

Working Paper no. **196**

**Mapping the Italian third sector: insights from web-augmented
administrative data**

Carlo Bottai
University of Milano-Bicocca

Francesco Trentini
University of Milano-Bicocca, CRISP and LABORatorio R. Revelli

November, 2025

Mapping the Italian Third Sector: Insights from Web-augmented Administrative Data ^{*†}

Carlo Bottai[‡]

Francesco Trentini[§]

November 2025

Abstract

Reliable data on the Italian Third Sector remain limited despite its socio-economic importance. This paper presents a methodology for assembling a comprehensive dataset on the sector by systematically exploiting administrative data from the Single National Register of the Third Sector (RUNTS). The proposed pipeline automates the acquisition and cleaning of data on more than 134,000 organisations through a tailored web-scraping procedure. To address common shortcomings of administrative data, such as incompleteness and self-reporting bias, a novel web-augmentation strategy is implemented, combining information extracted from organisational websites with a Generative AI-based classification system. This approach allows the identification of beneficiaries, territorial scope, and economic activities, and enables comparison between web-derived and self-declared information for the latter case. Using data for the second quarter of 2025, the paper provides an up-to-date picture of the Italian Third Sector and offers a new informational base for socio-economic research and co-programming of local policies.

Keywords: Social Economy, Third Sector, Administrative Data, Web-Augmentation, Generative AI, ICNPO.

JEL Codes: C81, C83, L31.

*We thank Anna Velyka for her contributions to the development of the classification pipeline. We are grateful to the participants of the 6th and 7th International Conferences on Advanced Research Methods and Analytics (CARMA 2024 and CARMA 2025), the 15th Annual Conference on Statistics at Istat, the Renaissance in Economics 2nd International Conference, and the participants of the DEFAP 2025 Seminar Series at the Department of Economics and Finance, University of Bari, and the LEA seminar – Multidisciplinary Territorial Policy Lab, DEMDI, University of Bari for their valuable comments and insights.

†We kindly acknowledge the University of Milano-Bicocca for financial support within the framework of the Bicocca Starting Grant 2023, ATESP2024-0007 “Big Data Platform per il Terzo Settore.”

‡Department of Economics, Management and Statistics, University of Milano-Bicocca. Email: carlo.bottai@unimib.it.

§Department of Statistics and Quantitative Methods, University of Milano-Bicocca; Interuniversity Research Center For Public Services (CRISP); Laboratorio “R. Revelli”. Email: francesco.trentini@unimib.it.

1 Introduction

The Social Economy, also known as Social and Solidarity Economy, is that segment of the economy that encompasses those organisations whose primary goal is to generate a positive social impact (Salamon and Sokolowski, 2016; European Commission, 2020; OECD, 2023; Bassi, 2025). Although the specific definition varies with the national legislative context, it is possible to trace common principles and features among Social Economy organisations in Europe (Monzon and Chaves, 2008; European Commission, 2021a). First, social and environmental objectives takes precedence over profit. Second, profits and surpluses are retained and reinvested within the organisation¹. Third, these organisations act in the collective or general interest and, fourth, they must have a democratic and participatory governance. The main legal forms that these organisations take up are co-operatives, mutual benefit societies, associations, and charities and foundations². They are active in a wide range of sectors, such as social assistance and healthcare, culture and education, finance and insurance, logistics, and manufacturing, providing goods and services. They may deliver goods and services either through market transactions or on a non-market basis, either in the form of in-kind contributions or monetary transfers. Also in light of the peculiar characteristics of these organisations, in the last decade, the European Commission (EC) has taken steps, within the framework of the ‘proximity and social economy ecosystem’, to support this economic model, facilitating access to financial markets and promoting a coherent regulatory framework, also spurring enterprise integration across Member States (European Commission, 2021b; European Commission et al., 2024)—an overview of the EU-level Social Economy policy is provided in the Appendix A. The Social Economy, in Europe in 2021, represents more than 4.3 million organisations and employs more than 11.5 million people—i.e., six percent of the EU employment—, with 912 billion euros in turnover (European Commission et al., 2024).

Comparably with other EU economies, the Social Economy is a primary actor in the Italian socio-economic system. According to the latest figures, Social Economy organisations constitute a complex and multifaceted aggregate of 406,709 organisations, employing more than 1.5 million workers, and activating more than 4.6 million volunteers (European Commission et al., 2024). The vast majority of the organisation consists of associations, representing 76.5% of the total organisations—311,423 units. Co-operatives account for 13.5% of the organisations—55,237 units—, of which 27.2% are social co-operatives—15,067 units. Despite the numerical prevalence of associations, co-operatives are the primary contributors to the Social Economy’s overall turnover (128,931.07 million euros) and employment (1,130,741 employees). Finally, foundations are an important, but numerically residual, actor—8,39 units.

In historical perspective, the Italian Social Economy must be understood as the sum of two components: mutual co-operatives on the one hand, and associations, social co-operatives, and

¹While in some cases the distribution of profits is allowed, there are strong limitations on the degree to which this can be done. Details are provided in Section 2.

²In all official documents the EC refers to organisations active in the social economy as ‘social enterprises’, independently of their legal form and characteristics.

social enterprises on the other (Mancino and Thomas, 2005; Ianes, 2016; Borzaga et al., 2023)³. In 2017, the second component was unified by the legislator in the so-called Third Sector through the Legislative Decree 117/2017⁴, known as “Codice del Terzo Settore” (CTS). Consequently, the Third Sector represents only a segment of the broader Italian Social Economy and complements the mutual co-operative component — which dominates in terms of employment and value added (Istat and Euricse, 2021)⁵. In order to qualify as an ETS, each organisation must be registered with the newly-constituted Registro Unico Nazionale del Terzo Settore (RUNTS), the first single national register of the Italian Third Sector. Managed by the Italian Ministry of Labour and Social Policies (MLPS), this register replaced hundreds of local, regional, and national registers. Moreover, with the aim of transparency and accountability of the ETS listed in the register, the RUNTS has been publicly accessible, via a web platform, since December 2023 (Italian Ministry of Labour and Social Policies and Italian Union of Chambers of Commerce, 2024).

Nevertheless, despite the economic and social relevance of the Third Sector, available evidence is still scarce and despite the public availability of the RUNTS, the difficulty of access has impinged on the adoption of this source for research purposes. The RUNTS has been used in institutional research (Italian Ministry of Labour and Social Policies and Italian Union of Chambers of Commerce, 2024; INPS, 2025), research by foundations aimed at characterising the general characteristics of the sector and, in parallel, assess the degree of implementation of the reform (Fondazione Terzjus, 2022, 2023, 2024; Generali and Fondazione Cattolica, 2024) and in the academic context in a single attempt to build an indicator of territorial proximity of ETS to the population in Rome (Angotzi et al., 2025).

In this work, we describe a pipeline for collecting and validating the RUNTS administrative data and produce a picture of the Italian Third Sector in the second quarter of 2025. From the methodological point of view, we propose a procedure to produce data suitable for scientific research. We also

³It is possible to trace a distinction between two alternative models of Social Economy in Europe. On one hand, there is the Anglo-Saxon model which envisages a charitable model primarily composed of associations and foundations, where the organisation's activities are driven by the non-profit motivation, and any surpluses must be automatically allocated to its statutory activities. On the other, there is the Francophone model, primarily composed of mutual benefit societies, co-operative enterprises, and associations, which are viewed as economic agents operating in the market but according to solidarity principles, thereby allowing for some profit to be redistributed among the enterprise's 'collective membership' (Mertens, 1999; Moulaert and Ailenei, 2005). Although Italy was historically closer to the Francophone model, as evident from the large number and socio-political importance of its co-operatives, today—in line with the rest of continental Europe—the country is experiencing a shift toward a mixed model where both traditions coexist, contaminating each other, and even giving birth to hybrid organisational forms (Borzaga et al., 2014).

⁴Decreto Legislativo 3 luglio 2017, n. 117 (Codice del Terzo settore), Pubblicato nella Gazzetta Ufficiale Serie Generale n. 179 del 02 Agosto 2017 – Supplemento Ordinario n. 43. <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2017-07-03;117!vig=>

⁵At the same time, there are concurrent classifications that are related but also differentiate from the concepts of Social Economy and Third Sector, among which the concept of Non-Profit Institutions (NPIs) stands out, in particular in the institutional work of the Italian National Institute of Statistics (Istat). Istat definition of NPIs revolves around the prohibition of profit distribution only, while there are no constraints on the type of interests that they can pursue and is therefore different from the definition of ETS. NPIs include a variety of organisations, such as associations, cooperatives and foundations but also Non Governmental Organisations, trade unions, political parties and some religious institutions. Istat produces a yearly census of NPIs that collects information on a variety of dimensions, such as revenues and expenditures for 2021 and the related budget items, type of market or non-market economic activity, the main type of funding (public or private), fundraising initiatives, the use of resources (Istat, 2025; Stoppiello et al., 2025). The latest figures show that in Italy 368,367 organisations are active, directly employing 949,200 workers in 2023 (Istat, 2025) and engaging 4.6 million volunteers in 2022 (Istat, 2024). The literature on the Non-Profit sector is quite extensive. Among recent contributions, see Della Queva et al. (2023) for an appraisal on the relevance of NPIs to local development in Southern Italy and peripheral areas and Lori (2024); Lori and Zandonai (2020) for an interesting discussion on the relation between Non-Profit Sector and Third Sector.

show that, despite the wealth of information in many domains, the administrative data is incomplete mainly due to the non-compulsory nature of the information provided by the ETS and is subject to self-reporting bias. To overcome these issues, we propose a strategy to web-augment the available information and classify, with a Generative AI (GenAI) based model, the economic activities carried out by each organisation, according to the International Classification of Nonprofit Organisations (ICNPO) taxonomy ([Salamon and Anheier, 1992, 1994, 1996](#)). These techniques align with practices in official statistics for using GenAI models to process available textual information ([Kwon et al., 2024; Ortame et al., 2025](#)). Moreover, our pipeline integrates the administrative information with the systematic extraction and analysis of online information. This process directly addresses the key methodological challenge of using Generative AI to produce structured datasets from unstructured, web data for socio-economic research ([Devetak and Mandel, 2023](#)).

The rest of the paper is organised as follows. Section 2 provides an essential historical and legal overview of the Italian Third Sector. Section 3 then illustrates the strengths and weaknesses of administrative data for research and details the data collection and cleaning pipeline, including the methodology for web-scraping the RUNTS register. Finally, the section closes by presenting the initial descriptive results on the Italian Third Sector as of the second quarter of 2025. Section 4 presents the novel web-augmentation strategy and the Generative AI expert system used for ICNPO classification and closes with an illustration of the initial results. Lastly, Section 5 concludes by summarising our contributions and discussing the broader implications for administrative data research.

2 The Italian Third Sector

The introduction of the CTS in 2017 profoundly reorganised the legislation on the Third sector under a unified framework. First, it identified a set of organisation types eligible to be part of the Italian Third Sector and formally qualified to be designated as Third Sector Organisation (Ente del Terzo Settore, ETS). The reform also introduced a set of activities of general interest (Art. 5 CTS), covering the areas of social and health services, education, culture and recreation, civil rights and cooperation, environment and community development, charity and economic solidarity, and emergency services (for the detailed taxonomy, see Table D.1 in the Appendix). Each organisation must pursue at least one of these objectives to use the ETS designation. In addition, each ETS must satisfy specific requirements regarding balance sheets and social and sustainability reports that must be produced for accountability reasons, depending on its legal form and size ([Italian Ministry of Labour and Social Policies and Italian Union of Chambers of Commerce, 2024](#)).

