community Interest Statements Dataset*. UK Third and Civil Society Sector Database. Available at: https://uk-third-sector-database.github.io/data/ (https://uk-third-sector-database.github.io/data/). Licensed under CC BY 4.0.

If you would like to learn more about this dataset and how it can be applied to your project or research programme, please contact research@brawdata.com (mailto:research@brawdata.com).

8. Changelog

Version	Date	Description
1.0	February 2026	Initial release: 62,088 CIC incorporations (2005–2025)

A1. Pipeline Architecture

The CIC 36 extraction pipeline transforms scanned PDF forms into structured, machine-readable data through five stages. Each stage is implemented as a separate Python module, allowing individual stages to be re-run independently as the pipeline is refined.

PDF Collection → OCR Page Detection → LLM Extraction → Quality Control → CSV Output

1. **PDF Collection** — CIC 36 forms are downloaded from the Companies House document filing system. Each CIC's incorporation documents are retrieved as multi-page PDF files.
2. **Page Detection** — Tesseract OCR scans each page of the PDF, identifying pages that contain CIC 36 form content using marker text detection. This reduces the input to the extraction model by approximately 91% in token count.
3. **Extraction** — The identified CIC 36 pages are submitted to the OpenAI batch API (gpt-4.1-mini), which extracts structured JSON from the form content using a defined schema with few-shot examples.
4. **Quality Control** — Automated heuristics detect hallucinated or low-quality extractions. Flagged records are re-submitted for extraction. A subset of results is compared against independent Gemini (Google) extractions as a ground truth check.
5. **Post-Processing** — Variable-length JSON extractions are reshaped into wide-format CSV (one row per CIC, with activity_1..10 and community_benefit_1..10 columns). Company numbers are joined with the project's organisation spine to add standardised UIDs.

A2. Page Detection

CIC incorporation documents filed with Companies House are typically multi-page PDFs that bundle several forms together — the memorandum and articles of association, Form IN01 (application for incorporation), and the CIC 36 community interest statement. Only one or two pages out of a document that may contain 20–50 pages are relevant for extraction. Sending entire documents to the language model would be prohibitively expensive and would introduce noise.

The page detection module uses Tesseract OCR to scan each page of every PDF and identify those containing CIC 36 form content. It searches for marker text patterns including “community interest statement”, “cic 36”, and “community benefit statement”. Pages matching any of these markers are extracted as separate images for submission to the language model. This filtering step reduces the total input by approximately 91% in token count, from an estimated 2.8 billion tokens to around 250 million tokens.

The page detection logic is implemented in `ocr_page_detector.py` and runs as a batch process across all downloaded PDFs. Pages that do not match any marker text are discarded. For documents where no CIC 36 pages are detected, the record is flagged for manual review.

A3. Extraction Model

The extraction stage uses OpenAI’s gpt-4.1-mini model, chosen for its combination of cost-effectiveness and high accuracy on structured extraction tasks. All extractions are submitted through the OpenAI Batch API, which provides a 50% cost reduction compared to synchronous API calls at the cost of higher latency (results are typically available within 24 hours).

The model extracts three top-level fields from each CIC 36 form using a structured output schema:

- **`beneficiaries`** (string or null) — the content of Section A, describing who will benefit from the company’s activities.
- **`activities`** (array of objects) — each object contains an `activity` field and a `community_benefit` field, capturing the activity/benefit pairs from Section B.
- **`surplus_use`** (string or null) — the statement on how surplus will be used for community benefit.

The extraction prompt (version 4) includes few-shot examples selected to cover challenging cases: forms with handwritten text, heavily degraded scans, and unusual layouts. For scanned documents, a two-pass pipeline is used: Pass 1 converts the page images to markdown text using the vision API, and Pass 2 extracts the structured fields from the markdown representation.

A4. Quality Validation

Quality validation uses a combination of automated heuristics and ground truth comparison to identify and correct extraction errors.

Ground truth comparison: A random subset of extractions was independently processed using Google’s Gemini model. The two sets of extractions were compared at the field level. Coverage thresholds were set at 98% for `beneficiaries` and `activities` fields, and 93% for `surplus_use` (which has a higher legitimate null rate because many applicants leave this section blank on the original form).

Hallucination detection: Automated heuristics scan the extracted text for suspicious patterns that suggest the language model generated plausible but incorrect content. These patterns include generic placeholder text (e.g., "the company will carry out activities for the benefit of the community" without specific details), repeated verbatim phrases across unrelated companies, and extractions that are implausibly long or short relative to the source document.

Re-extraction: Records identified as likely hallucinations or low-quality extractions were re-submitted to the language model with adjusted prompts. In cases where re-extraction did not improve quality, the record was flagged and the extraction was retained but marked as lower confidence.

A5. Post-Processing

Wide-format reshape: Each JSON extraction contains a variable-length `activities` array (1–10 elements). The post-processing step converts these into a fixed wide-format structure with columns `activity_1` through `activity_10` and `community_benefit_1` through `community_benefit_10`. Companies with fewer than 10 activities have null values in the unused columns.

Spine UID join: Each company number is matched against the project's master organisation spine — a cross-register lookup table that assigns standardised unique identifiers (UIDs) in the format `GB-COH-{company_number}`. This allows the CIC 36 dataset to be linked to other datasets in the UK Third Sector Database ecosystem using a consistent identifier scheme.

Output: The final output is a single CSV file with one row per CIC and columns for all extracted fields. The file is encoded as UTF-8 with standard comma delimiters. Missing values are represented as empty strings.

A6. Reproducibility

The extraction pipeline code is available in the project's public repository at github.com/uk-third-sector-database/tso-database-builder (https://github.com/uk-third-sector-database/tso-database-builder), with the CIC 36 pipeline code located in the `openai-api` subdirectory.

The pipeline requires Python 3.12 or later, with dependencies managed using uv (https://docs.astral.sh/uv/). Key dependencies include:

- `openai` — for API access to gpt-4.1-mini
- `PyMuPDF` (fitz) — for PDF page rendering
- `pytesseract` — for OCR-based page detection
- `python-dotenv` — for environment variable management

All paths, model settings, prompt versions, and quality thresholds are defined in a central `config.py` file. Environment-specific settings (such as the OpenAI API key) are loaded from a `.env` file that is not committed to version control.

> **Note:** Running the full pipeline requires an OpenAI API key and access to the source PDF files downloaded from Companies House. The PDF files are not included in the repository due to their size (approximately 180 GB). The final extracted CSV dataset is available for download from the project website.

Document generated: March 2026

UK Third and Civil Society Sector Database — uk-third-sector-database.github.io (https://uk-third-sector-database.github.io/)