The seven types of organisation listed in the CTS are organised in a corresponding number of Sections of the register which includes Organizzazioni di Volontariato (OdV), Associazioni di Promozione Sociale (APS), Enti Filantropici (EF), Imprese Sociali (IS), Società di Mutuo Soccorso (SoMS), Altri Ente del Terzo Settore (AE) and Reti Associative (RA). Each of them has specific legal requirements regarding its governance and the composition of its members and workforce—see Table 1 for an overview. The OdV are typical expressions of civic solidarity and altruistic organisations. Their primary function is to provide services to the general public or disadvantaged groups, acting as a direct extension of the welfare state in areas such as emergency services, social assistance, and

civil protection. They work with a prevalence of volunteer labour, which must be the dominant component of their workforce. Funding is heavily dependent on donations, public grants, and fundraising events. Despite the large degree of similarity to the OdV, the APS are characterised by a mutualistic or community-building focus, which benefits their members as well as the wider public, fostering social cohesion and community engagement. Although they rely on volunteers, they have less stringent limitations regarding paid workers. The EF, typically structured as foundations or associations, are the primary capital-movers of the Third Sector. Their main role is not direct service provision, but rather enabling the work of others through funding and coordination. Their funding model is largely based on income from endowments and donations, and it requires a professional (paid) workforce to manage it with the aim of channelling resources to the rest of the Third Sector. The IS, including the social co-operatives, are the most business-oriented organisations within the Third Sector, combining to achieve social impact through commercial activity. They operate as a traditional business with paid staff and are allowed for limited profit distribution with strict caps, a unique feature among ETS⁶. The SoMS are mutual societies, that provide insurances, healthcare, and pensions for their members and families. Instead, the AE are a residual category for any organisation that meets the general requirements of an ETS but does not fit into one of the more specific legal forms. It can host innovative organisations that do not fit the other types provided by the CTS. Finally, the law also introduced the category of RA, which are second-level organisations, meaning that their members are not individuals, but other ETS, which they serve by acting as hubs for coordination, representation, and support. The RA may opt for obtaining a second qualification by registering in another RUNTS section, meaning that they can either operate as pure network coordinators or run initiatives as an APS, OdV or any other type of ETS.

The ETS designation provides the organisation with a range of significant advantages. Primarily, it facilitates access to crucial funding mechanisms, including the 5×1000 donations, exclusive public procurement, and subsidised credit at both national and European levels. Moreover, registered organisations benefit from substantial fiscal and tax advantages, such as exemptions or reductions on stamp duty, registration, local taxes like the Italian municipality property tax (IMU), and the ability to apply subsidised tax regimes for commercial activities. Furthermore, registration entitles organisations to collaborate with public bodies, offering a preferential path in administrative procedures and the possibility to enter into agreements with the public administration for activities of general interest. Finally, the process simplifies the procedure for obtaining legal personality while promoting greater transparency and visibility toward members, citizens, and institutions.

3 The RUNTS as Administrative Data Source for Statistical Use

Administrative data have been widely adopted in economic research. Labour market and healthcare data are the most widely used in both institutional (Berton and Paruolo, 2026) and academic contexts (e.g., Ardito et al., 2023, 2025). Administrative data are a specific class of big data (Connelly et al., 2016; Leombruni, 2024), with whom they share the advantages of high volume and high frequency, they usually have complete coverage of defined populations, measurement-error is limited, the

⁶Information on IS is obtained from the business register, with which registration is mandatory. Nevertheless, they are also required to submit further information to the RUNTS, such as the general-interest activities they undertake and the number of volunteers (Italian Ministry of Labour and Social Policies and Italian Union of Chambers of Commerce, 2024)

Table 1: ETS types and main characteristics as of D.Lgs. 117/2017, Codice del Terzo Settore (CTS)

Organisation Type	Acronym	Economic Function	Target Beneficiaries	Governance and Capital	Key Funding Sources	Labour Input Mix
Organizzazioni di Volontariato	OdV	Provision of solidarity and social services to third parties	General public and disadvantaged groups	Membership-based; at least majority of members must be volunteers	Fundraising, donations, contributions from members, public or private grants	Predominantly volunteer-based; paid workers $\leq 50\%$ of volunteers (Art. 33 CTS)
Associazioni di Promozione Sociale	APS	Mutual and general interest activities	Members, their families, and the general public	Democratic and membership-based; open participation	Membership fees, donations, fundraising, and income from institutional or secondary activities	Volunteer prevalence; paid workers $\leq 50\%$ of volunteers or $\leq 20\%$ of members (Art. 36 CTS)
Enti Filantropici	EF	Grant-making and philanthropic support	Disadvantaged individuals and other ETS	Foundation or association form; governance ensures transparency and purpose consistency	Donations, endowments, bequests, and fundraising (prevalent)	No specific ratio required; may use paid staff for operational tasks
Imprese Sociali	IS	Market-based production of goods and services with social impact	Community or disadvantaged groups	Any private legal form; profit distribution allowed within statutory limits	Revenues from market activities (prevalent), plus public/private contributions	Primarily paid employees; limited volunteer involvement possible
Società di Mutuo Soccorso	SoMS	Mutual aid, welfare and insurance-type services	Members and their families	Membership-based and non-profit; democratic participation of members	Membership fees (prevalent), returns from collective funds and assets	Mainly member-driven; may employ staff for administrative or technical support
Altri ETS	AE	Flexible pursuit of civic, solidaristic and social utility aims	Unrestricted (depending on statutory purpose)	Flexible legal form; must comply with non-profit and participation principles	Any lawful source (donations, fees, activities); no prevalence required	No specific ratio; balance between volunteers and professionals based on activities
Reti Associate	RA	Coordination, representation, and support for member ETS	Member ETSs and their networks	Second-level organisations; at least 100 ETS or 20 foundations in 5+ regions	Membership fees, contributions from affiliates, and project-based funding	No specific ratio; staff proportional to coordination needs

costs of collection are paid by the primary users, and they entail the possibility to cover populations that are hard to reach through surveys. Nonetheless, they also have inherent limitations which make the data generating process very complex and not linear, such as the lack of documentation of data input and processing procedures, lack of metadata, the adoption of idiosyncratic non-standardised taxonomies, which also complicates external data linkage, and a large user base with heterogeneous proficiency in the use of the data input procedure (Harron et al., 2017). Given their advantages in comparison to traditional data sources (mostly surveys), administrative data have also been widely adopted in official statistics, and national statistical offices are increasingly replacing survey data with administrative data as input in the production of official statistics (Di Bella, 2022). However, their integration requires new statistical techniques compared to the usual ones, as well as new competences and processes in national statistical institutes (Pfeffermann et al., 2015).

Among other things, administrative systems prioritise immediate operational or legal mandates. Therefore, their focus is on efficient transaction processing rather than accuracy (Italian Ministry of Labour and Social Policies and Italian Union of Chambers of Commerce, 2024). This can manifest itself as missing or incomplete records, data entry errors—by clerks or users—, or inconsistencies that are often challenging to detect and rectify. Errors may be at random, or even systematic—reflecting a flawed administrative process. Furthermore, administrative data often lack information on quality and paradata typically collected in surveys, like response rates on non-compulsory form fields. Our methodological pipeline is specifically designed to address these inherent issues in processing raw registry data into datasets ready for statistical use (Jones et al., 2019).

3.1 Scraping of the Administrative Data

As said in the Introduction, the main goal of this paper is to describe a pipeline for collecting and validating the RUNTS administrative data, and produce data suitable for scientific research. Therefore, the obvious first step of this pipeline is the collection of the raw data from the administrative register—the interested reader can find an overview of the whole pipeline in the Appendix C.

Through the RUNTS website⁷, the MLPS provides a constantly updated spreadsheet containing, for each ETS, its tax code (codice fiscale), business name, legal type, the municipality of the headquarters, the date of enrolment in the register and the general details of the legal representative⁸.

However, the register is much richer in terms of information about each individual ETS, including details about the number of workers and volunteers activated by the organisation, the economic activity carried out by the organisation, and the composition of the bodies that govern the organisation.

Nevertheless, all these pieces of information can be accessed only through a non-machine-friendly web platform⁹, for which no web application programming interface (web API) exists. To automatically access and systematically collect these data, we developed a scraper that, moving from the list of organisations in the spreadsheet, simulates human behaviour to systematically explore the RUNTS platform to search for such additional details. The scraper queries the web platform about a specific tax code, and if any results are returned, opens the ‘ETS details’ page, in which

⁷<https://servizi.lavoro.gov.it/runts>

⁸<https://servizi.lavoro.gov.it/runts/it-it/Lista-enti>

⁹<https://servizi.lavoro.gov.it/runts/it-it/Ricerca-enti>

all the additional information about that specific organisation can be found. Lastly, it collects the information and stores it in a JSON-like format. The data are subsequently structured in a PostgreSQL database to facilitate their accessibility.

A primary obstacle to a systematic collection of the information is that most of the information is provided by the organisations on a voluntary basis. Most fields are not mandatory, free-form text, and are neither validated at the input stage nor post-processed ex-post. Therefore, the scraper cannot exploit a predefined schema and some taxonomies to guide its search for the information contained on a ‘ETS details’ page. Rather, it must collect as many information pieces as possible and try and restructure them ex-post, searching for similarities among the organisations.

Particularly challenging is, then, the fact that the RUNTS web platform is not easy to read by a machine (Berners-Lee and Fischetti, 1999; Berners-Lee et al., 2001; Shadbolt et al., 2006). In the HTML code of the ‘ETS details’ web page, meta-data are provided only partly, and, even more, some important parts of the meaning can only be understood by the human reader by looking at the rendering and typesetting of the web page. To provide examples, the main details of the ETS are presented in a dictionary-like manner, which corresponds to two HTML tags: one for the key of the dictionary and the other for the corresponding value—see Figure 1, lines 10–15. However, the web page is not consistent with this pattern. Other pieces of information are provided as strings, with no dictionary-like structure and, even worse, within one HTML tag of class `ente_testo-1g`; that is the same as that of the ‘value’ element of the previous example—see Figure 1, lines 28–30. This would result in a machine reading the HTML code of this page being ‘confused’ about the real meaning and role of the elements of class `ente_testo-1g`. And this is just one of the several examples that we may provide.

Other portions of the page are even more challenging to scrape. For example, the information about the members of governing bodies of the organisation is presented in two columns, one for the information about the person and the other with that for her or his positions performed in the organisation. Each piece of information is provided in a dictionary-like structure, but this time the key can be deduced only by looking at the textual content of the first of the two HTML tags and no meta-data is clearly provided—see Figure 2, lines 6–7.

These examples point to a more general consideration regarding the accessibility and availability of administrative data for research purposes. The intention of the legislator in declaring public and free access to the data¹⁰ must be understood as directed to for purposes more limited than a systematic exploration, as opposed to what is necessary for scientific research purposes.

Notwithstanding these challenges, the scraper is able to collect (whenever available) information about each ETS, including but not limited to:

Business profile: Business name, tax code, incorporation date, enrolment date, section of the register, legal form, certified e-mail address, and website address.

Location Full address of the registered office.

Economic Activity: Eligibility for the 5×1000 tax declaration, and the economic activity carried out,

¹⁰“Il sito del RUNTS è pubblico. [...] L’accesso alla consultazione delle informazioni e degli atti presenti nel RUNTS è libero [The RUNTS website is public. [...] Access to the consultation of the information and documents contained in the RUNTS is free.]” (Decreto Ministeriale 106/2020).

Figure 1: Example of a portion of a RUNTS portal's page that contains information about the business profile of an ETS (top part) and the corresponding underlying HTML code (bottom part).

 Ministero del Lavoro e delle Politiche Sociali

IT ▼

RUNTS

Registro Unico
Nazionale
Terzo Settore

Scopri il RUNTS

News

Normativa

Manuale utente

FAQ

Accedi al registro

Lista enti

Ricerca enti

Ricerca enti > L'ASSOCIAZIONE APS

L'ASSOCIAZIONE APS

Repertorio: 99999 Codice fiscale: 90909090909

Iscritto il 01/01/2024

Iscritto nella sezione in data 01/01/2024

Dati ente

Sezione ASSOCIAZIONI DI PROMOZIONE SOCIALE	Forma Giuridica ASSOCIAZIONE Email PEC LASSOCIAZIONE@PEC.IT	Atto costitutivo 01/01/2024 Ultimo aggiornamento statuario 01/06/2024
Iscrizione come rete associativa No	Sito internet WWW.LASSOCIAZIONE.NET	
Codice fiscale 90909090909		

```
<div class="row">
  <div class="col-sm-12 col-md-12 col-lg-12">
    <span id="dnn_ctr448_View_spnDenominazione"
      class="ente_titolo-xl ente_bold">
      L'ASSOCIAZIONE APS
    </span>
  </div>
</div>
<div class="row">
  <div class="col-sm-12 col-md-12 col-lg-2">
    <span class="ente_titolo-lg ente_bold">
      Repertorio:
    </span>
    <span id="dnn_ctr448_View_spnRepertorio" class="ente_testo">
      99999
    </span>
  </div>
  <div class="col-sm-12 col-md-12 col-lg-10">
    <span class="ente_titolo-lg ente_bold">
      Codice fiscale:
    </span>
    <span id="dnn_ctr448_View_spnCodiceFiscale" class="ente_testo-lg">
      90909090909
    </span>
  </div>
</div>
</div>
<div class="row">
  <div class="col-sm-12 col-md-12 col-lg-12">
    <div id="dnn_ctr448_View_spnIscrittoIl"
      class="ente_testo-lg">
      Iscritto il 01/01/2024
    </div>
    <div id="dnn_ctr448_View_spnIscrittoIlRete"
      class="ente_testo-lg">
    </div>
    <div id="dnn_ctr448_View_spnIscrittoIlSezione"
      class="ente_testo-lg">
      Iscritto nella sezione in data 01/01/2024
    </div>
  </div>
</div>
```

Figure 2: Example of a portion of a RUNTS portal's page that contains information about the governing bodies members of an ETS (top part) and the corresponding underlying HTML code (bottom part).

Persona 1	Carica 1
Tipo Persona Fisica	Carica LEGALE
Rappresentante legale Sì	RAPPRESENTANTE
Codice fiscale ABCABC12Z89X456Y	Data nomina 01/01/2023
Nome MARIO	
Cognome ROSSI	
Stato	
Data di nascita 01/01/1980	
Provincia RM	
Comune ROMA	


```

1  <!-- Colonna 1 -->
2  <div class="ente_grid3Columns">
3    <div class="ente_grid3Columns-column">
4      <span class="ente_testo-lg blu ente_bold">Persona 1</span>
5      <div>
6        <span class="ente_testo-lg ente_bold">Tipo</span>
7        <span class="ente_testo-lg">Persona Fisica</span>
8      </div>
9      <div>
10       <span class="ente_testo-lg ente_bold">Rappresentante legale</span>
11       <span class="ente_testo-lg">Sì</span>
12     </div>
13     <div>
14       <span class="ente_testo-lg ente_bold">Codice fiscale</span>
15       <span class="ente_testo-lg">ABCABC12Z89X456Y</span>
16     </div>
17     <div>
18       <span class="ente_testo-lg ente_bold">Nome</span>
19       <span class="ente_testo-lg">MARIO</span>
20     </div>
21     <div>
22       <span class="ente_testo-lg ente_bold">Cognome</span>
23       <span class="ente_testo-lg">ROSSI</span>
24     </div>
25   </div>
26
27   [...]
28
29   <div>
30     <span class="ente_testo-lg ente_bold">Comune</span>
31     <span class="ente_testo-lg">ROMA</span>
32   </div>
33 </div>
34 <!-- Colonna 2 -->
35 <div class="ente_grid3Columns-column">
36   <span class="ente_testo-lg blu ente_bold">Carica 1</span>
37   <div>
38     <span class="ente_testo-lg ente_bold">Carica</span>
39     <span class="ente_testo-lg">LEGALE RAPPRESENTANTE</span>
40   </div>
41   <div>
42     <span class="ente_testo-lg ente_bold">Data nomina</span>
43     <span class="ente_testo-lg">01/01/2023</span>
44   </div>
45   <br>
46 </div>
47 </div>

```

classified both according to Art. 5 CTS (see Table D.1 in the Appendix) and the International Classification of Nonprofit Organisations (Table D.2).

Governing Bodies: Full names of natural persons who hold positions in the governing bodies, comprising the legal representative(s), for whom additional information is provided (including their tax code). For each person, a list of the positions performed is provided.¹¹ List of governing bodies, each with its appointment date and size.

Organisational Capacity: Number of members, paid workers, and volunteers.

Documents: Date of incorporation, statute, annual financial statements, and corporate sustainability and responsibility statements, among others.

3.2 The Italian Third Sector Through the Lens of the RUNTS

Using data collected in the second quarter of 2025, we are now able to provide a comprehensive picture of the Italian Third Sector as emerging from the RUNTS¹².

The quantitative analysis of the RUNTS's seven organisation types reveals a heterogeneous composition of the sector (Table 2). APS constitute the largest category and, with 62,584 organisations, represent 46.42% of the total organisations. They are followed by OdV, which, with 38,237 organisations, make up 28.36%. Combined, these two categories represent nearly three-quarters of all Italian ETS, confirming the central role of volunteerism and civic association in the country (Zamagni et al., 2004; Borzaga and Ianes, 2006; Burroni et al., 2017; Istat, 2025). An interesting finding is the substantial presence of IS among the Italian ETS. With 22,520 organisations (16.70%), their presence highlights the consolidated market-oriented and productive dimension of the Italian Third Sector. The remaining typologies are numerically smaller, highlighting their specialised nature and peculiar functioning.

Therefore, an interesting picture of the sector emerges. The Italian Third Sector is characterised by a major associative base and a number of specialised organisations that complete the ecosystem by providing services and financing to the rest of the sector, with the addition of a robust segment of market-oriented companies¹³. Looking at the year of establishment of the ETS, their average age is 17 years—year of establishment 2008 (see Table 3). However, this figure shows significant variability. All types of ETS, with the exception of RA, include at least one organisation founded before 1900.

¹¹As interesting as they are, some of the details included in the administrative register are particularly sensitive, and their collection goes beyond the legitimate interests for which they were collected by the public administration. Therefore, we designed the scraper to compromise between two necessities: to preserve as much information as possible and protect information privacy. To provide an example, the sex, and year and county of birth of the legal representative(s) of each organisation is extracted from their tax code (exploiting a peculiarity of the Italian case). Those details are then stored in our database—being interesting for scientific analysis purposes—, while the full tax code has been pseudonymised through a hash function to prevent the identification of individuals.

¹²The data required an extensive cleaning and validation phase due to numerous issues, including significant misreporting. For instance, we identified tax codes that failed validation with the Italian Revenue Agency (Agenzia delle Entrate) and temporal inconsistencies, such as statutory updates dated prior to the organization's establishment or set in the future. Additionally, some organisations declared ICNPO activities, like trade union or political activities, that are forbidden under the CTS legislative framework.

¹³It is important to note that the picture should be complemented by Foundations of banking origin and Savings Banks that also play an important role in financing Third Sector activities.

Table 2: Number of ETS by organisation type. 2025Q2.

Organisation type	N	N%
Ass. di Promozione Soc. (APS)	62,584	46.42%
Org. di Volontariato (OdV)	38,237	28.36%
Imprese Sociali (IS)	22,520	16.70%
Altri ETS (AE)	11,186	8.30%
Enti Filantropici (EF)	178	0.16%
Soc. di Mutuo Soc. (SoMS)	103	0.08%
Rete associativa (RA) [†]	7	0.01%
Total	134,815	100%

Notes: Percentages may not sum to 100 due to rounding.

[†] RA includes ETS that are not registered to any other RUNTS section to avoid double counting. Including RAs also registered to other RUNTS sections increases their total number to 63.

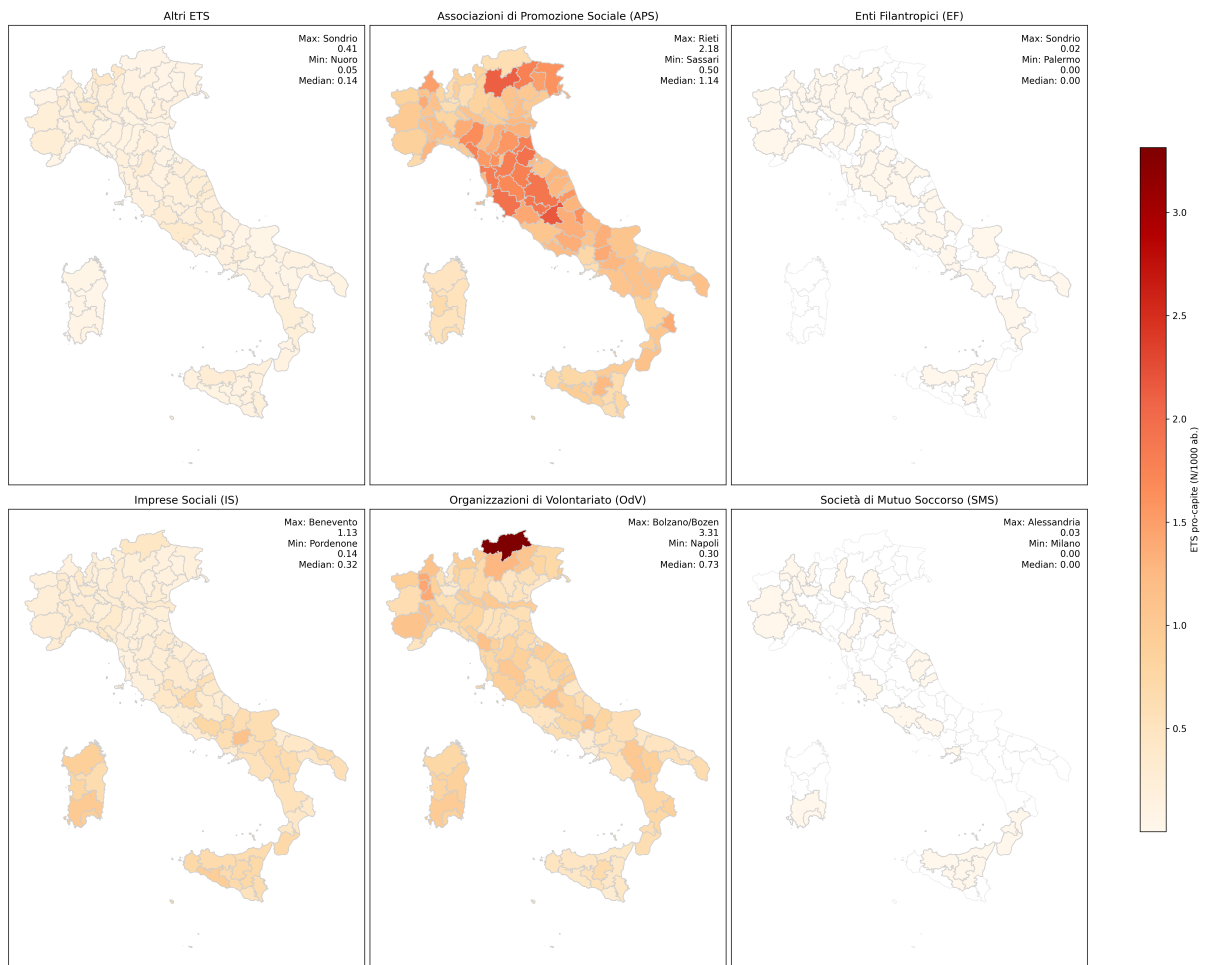
Table 3: Registered ETS by decade of establishment. 2025Q2.

Decade of establishment	N	N%
Before 1950	1,872	1.4%
1950–59	1,177	0.9%
1960–69	1,736	1.3%
1970–79	3,599	2.7%
1980–89	7,544	5.6%
1990–99	14,362	10.7%
2000–09	21,570	16.0%
2010–19	32,855	24.4%
2020–25	50,100	37.2%
Total	134,815	100.0%

Particularly long-lived among them are the SoMS, over a quarter of which were established before the beginning of the twentieth century.

The Geography of the Italian Third Sector The data show that the Third sector has capillary diffusion on the Italian territory. With at least one organisation registered in 7,437 municipalities, corresponding to 94.2% of the 7,900 municipalities that are present in Italy in the second quarter of 2025, the evidence shows extensive penetration and underscores the sector's foundational role within local communities. The national average is 42 ETS per municipality, with substantial variation in relation to the type of services locally available. Service-provision centres, which are cities or groups of cities that jointly offer essential services in areas of health, education, and connections (Istat, 2022), have the highest concentration of ETS, 813 on average. The top ten cities have 17,775 ETS; i.e., 14.9% of the total. The number of ETS decreases exponentially with the distance from service-provision centres as the average drops to 28 organisations in outlying areas, 23 in intermediate, 20 in peripheral, and 13 in ultra-peripheral municipalities. Despite the high density in urban hubs, the presence of ETS is pervasive on the Italian territory and is an asset to social cohesion and development in more marginalised territories.

Figure 3: ETS density by section and province. 2025Q2.



The Governance of the Italian Third Sector The governance structure of the Italian ETS is characterised by a mature, predominantly male, and embedded in local communities leadership profile. Legal representatives have an average age of 58 years, with a significant gender imbalance (62.2% male). Notably, female leaders are, on average, younger than their male counterparts (55 years old) and have a tendency to share responsibilities with other women. The localism of the sector is confirmed by the fact that nearly two-thirds (65%) of legal representatives operate within an ETS located in their province of origin.

Regarding formal governance, a substantial majority of organisations (61.1%) have structured formal governance, reporting the presence of at least one administration, control, or auditing body. This arrangement is prevalent within OdV and APS, which together account for 90% of the ETS with structured management.

The Economic Activities of the Italian Third Sector The analysis of the activities carried out by ETS provides essential insight into the composition in terms of the main areas of activity of the Italian Third Sector. Following the legal provisions of the CTS, each ETS must pursue at least one activity of general interest, thereby ensuring alignment between organisational purpose and collective welfare. The RUNTS requires each ETS to self-declare all the activities of general interests pursued by each organisation according to two classifications: the Art. 5 CTS classification (see Table D.1 in the Appendix) and the ICNPO (see Table D.2), which allows for international comparability.

The evidence concerning the activities of general interest ex Art. 5 pursued by the ETS highlights the strong cultural and educational vocation of the sector (see Table 4). The most widespread domains are cultural, artistic, and recreational activities of social interest (68.5%) and education and training (42.4%), followed by social services (26.6%) and charitable activities (25.4%). Environmental protection (21.2%) and the promotion of legality and civil rights (around 24%) also represent significant areas of engagement, confirming the multidimensional nature of social utility pursued by Italian ETS, also supported by the fact that only a few categories have low shares across the distribution. Moreover, it is interesting to note that only 26,362 ETS (30.2%) declare that they perform a single activity, being the average number activities 4.3 (median 3), revealing the multidimensionality of ETS activities.

The picture emerging from the ICNPO classification confirms the prevalence of cultural, recreational, and social activities, alongside a widespread presence of organisations pursuing educational, environmental, and solidarity-based objectives observed under the classification ex Art. 5 CTS.

The majority of the ETS operate in the area of Culture and Recreation (ICNPO group 1000), which alone accounts for more than half of all declared activities (64.1%). Within it, Cultural and Arts (ICNPO section 1100) (45.0%) and Other Recreational and Social Clubs (1300) (44.0%) are particularly prominent, reflecting the centrality of social aggregation and community participation among ETS. Sports organisations (1200) also represent a significant share (12.4%). The second most relevant field concerns Social Services (4000), which accounts for around 23.9%, encompassing both general social assistance and emergency services. These are followed by the area of Education and Research (2000), in which 14.1% of ETS are active; interestingly 7.8% of ETS pursue Vocational and Adult Education (2300) objectives, confirming the sector's contribution to lifelong learning and skills development, as well as social inclusion.

Table 4: Number of ETS by general interest activities ex Art. 5, CTS. 2025Q2.

Code	Description	N	N%
A	Social services & interventions, includ. services for people with disabilities	23,285	26.6%
B	Health interventions and services	8,750	10.0%
C	Socio-sanitary and healthcare services	13,063	14.9%
D	Education, vocational train., and cultural activities with educ. purpose	37,083	42.4%
E	Environmental protection and the sustainable use of natural resources	18,556	21.2%
F	Protection and enhancement of cultural heritage and landscapes	21,383	24.5%
G	University and post-graduate education	4,829	5.5%
H	Scientific research of significant social interest	8,931	10.2%
I	Cultural, artistic, or recreational activities of social interest	59,912	68.5%
J	Non-profit community radio broadcasting	3,722	4.3%
K	Tourism of social, cultural, or religious interest	22,345	25.6%
L	Extra-curricular training to prevent school dropout and combat bullying	20,201	23.1%
M	Support services for the third sector	3,287	3.8%
N	Development coop. aimed at eradicating poverty and reducing inequality	4,289	4.9%
O	Fair trade and solidarity-based commerce	2,228	2.5%
P	Job placement or re-entry into the labour market	4,356	5.0%
Q	Social and affordable housing	7,132	8.2%
R	Humanitarian reception and social integration for migrants	9,939	11.4%
S	Social farming and agriculture	4,200	4.8%
T	Amateur sports activities	15,251	17.4%
U	Charitable activities, including donations of funds, goods, or services	22,243	25.4%
V	Promotion of a culture of legality, peace, and non-violence	20,883	23.9%
W	Promotion and protection of human, civil, social, and political rights	21,468	24.6%
X	International adoption services	394	0.5%
Y	Civil protection and emergency services	5,181	5.9%
Z	Redevelop. of unused public or confiscated from organised crime assets	13,347	15.3%
Number of ETS declaring at least one activity		87,429	

Notes: Since each ETS can declare more than one activity, column N% presents the share of ETS that performs the corresponding activity.

Philanthropic Intermediaries and Voluntarism Promotion (8000) (19.2%) is highly relevant, with a particular importance of Volunteerism Promotion (8200) (17.1%). An important share of ETS are also active in environmental activities (5000) (14.6%), mainly in Environmental Protection (12.1%). A non-negligible share of ETS also operates in domains related to Law, Advocacy, and Politics (7000) (7.3%), and International Activities (9000) (2.8%), thus highlighting the pluralism of purposes and the articulation of the Third Sector across different dimensions of general interest.

In terms of territorial distribution, it is interesting to note that the most widespread activity according to the ICNPO classification is culture and leisure, which is offered by at least one ETS in 82.1% of municipalities. It is followed by social services (present in 59.7% of municipalities), environmental protection (58.7%), and the activities of foundations, charities, and volunteer organisations (55.2%). The least common activities are those related to international cooperation, for which at least one ETS is active in 13.2% of municipalities.

In both cases, a number of limitations of these data emerge. First, coverage is partial, as only a subset of the ETS declares at least one economic activity ex Art. 5 CTS or according to the ICNPO—87,429 for the classification ex Art. 5 CTS and 92,134 for the ICNPO, over a total of 135,815. Moreover, the level of detail of the ICNPO is not consistent across ETS, since they are allowed to choose among group- or section-level categories. Finally, a large number of ETS choose the residual category 12000 which can be seen as a residual category. To address these limitations, validate the RUNTS data, and enrich the database with new dimensions of interest, we introduce an exercise of web-augmentation, which the rest of the paper presents.

4 Enrichment of the Administrative Data with Web-augmentation and Machine Learning Techniques

A frequent and particularly pernicious limitation encountered in administrative data is the presence of optional, non-compulsory fields in the administrative forms that are the legal basis of the data observed (Wallgren and Wallgren, 2022). This lack of compulsoriness often leads to violations of *missing completely at random* (MCAR) assumptions. At the same time, administrative data often miss the information needed to control for the factors affecting these different behaviours by the respondents, leading to *missing not at random* (MNAR) situations. For example, respondents may systematically decline to fill in optional fields related to sensitive or complex information, precisely because those values are perceived as too high/low or too sensitive. In addition, the outcome of the administrative process can differ depending on whether the respondent was the individual itself or someone acting on behalf of this individual. This non-compulsory nature of the form thus inherently introduces a systematic selection bias into the observed data. This MNAR mechanism fundamentally undermines one of the chief advantages of administrative registers: their comprehensive coverage, since statistical estimates derived from that field will suffer from non-response bias. Furthermore, standard imputation techniques become inadequate or invalid under MNAR conditions.

In the RUNTS case, a key example of this situation is information about the main economic activity carried out by each organisation. Respondents are allowed to indicate their economic activities from a list of ICNPO group- and section-level categories (see Table D.2 in the Appendix). However, this field in the form is not compulsory and, referring to the second quarter of 2025, only

Table 5: Number of ETS by ICNPO category. 2025Q2

Code	Description	N	N%
1000	Culture and Recreation	10,215	11.1%
1100	Culture and Arts	41,420	45.0%
1200	Sports	11,451	12.4%
1300	Other Recreation and Social Clubs	40,532	44.0%
2000	Education and Research	2,181	2.4%
2100	Primary and Secondary Education	2,471	2.7%
2200	Higher Education	1,628	1.8%
2300	Vocational and Adult Education	7,158	7.8%
2400	Research	3,972	4.3%
3000	Health	2,541	2.8%
3100	Hospitals and Rehabilitation	541	0.6%
3200	Nursing Homes	261	0.3%
3300	Mental Health and Crisis Intervention	296	0.3%
3400	Other Health Services	7,760	8.4%
4000	Social Services	5,069	5.5%
4100	Social Services (excluding emergencies)	13,732	14.9%
4200	Emergency and Relief Services	5,294	5.7%
4300	Income Support and Material Assistance	4,034	4.4%
5000	Environment	2,111	2.3%
5100	Environmental Protection	11,185	12.1%
5200	Animal Protection	3,530	3.8%
6000	Development and Housing	367	0.4%
6100	Economic, Social, and Community Development	8,412	9.1%
6200	Housing	1,080	1.2%
6300	Employment and Training	2,663	2.9%
7000	Law, Advocacy, and Politics	531	0.6%
7100	Civic and Advocacy Organisations	6,400	6.9%
7200	Legal Services	553	0.6%
7300	Political Organisations	19	0.0%
8000	Philanthropic Intermediaries and Voluntarism Promotion	2,202	2.4%
8100	Grant-making Foundations	1,842	2.0%
8200	Volunteerism Promotion	15,773	17.1%
9000	International Activities	425	0.5%
9100	International Economic and Humanitarian Assistance	2,324	2.5%
12000	Other	1,500	1.6%
12100	Not Elsewhere Classified	15,093	16.4%
Number of ETS declaring at least one ICNPO category		92,134	

Notes: Since each ETS can declare more than one activity, column *N%* presents the share of ETS that performs the corresponding activity. Moreover, we present section- and group-level classes together because ETS can declare activity classes at either or both levels.

67% of the respondents provided this information. Moreover, the number of groups and sections specified by a single ETS is various, going from one, in 7,646 cases (42.8 percent of the total), to twenty (i.e., a large part of the possible twenty-five section-level categories offered by the ICNPO taxonomy) in six cases. To try and tackle this limitation, we explored the possibility of detecting and determining the main economic activity of each ETS based on their online behaviour.

This exercise is similar, in spirit, to previous studies, which investigated the possibility of automatically identifying an appropriate category of organisations of the Social Economy ([Rutherford et al., 2025a,b](#)). Among others, [Litofcenko et al. \(2020\)](#) focus on the classification of Austrian non-profit organisations (NPOs) using the ICNPO taxonomy. The authors tackle the problem from a peculiar view-point, that is, when only the organisation's name is available as input data. The researchers compare two semi-automated methods, a rule-based and an ML method that uses curated keywords. The authors conclude in favour of the rule-based classifier with a twofold reasoning. First, the rule-based method offered a higher degree of transparency compared to the ML approach. Second, it also performs comparably well as manual human coding—in terms of precision and sensitivity—, while offering significant labour saving advantages.

In contrast, [Ma \(2021\)](#) applies state-of-the-art computational linguistics to massive administrative datasets. The author develops a high-performing ML classifier, employing the Bidirectional Encoder Representations from Transformers (BERT) model, to remap the U.S. non-profit sector—i.e., more than 439,000 NPOs—based on the U.S. National Taxonomy of Exempt Entities (NTEE) taxonomy. This research achieves astonishing results, with a final classification almost perfectly aligned with that of human coders. Crucially, this classifier was used to reassign multiple category codes to each NPO and uncover organisational activities previously ignored in the manual classification.

Finally, [Damm and Kane \(2024\)](#) present a novel, more granular classification taxonomy for UK charities. This taxonomy is specifically designed to better capture the local context and critical sub-sectors that existing schemes poorly represent. Their classification also allows for multiple classes per charity, providing a higher level of detail than traditional single-code systems. Similarly to the rule-based approach discussed previously, the researchers applied their classification to charity registers using automated keyword search rules. This keyword search method proved to be a sufficiently accurate and transparent way to implement the new, detailed classification system, offering a richer picture of the voluntary sector.

As said, our aim is in line with these publications. However, shifting from these approaches, and in line with a growing economic literature which involves the analysis of the corporate website of for-profit businesses (see, e.g., [Kinne and Axenbeck, 2020](#); [Daas and van der Doef, 2020](#); [Ashouri et al., 2022](#); [Crosato et al., 2023](#); [Rammer and Es-Sadki, 2023](#); [Bottai et al., 2025](#)), our classifier uses information from the organisational website of each ETS—with which we web-augment the RUNTS administrative data. Indeed, the increasing availability of digital information allows for text to be used as a valuable complement to structured administrative data in economic research ([Gentzkow et al., 2019](#)) and, in this sense, the web is a timely, large-volume, and flexible data source that is essential for advancing research in the social sciences ([Askitas and Zimmermann, 2015](#)).

Several attempts exist to use corporate websites as the source to inform classifiers aimed at identifying the economic activity carried on by for-profit businesses. Indeed, as underlined by [Domenech et al. \(2012\)](#); [Blazquez and Domenech \(2018\)](#), companies' websites are functional for

marketing and the dissemination of business activities. Therefore, content analysis may allow for an accurate estimation of their economic activity. For example, [Berardi et al. \(2015\)](#) tackle this problem by using supervised ML techniques to automatically classify corporate websites. They built a hierarchical classification system and tested its accuracy on a dataset of more than 20,000 URLs classified according to a 2-level taxonomy of 216 leaf classes explicitly designed for market research. Particularly interesting is the use, as predictors, of words extracted from the most common queries, run in the previous months on Amazon Alexa, that directed traffic to the website of the company in question. This source shows astonishing results in predicting the companies' macro-class of economic activity. The researchers concluded that these queries are highly descriptive of the content of the websites to which they are directed. This information density helps compensate for the scarcity of endogenous information often available for rare or specialised business sectors, a problem that severely affects infrequent classes.

Instead, [Kühnemann et al. \(2020\)](#) investigate the automatic classification of businesses' main economic activity, according to the Statistical Classification of Economic Activities in the European Community (NACE) taxonomy, using text scraped from their corporate websites. The main focus of the paper is on reducing the time-consuming manual effort of national statistical institutes in labelling firms. The authors tested a so-called 'knowledge-based approach' which sought to leverage human-defined expertise in two ways. First, they use the presence of pre-defined, domain-specific keywords and synonyms, instead of raw words, as features. Second, they implement a two-level hierarchical classification to mimic the workflow of a manual NACE editor. Ultimately, the study found that this expert-driven approach did not improve performance compared to a simpler, purely data-driven model using basic character n -grams, leading to the conclusion that straightforward text-mining features can be more effective for this task.

Finally, [Fernández Navarro et al. \(2025\)](#) propose an automated methodology for the fine-grained classification of economic activity, using a sample of 151,180 Spanish firms. The approach uses textual content extracted from a broad selection of web pages of corporate websites of these firms as the primary input. The purpose is to automatically classify the main economic activity carried out by each firm according to the four-digit CNAE 2009 taxonomy. Unstructured text was first transformed into high-dimensional semantic embeddings using an OpenAI model (text-embedding-3-small). These embeddings are used as features for the final classifier. Among the several ML models tested, a Hierarchical Support Vector Machine model is the best-performing one, achieving promising results on the test set.

Alike for-profit businesses, ETS use their organisational websites as communication hubs for stakeholders. In both cases, these digital platforms serve as readily-accessible, dynamic data complements to traditional administrative records. Therefore, we exploit the organisational websites of Third Sector organisations to inform a ML model and use this richer content of information to assign a few ICNPO group- and section-level categories to each organisation. Among others, a clear limitation of this approach relates to the availability and quality of these organisational websites: not all ETS have their own website, not for all ETS who own one we have been able to identify this website, and not all websites that we have been able to detect are of comparable quality in terms of information provided or structured in a way such that our algorithm is equally good at exploiting their information content. Still, as we will show, this exercise leads to some interesting results.

4.1 Search for the Websites of the Organisations

To classify our ETS, we start by trying and assigning a website to each organisation in the register¹⁴. To do so, we rely on the websites self-declared by the organisations on the RUNTS and on the hits returned by Microsoft Bing APIs, searching for the name and tax code of each organisation.

For each result obtained from the query (or for the self-declared website, if present), we developed a crawler that downloads up to 20 web pages from the website. The crawler starts from the homepage of the website and collects all the hyperlinks in this web page that point to other pages of the same website (i.e., internal hyperlinks with a page-depth equal to one). It prioritises pages like ‘about us,’ ‘contacts,’ or ‘privacy.’ For the others, it sorts them according to the length of their URL, under the presumption that the shorter a URL, the more the corresponding page will be a meaningful-information carrier—see [Abbasiharofteh et al. \(2023\)](#) for a similar approach.

For the list of websites detected via the search engine, we (*a posteriori*) verify the accuracy of the match between the website and the corresponding ETS, using the downloaded web pages. To do so, we consider as correctly matched to an ETS the first website (if any), on the list returned by the search engine, on which we detected at least one of the following pieces of information referring to the corresponding ETS: tax code, or location (municipality and county), full address (in a form like ‘Main street 123’), or postcode of its headquarters—see [Barcaroli et al. \(2016\)](#); [Bottai et al. \(2022\)](#) for similar approaches.

4.2 Classification of the Economic Activity of the Organisations

As mentioned above, the final goal of this web-augmentation procedure is to try and identify the economic activity performed by an ETS. To do so, we developed an ML (Generative AI) classifier ([Vaswani et al., 2017](#); [Ortame et al., 2025](#)) that takes advantage of information from the textual content of a predefined subset of web pages from the ETS organisational website, previously identified, fetched, and archived by the crawler described above.

The salient set of pages is selected according to priority rules, based on an ex-ante score assigned to each website page. The top priority is given to the pages ‘what we do,’ ‘about,’ ‘mission,’ ‘activity,’ and ‘services.’ In case none of the relevant pages is found, all pages are bound and used jointly.

The classifier consists of well-structured and targeted questions for the ‘GPT-4o mini’ Large Language Model (LLM), as provided by the Microsoft Azure OpenAI Service and whose ‘temperature’ parameter is set to zero, to avoid erratic behaviours. This is today the OpenAI’s most cost-efficient small multimodal AI model. Compared to existing rivals, GPT-4o mini is considered to excel in terms of performance-to-cost ratio, outperforming competing small models on textual reasoning benchmarks. On the other hand, it offers a smaller context window (128K tokens) and a slower response time.

To perform our task, the GenAI classifier is instructed to return one (or more) ICNPO categories, a confidence score, and a short text to justify the choice it made. Specifically, we proceed in two

¹⁴We recognise that multiple websites can be associated to the same organisation and we acknowledge this as a limitation of the current approach, but this is mostly true only for large organisations that are, in this context, a minority.

steps. First, we ask the system prompt the following statement (*first prompt*)¹⁵:

```
You are an expert in analysing non-profit organisations. Your task is to extract important information from the text provided, focusing on (1) the mission of the organisation; (2) its goals; (3) areas of intervention; (4) activities performed; (5) beneficiaries; (6) and main geographical area of activity.
```

Then, we ask the system prompt to perform the following task (*second prompt*):

```
You are an expert in classifying non-profit organisations according to the ICNPO standard. Please, analyse the information previously extracted to assign the most appropriate ICNPO categories among the following:
```

```
1000 CULTURE AND RECREATION
```

```
- 1100 Culture and Arts
```

```
- 1200 Sports
```

```
[...]
```

```
RETURN A PYTHON DICTIONARY with this format:
```

```
{
  'classifications': [
    {
      'macro_category': 'XX000',
      'macro_category_confidence': 0.95,
      'justification': 'brief explanation of this choice',
      'subcategories': [{
        'code': 'XX100',
        'confidence': 0.90,
        'justification': 'brief explanation of this choice'
      }]
    }
  ]
}
```

```
CLASSIFICATION CRITERIA:
```

1. High confidence (> 0.8): activities clearly central to the mission and objectives
2. Medium confidence (0.6-0.8): activities mentioned but not central
3. Ignore classifications with confidence < 0.6
4. An organisation can belong to more than one category

The two prompts together use a two-stage expert system pipeline, which consists of decomposing a complex task into modular stages. This approach is designed to move from an unstructured textual source to standardised, quantitative data. This structure ensures that the final output is both rich in detail and comparable across different ETS, pushing for a rigorous scientific investigation. The pipeline simulates the workflow of a domain expert by performing sequential tasks: initial data extraction followed by formal classification.

The core function of the first prompt is the systematic synthesis of data according to a highly

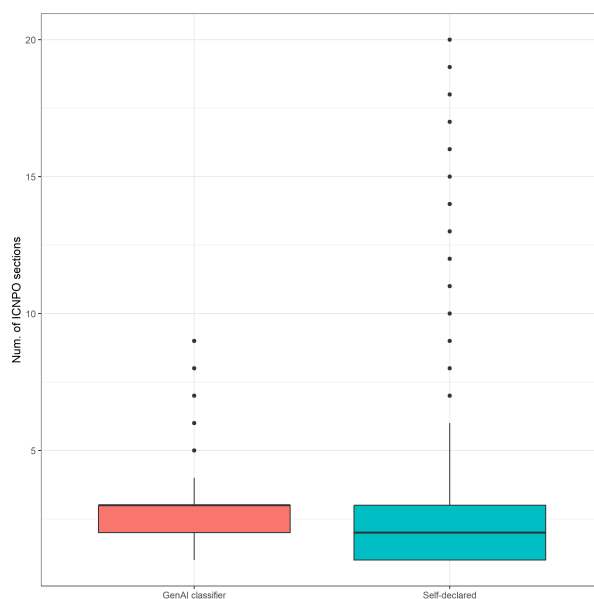
¹⁵The interested reader will find the prompting texts in the Appendix [E](#).

structured output schema. The system is instructed to delineate six fundamental attributes—mission, goals, areas of intervention, activities performed, beneficiaries, and geographical area of activity—from the text of the web page provided. This explicit, six-point structure converts an open-ended generation task into an information-extraction and slot-filling problem (Xu et al., 2024), which minimises ambiguity and is critical to ensure the systematic, objective, and comparable nature of the resulting organisational profiles. Asking the language model to focus on specific elements is a constraint-based prompting style, crucial for data integrity: it ensures that all required variables are addressed and maximises the consistency and utility of the extracted information for subsequent analysis.

Instead, the second prompt uses the structured data generated from the first one to perform a standardised classification to assign the most appropriate categories from the ICNPO taxonomy. This is a structured-output prompt, designed to execute a standardised classification and enforce transparent, evidence-based reasoning on the data derived from the previous step. Asking the second prompt to only analyse the information previously extracted is critical to prevent potential drift back to the original source text, ensuring a classification purely based on the structured, intermediary data. Instead, the constraint on the output forces the model to structure its final answer for machine readability and automated parsing, which is essential for integration into a research database. Moreover, the requirement to self-explain and provide a justification for each ICNPO category selected by the model is an attempt to prevent hallucinations and be able to control, ex-post, the quality of the classification. Finally, the prompt explicitly links a quantitative confidence score to the centrality of the activity, while the allowance for multiple categories addresses the real-world complexity of multi-mission organisations.

The prompting style used in both cases is often known in the literature as ‘role-based.’ This style has been shown to be effective in improving task performance in reasoning and explanation, particularly because it functions as a ‘cognitive anchor’ for the language model and narrows the scope of the request down to a specific area of its knowledge base (see, e.g., Kong et al., 2024). Role-based prompting assigns a persona—like ‘paralegal’ or ‘engineer’—to an AI model to guide the style, tone, and focus of responses. Indeed, according to Shanahan et al. (2023), role-based prompting, like conditioning in probability theory, collapses what they call ‘superposition of simulacra’ and biases the model’s token prediction process to sample from a subset of its massive knowledge base. As such, role-based prompting can help generate specialised or expert-level content and improve accuracy in tasks that require domain knowledge. Although its effectiveness has been recently questioned in the context of state-of-the-art models, role-based prompting is still considered capable of enhancing open-ended tasks like our own. On the other hand, a well-recognised limitation, relevant for our application, is that the effectiveness of this prompting strategy depends heavily on the quality of the role representation in the model’s training data. Among the alternative strategies with which we experimented, this prompting style was found to be the most effective. However, further exploration is needed to properly assess its effectiveness (Zheng et al., 2024). With specific reference to official statistics applications, the advantages of role-based prompting are widely discussed in the literature (see, e.g., High-Level Group for the Modernisation of Official Statistics (UNECE) et al., 2025, p. 16) and, in recent years, statistical offices have experimented with strategies similar to ours to classify the main economic activity of for-profit businesses (see, e.g., Kühnemann et al., 2020; Faria, 2025).

Figure 4: Distribution of the number of section-level ICNPO categories self-reported by each ETS (box-plot on the right) or assigned by the GenAI classifier (box-plot on the left)



Notes: This figure considers only the 17,862 organisations for which both classifications are available.

4.3 Insights from the Web-augmentation Exercise

The results of the classification exercise show promise, even though limitations emerged that will be discussed at the end of the section together with strategies to overcome them and future research directions.

We managed to assign at least a section-level ICNPO category to 19,680 ETS (about 15% of the total), mostly due to the lack of an organisational website for the others—either because it does not exist at all or because we did not manage to identify or fetch it.

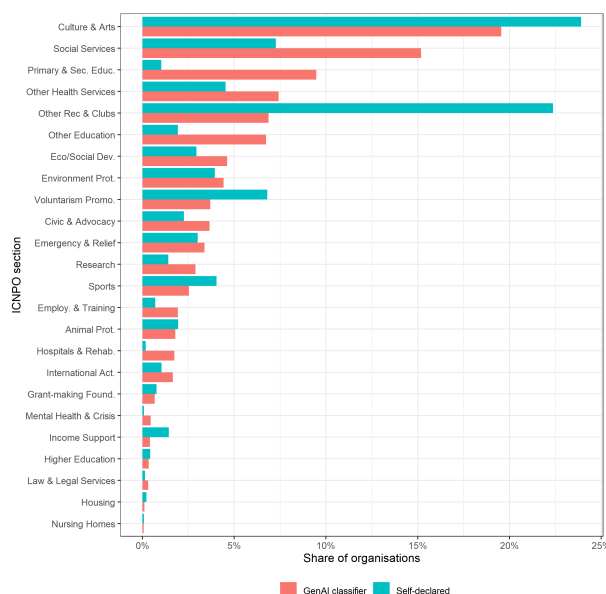
We can compare this classification with the self-declared section-level ICNPO categories, available for 88,043 organisations.¹⁶ Of these organisations, in 17,862 cases we have ICNPO categories both self-declared and assigned by the GenAI classifier. The results show that, if we look at the distribution of the number of section-level ICNPO categories self-declared by an individual ETS (Figure 4, box-plot on the right), we can see a larger variety (with an IQR of 2, from 1 to 3 categories per ETS) and a greater number of outliers in the upper whiskers. Instead, the GenAI-based classifier (box-plot on the left) tends to assign the same number of section-level ICNPO categories to all the organisations (2.60 on average, compared to 2.48 self-declared categories). However, the classifier mostly avoids cases with many categories assigned to a single organisation.

Moreover, in the self-declared data, the distribution of section-level ICNPO categories is much more unevenly distributed, with a high concentration in a few categories—see Figure 5¹⁷. Instead, the GenAI-based classifier tends to use the possible categories more evenly. This evidence is compatible with the fact that the members of the Third Sector organisations—governing bodies, volunteers, workers, or even professional representatives—may not report in this field of the RUNTS the set of

¹⁶Please note that, even though 92,134 ETS declared some ICNPO category, 4,091 of them only self-declare a group-level category. Here, these last have been dropped for comparability.

¹⁷Figure E.3 in the Appendix reports the same information in the form of two treemaps

Figure 5: Distribution of ETS per section-level ICNPO category using either the self-reported information (blue bars) or that provided by the GenAI classifier (red bars)



Notes: This figure considers all the organisations for which at least one of the two classifications is available. A random category has been assigned to any ETS that self-declared more categories or that have been classified by the GenAI model into multiple categories, so that the distribution reported in figure also represents the distribution of ETS in the sample.

economic activities actually performed by the organisation—as they should, given the specific focus of the question asked by the public administration—but rather use this field to identify and signal to the public a thematic area of intervention that represents their objectives, mission, and vision.¹⁸ On the contrary, the ML classifier is based on observational data—even though they themselves are self-declared. The reliability of the results is also enhanced by the rationale we adopt in the selection of the relevant knowledge base used for this classification activity, which leverages a set of prioritised web pages that are believed to carry relevant and high-quality information for the task.

Although encouraging, these initial findings still suffer from several important limitations. First, we have been able to identify a website and fetch some of its pages only for a minority of the ETS listed on the RUNTS. Any future improvement cannot ignore a resolution of this point. We should both assess how many ETS actually have an organisational website—to be able to quantify the significance of the missing information in our data—and further improve the ability of our pipeline to detect and fetch these web pages. Second, we need to find a solid, statistical method to validate our results, which essentially means determining a valid and trustable ‘ground truth’ to compare our results with. The issue here is that the self-declared information is likely biased, for reasons discussed above, and therefore invalid. An alternative has to be identified. Finally, the GenAI classifier used in the paper suffers from black-box limitations widely discussed by the literature. To partly overcome this limitation, we need to compare the results of this model with those of alternative ones. However, to do so, we need to solve the lack of ground truth problem already discussed.

¹⁸The validity and reasonableness of this hypothesis have been confirmed to us, in a private conversation, by several field experts and practitioners.

5 Discussion and Conclusions

In this work, we present a novel, integrated data set on the Italian Third Sector, built by extracting and validating information from the national register of Third Sector organisations (RUNTS) and by enriching it with web data. The procedure combines data acquisition, cleaning, and validation. Then by adopting a structured GenAI classification strategy, we produce the first comprehensive portrait of Italian Third Sector organisations using web-enriched, scraped administrative data. The resulting database offers an up to date quantitative map of more than 130,000 ETS which can be leveraged for statistical analyses.

The descriptive evidence highlights the heterogeneity and territorial capillarity of the sector and shows the coexistence of various entrepreneurial models within the Italian Third Sector, spanning from associative volunteering-based organisations to more market-oriented organisations. The data also reveal marked regional differences in density and specialisation. These patterns provide a factual basis regarding the contribution of the Third Sector to welfare provision and local development.

At the same time, the analysis uncovers several measurement challenges that are typical of administrative sources. The RUNTS fields are partly optional, the degree of completeness varies across sections, and a large share of organisations do not declare their activity classification. Integration with web data mitigates some of these limitations by retrieving additional descriptive information and by assigning ICNPO categories to previously unclassified organisations. Future developments will be devoted to the exploration of additional dimensions of the data yet to be explored in detail, regular updates of the dataset. Most importantly, the availability of social and sustainability reports constitutes an important turning point in the enrichment strategy as they are more comprehensive and focused documents compared to webpages. The availability of these reports in the RUNTS is extensive and can provide a valuable information source on a variety of topics: type of activities and projects in which the ETS is involved and/or it directly organises, the content of these initiatives, funding channels, beneficiaries, and area of activity.

In this way, the project contributes to the construction of a timely and detailed database on the Italian Third Sector and demonstrates the potential of integrating administrative and digital data in socio-economic research. From an analytical point of view, the enriched RUNTS can support studies on the spatial distribution, governance, and economic role of the Third Sector, enable comparisons with for-profit enterprises and a comprehensive picture of the services offered by public administrations and ETS. Moreover, the data can be linked to external sources.

The combination of administrative and web data opens promising avenues for policy, as well. In so-doing, apart from its role as a monitoring device of administrative nature, the RUNTS can also be valorised as an informative support for policy design. The emerging knowledge can support co-programming practices and create value for ETS, for which the activities to update the RUNTS constitute a relevant administrative burden in their daily activities.

Moreover, the constructed data provides public administrations with an instrument that empowers its outreach capacities in monitoring the fabric of the local Third Sector. This knowledge can be used to inform co-programming initiatives in a context in which the collaboration between local administrations and ETS is limited (CNEL, 2025) and therefore supports the actual implementation of the circular subsidiarity model. In this model the provision of public services is programmed and implemented *jointly* by Public Administrations and the Third Sector. The collaborative gov-

ernance is founded on the principles of co-programming and co-design, where the two parties work together from the outset to identify community needs and define priorities strategies to address them, and jointly design and implement interventions. The Italian Constitutional Court (C. Cost., Sentenza 131/2020) has stressed that this model, implicit in art. 55 CTS, [...] is not based on the payment of prices or remuneration from the public party to the private one, but rather on the convergence of objectives and the pooling of public and private resources for the joint planning and design of services and initiatives aimed at raising the levels of active citizenship, cohesion and social protection, within a relational sphere that goes beyond mere utilitarian exchange". Moreover, the Constitutional Court adds in the same judgement that "[...] public authorities [are entrusted] with the task of ensuring,[...], the active involvement of Third Sector Organisations (ETS) in the planning, design and organisation of interventions and services in the areas of activities of general interest defined by Article 5 of the same Code." Therefore, even if not used for research purposes, appropriately prepared RUNTS data can serve local public authorities as a tool to effectively reach out to existing ETS and to proactively involve them in the formulation of local public policies.

References

- Abbasiharofteh, M., Kinne, J., and Krüger, M. (2023). Leveraging the digital layer: the strength of weak and strong ties in bridging geographic and cognitive distances. *Journal of Economic Geography*, 24(2):241–262.
- Angotzi, D., Macigno, L., Giusti, G., Notarstefano, G., Santone, F. S., and Terzo, G. (2025). A territorial proximity measurement of non-profit activity: Evidence from the Third Sector Single National Register (RUNTS) data. In Pollice, A. and Mariani, P., editors, *Methodological and Applied Statistics and Demography II. SIS 2024*, Italian Statistical Society Series on Advances in Statistics (ISSAS), pages 434–439. Springer Nature.
- Ardito, C., Berton, F., and Pacelli, L. (2023). Combined and distributional effects of EPL reduction and hiring incentives: an assessment using the Italian “Jobs Act”. *The Journal of Economic Inequality*, 21(4):925–954.
- Ardito, C., Leombruni, R., Costa, G., and d’Errico, A. (2025). Mortality and risk of cardiovascular diseases by age at retirement in three Italian cohorts. *Big Data Research*, 41:100543.
- Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., and Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. *Data in Brief*, 42:108246.
- Askitas, N. and Zimmermann, K. F. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, 36(1):2–12.
- Barcaroli, G., Scannapieco, M., and Summa, D. (2016). On the use of internet as a data source for official statistics: a strategy for identifying enterprises on the web. *Rivista Italiana di Economia Demografia e Statistica*, 70(4):25–41.
- Bassi, A. (2025). The terminological debate. In Bassi, A., Aquino Alves, M., and Cordery, C., editors, *The Future of Third Sector Research: From Theory to Definitions, Classifications and Aggregation Towards New Research Paths*, Nonprofit and Civil Society Studies, chapter 5, pages 49–66. Springer Nature.
- Berardi, G., Esuli, A., Fagni, T., and Sebastiani, F. (2015). Classifying websites by industry sector: a study in feature design. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC ’15, pages 1053–1059, New York, NY, USA. Association for Computing Machinery.
- Berners-Lee, T. and Fischetti, M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor*. Harper.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43.
- Berton, F. and Paruolo, P., editors (2026). *Data-driven learning in the EU: from administrative sources to public value*. Springer.

- Blazquez, D. and Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130:99–113.
- Borzaga, C., Bodini, R., Carini, C., Depedri, S., Galera, G., and Salvatori, G. (2014). Europe in transition: The role of social cooperatives and social enterprises. Euricse Working Papers 69/14, Euricse.
- Borzaga, C., Calzaroni, M., Fontanari, E., and Lori, M. (2023). Social Economy in Italy: Dimensions, dynamics and characteristics. Euricse Working Papers 125/23, Euricse.
- Borzaga, C. and Ianes, A. (2006). *L'economia della solidarietà: Storia e prospettive della cooperazione sociale*. Donzelli Editore.
- Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., and Liberati, C. (2022). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pages 338–344. Association for Computing Machinery.
- Bottai, C., de Rassenfosse, G., and Raiteri, E. (2025). A new approach to measuring invention commercialization: An application to the SBIR program. *Research Policy*, 54(9):105302.
- Burroni, L., Ramella, F., and Trigilia, C. (2017). *Fondazioni e sviluppo locale*. Donzelli Editore.
- CNEL (2025). Relazione sui servizi pubblici 2025. Technical report, Consiglio Nazionale dell'Economia e del Lavoro.
- Connelly, R., Playford, C. J., Gayle, V., and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59:1–12.
- Crosato, L., Domenech, J., and Liberati, C. (2023). Websites' data: a new asset for enhancing credit risk modeling. *Annals of Operations Research*, pages 1–16.
- Daas, P. J. and van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, 36(4):1239–1251.
- Damm, C. and Kane, D. (2024). Classifying UK charities' activities by charitable cause: a new classification system. *Voluntary Sector Review*, 15(2):249–275.
- Della Queva, S., Nicosia, M., and Stoppiello, S. (2023). Il settore non profit nelle Aree interne del Mezzogiorno: una prima analisi descrittiva. *Autonomie locali e servizi sociali*, XLVI(3):367–386.
- Devetak, M. and Mandel, A. (2023). Using Large Language Models to generate socio-economic datasets.
- Di Bella, G. (2022). The use of administrative data in Istat. United Nations Economic Commission for Europe (UNECE) Expert Meeting on Statistical Data Collection – Towards a New Normal?
- Domenech, J., de la Ossa, B., Pont, A., Gil, J. A., Martinez, M., and Rubio, A. (2012). An Intelligent System for Retrieving Economic Information from Corporate Websites. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 573–578. IEEE Computer Society.

- European Commission (2020). *Social enterprises and their ecosystems in Europe: comparative synthesis report*. Publications Office of the European Union, LU.
- European Commission (2021a). Building an economy that works for people: an action plan for the social economy. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions.
- European Commission (2021b). *Impact of the European Commission's Social Business Initiative (SBI) and its follow-up actions: final report*. Publications Office of the European Union, LU.
- European Commission, Euricse, Spatial Foresight, Carini, C., Galera, G., Tallarini, G., Chaves Avila, R., Sak, B., and Schoenmaeckers, J. (2024). *Benchmarking the socio-economic performance of the EU social economy: improving the socio-economic knowledge of the proximity and social economy ecosystem*. Publications Office of the European Union.
- Faria, T. (2025). From NACE 2008 to NACE 2025: Retraining an ML model in production using Large Language Models. United Nations Economic Commission for Europe (UNECE) Generative AI and Official Statistics Workshop 2025.
- Fernández Navarro, J., Debón, A., and Domenech, J. (2025). Classification of companies' economic activity: A web content and machine learning-based approach. In *Proceedings of the 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025)*. Editorial de la Universitat Politècnica de València.
- Fondazione Terzjus (2022). *Dal non profit al terzo settore: una riforma in cammino. 2° rapporto sullo stato e le prospettive del diritto sul terzo settore in Italia*. Terzjus Report. Editoriale Scientifica.
- Fondazione Terzjus (2023). *Dalla regolazione alla promozione: una riforma da completare. 3° rapporto sullo stato e le prospettive del diritto del terzo settore in Italia*. Terzjus Report. Editoriale Scientifica.
- Fondazione Terzjus (2024). *A due passi dalla meta: verso il completamento della riforma. 4° rapporto sullo stato e le prospettive del diritto del terzo settore in Italia*. Terzjus Report. Editoriale Scientifica.
- Generali and Fondazione Cattolica (2024). Protagonista della crescita e della coesione sociale del Paese: evoluzione e sfide. Rapporto Terzo Settore 2024.
- Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.
- Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 4(2):205395171774567.
- High-Level Group for the Modernisation of Official Statistics (UNECE), Abu Awad, Y., Anwar, F., Choi, I., Clarke, C., del Monaco, A., Engdahl, J., Jadeau, J., Kipkeeva, A., Leinweber, A., Letardi, S., Lo Giudice, G., Massoli, P., McLaughlin, J., Senanayake, P., Silipo, M., Sirello, O., Solomon, J., Tessitore, C., Trace, S., Vaiciulis, V., Yoon, J., and Wang, L. (2025). Generative AI for Official Statistics. HLG-MOS Report, United Nations Economic Commission for Europe (UNECE).

- Ianes, A. (2016). Exploring the origins of social enterprise: Social co-operation in the Italian welfare system and its reproduction in Europe (from the 1970s to the present). Euricse Working Papers 88/16, Euricse.
- INPS (2025). XXIV Rapporto annuale.
- Istat (2022). The geography of inner areas in 2020. Territories between potential and weaknesses. Press release, Italian National Institute of Statistics.
- Istat (2024). Struttura e profili del settore non profit: anno 2022. Press release, Italian National Institute of Statistics.
- Istat (2025). Struttura e profili del settore non profit. Anno 2023. Press release, Italian National Institute of Statistics.
- Istat and Euricse (2021). L'economia sociale in Italia: dimensioni, caratteristiche e settori chiave. Rapporto di ricerca, Istat and Euricse.
- Italian Ministry of Labour and Social Policies and Italian Union of Chambers of Commerce (2024). *Rapporto 2024 sul Registro Unico Nazionale del Terzo Settore*. Editoriale Scientifica, Napoli.
- Jones, K. H., Heys, S., Tingay, K. S., Jackson, P., and Dibben, C. (2019). The good, the bad, the clunky: Addressing challenges in using administrative data for research. *International Journal of Population Data Science*, 4(1).
- Kinne, J. and Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. *Scientometrics*, 125(3):2011–2041.
- Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. (2024). Better zero-shot reasoning with role-play prompting. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.
- Kühnemann, H., van Delden, A., and Windmeijer, D. (2020). Exploring a knowledge-based approach to predicting NACE codes of enterprises based on web page texts. *Statistical Journal of the IAOS*, 36(3):807–821.
- Kwon, B., Park, T., Perez-Cruz, F., and Rungcharoenkitkul, P. (2024). Large language models: a primer for economists. *BIS Quarterly Review*, pages 37–52.
- Leombruni, R. (2024). Interviewing Administrative Records. A conceptual map for the use of Big Data for economic research. *Statistica Applicata – Italian Journal of Applied Statistics*, page 1.
- Litofcenko, J., Karner, D., and Maier, F. (2020). Methods for Classifying Nonprofit Organizations According to their Field of Activity: A Report on Semi-automated Methods Based on Text. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 31(1):227–237.
- Lori, M. (2024). Il settore non profit in Italia. In *Le regole del gioco. Proposte di trasformazione per uno sviluppo integrale*.

- Lori, M. and Zandonai, F. (2020). Tempo di bilanci (e di mutamenti?): il ruolo delle istituzioni non profit nella geografia del welfare territoriale. *Politiche Sociali*, VII(3):459–484.
- Ma, J. (2021). Automated coding using machine learning and remapping the US nonprofit sector: A guide and benchmark. *Nonprofit and Voluntary Sector Quarterly*, 50(3):662–687.
- Mancino, A. and Thomas, A. (2005). An Italian pattern of social enterprise: The social cooperative. *Nonprofit Management and Leadership*, 15(3):357–369.
- Mertens, S. (1999). Nonprofit organisations and Social Economy: Two ways of understanding the Third Sector. *Annals of Public and Cooperative Economics*, 70(3):501–520.
- Monzon, J. L. and Chaves, R. (2008). The European Social Economy: Concept and dimensions of the Third Sector. *Annals of Public and Cooperative Economics*, 79(3–4):549–577.
- Moulaert, F. and Ailenei, O. (2005). Social Economy, Third Sector and solidarity relations: A conceptual synthesis from history to present. *Urban Studies*, 42(11):2037–2053.
- OECD (2023). *What is the social and solidarity economy? A review of concepts*. Number 2023/13 in OECD Local Economic and Employment Development (LEED) Papers. OECD Publishing.
- Ortame, F., Bruno, M., and Massacci, G. (2025). Transformer-based models for official statistics: a tutorial. United Nations Economic Commission for Europe (UNECE) Generative AI and Official Statistics Workshop 2025.
- Pfeffermann, D., Eltinge, J. L., and Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics. *Journal of Survey Statistics and Methodology*, 3(4):425–483.
- Rammer, C. and Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators – a literature review. *Technological Forecasting and Social Change*, 197:122874.
- Rutherford, A. C., Lepere-Schloop, M., and Perai, N. A. A. (2025a). Part I: The classification of civil society, nonprofit and third sector organizations. In Bassi, A., Aquino Alves, M., and Cordery, C., editors, *The Future of Third Sector Research: From Theory to Definitions, Classifications and Aggregation Towards New Research Paths*, Nonprofit and Civil Society Studies, chapter 9, pages 101–111. Springer Nature.
- Rutherford, A. C., Lepere-Schloop, M., and Perai, N. A. A. (2025b). Part II: Turning a critical lens on nonprofit classification: Opportunities and challenges in the digital age. In Bassi, A., Aquino Alves, M., and Cordery, C., editors, *The Future of Third Sector Research: From Theory to Definitions, Classifications and Aggregation Towards New Research Paths*, Nonprofit and Civil Society Studies, chapter 10, pages 113–124. Springer Nature.
- Salamon, L. M. and Anheier, H. K. (1992). In search of the non-profit sector II: The problem of classification. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 3(3):267–309.
- Salamon, L. M. and Anheier, H. K. (1994). Measuring the non-profit sector cross-nationally: a comparative methodology. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 4(4):530–554.

- Salamon, L. M. and Anheier, H. K. (1996). The international classification of nonprofit organizations: ICNPO–revision 1, 1996. Working Papers of the Johns Hopkins Comparative Nonprofit Sector Project 19, The Johns Hopkins Institute for Policy Studies.
- Salamon, L. M. and Sokolowski, S. W. (2016). Beyond nonprofits: Re-conceptualizing the Third Sector. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 27(4):1515–1545.
- Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101.
- Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role play with large language models. *Nature*, 623(7987):493–498.
- Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99.
- Stoppiello, S., Queva, S. D., and Nicosia, M. (2025). Official Statistics on Nonprofit Sector in Italy: The Permanent Census Design and New Perspective for Data Analysis. In Pollice, A. and Mariani, P., editors, *Methodological and Applied Statistics and Demography II*, pages 453–457. Springer Nature Switzerland, Cham. Series Title: Italian Statistical Society Series on Advances in Statistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wallgren, A. and Wallgren, B. (2022). *Register-based Statistics*. John Wiley & Sons.
- Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2024). Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357.
- Zamagni, V., Battilani, P., and Casali, A. (2004). *La cooperazione di consumo in Italia*. Il Mulino.
- Zheng, M., Pei, J., Logeswaran, L., Lee, M., and Jurgens, D. (2024). When “a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Appendices

A EU-level Social Economy Policy

The European Commission (EC) has sought to define the boundaries of the European Social Economy and support its development across the member States ([European Commission, 2021b](#); [European Commission et al., 2024](#)). The EC launched the Social Business Initiative (SBI) in 2011. Its primary objective was to establish an EU-level action plan with measures to create a favourable environment for social enterprises (SEs). The original SBI action plan included eleven key actions structured around three main objectives, namely improved access to funding, increased visibility of social entrepreneurship, and an improved legal environment. After 2016, two pillars were added: (i) social innovation, new technology, and business models, and (ii) the international scaling of SEs ([European Commission, 2021b](#)). The Social Economy Action Plan was presented in December 2021 and framed as a strategic follow-up to the SBI, designed to contribute to an inclusive recovery and the green and digital transitions after the COVID-19 pandemic. It pivoted around three main objectives. The first concerned the adaptation of regulatory frameworks to create a favourable environment for the development of these organisations. This objective was focused on the European Council recommendations and guidance for Member States on taxation frameworks for SEs and promoting socially-responsible public procurement. The second objective was to support the scaling of Social Economy organisations with increased financial support for the 2021–2027 period, including financial products catered to Social enterprises under the InvestEU programme, a European Competence Centre for social innovation (2022) and a centralised information hub on European funding (2023). Finally, the third objective was the recognition of the social economy, promoting qualitative and quantitative studies on the social economy across the EU, capacity building activities, and support for SEs through the green and digital transitions.

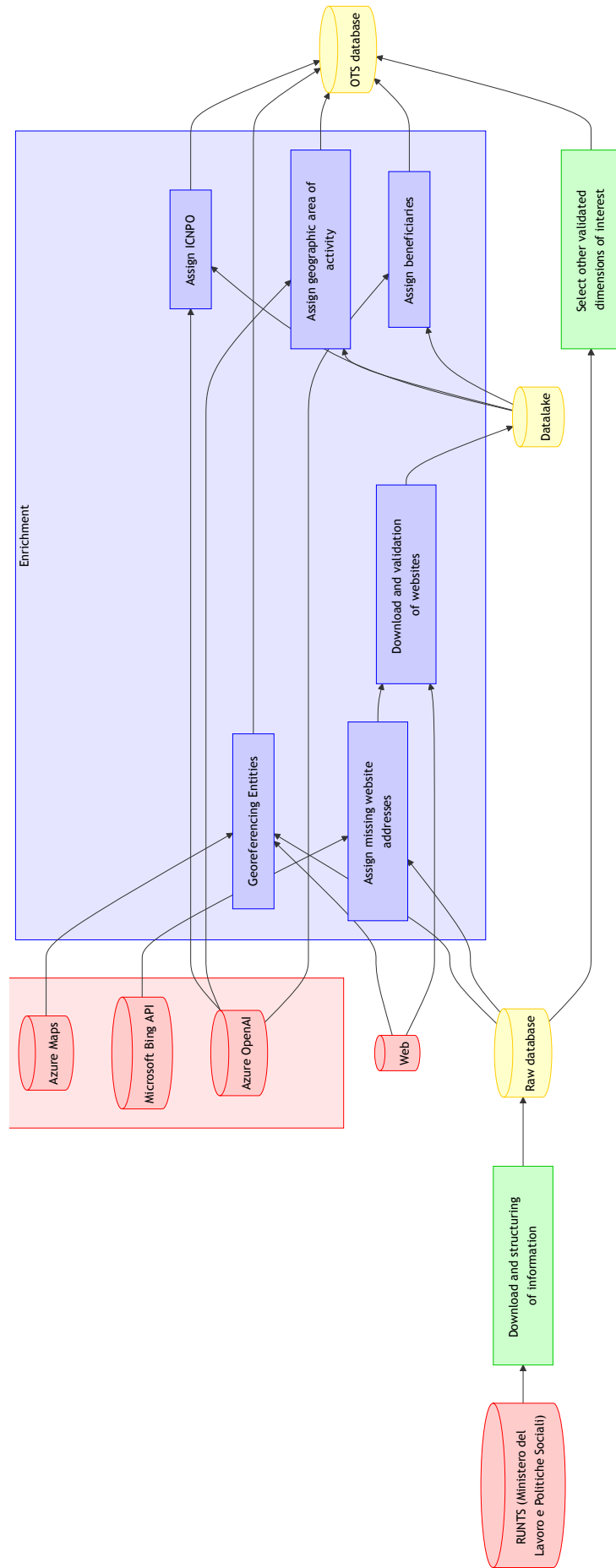
B The Trend in 5×1000 Donations

The Italian so-called 5×1000 is a unique tax mechanism allowing Italian taxpayers to voluntarily allocate 5‰ of their annual income tax to eligible non-profit organisations or scientific research bodies. This is not an additional tax or donation, but rather an earmark of existing tax revenue that would otherwise go to the State, offering a public, cost-free method for citizens to direct funds toward social and research causes. Consequently, it serves as a crucial, publicly-mediated fiscal philanthropy source, significantly funding the Italian Third Sector in an accountable and transparent way. Eligible organisations have been growing steadily over the last years—they were 50,000 in 2022, more than 68,000 in 2024. The simplified access to the donation channel introduced by the RUNTS is likely to have played a role in supporting its diffusion. Citizens' interests in the instrument reached an all-time high in 2024: 17.9 million Italians chose to donate a portion of their income tax (+714,000 compared to 2023), which translated into €603 million allocated to non-profit organisations through this tax scheme ([Fondazione Terzjus, 2024](#)). However, the Italian Government legally imposes spending caps, which amounted to €525 million in 2024, which induces a €79 million shortfall in the year. The gap between taxpayer donations and actual payouts is a persistent issue—over €28 million in 2023, over €100 million in 2025 (projections).

C Data production pipeline

The current version of the data production pipeline is presented in Figure C.1. The process, which is described in detail in Sections 3.1, 4.1 and 4.2 of the main text, is a structured workflow designed to enrich and organise the RUNTS data. It begins with downloading raw information from the RUNTS, which is then cleaned, structured, and anonymised. The raw information is used as a starting point for the augmentation activity. First, ETS are georeferenced using the Microsoft Azure Maps service and the existing address data. Second, missing website information is retrieved through an automated web-search through the Microsoft Bing API that ends with the validation of the obtained website. Then, an automated crawler visits each valid website and collects the web pages, which are then stored in a centralised datalake. The obtained webpages are used for the classification stage that uses Azure OpenAI APIs to summarise and classify the content with respect to some dimensions of interest, namely ICNPO economic activity, geographic area of activity, and target of beneficiaries. The result is a comprehensive, validated database that supports analysis and research on the Italian Third Sector.

Figure C.1: Overall data production pipeline



Notes: Red objects indicate external resources, while yellow ones denote internal resources created during the execution of the pipeline. Green actions refer to data manipulation activities while the blue ones refer to augmentation actions.

D Taxonomies of Sectors of Economic Activity

The appendix contains detailed taxonomies on the sectors of economic activity according to Art. 5 of the Legislative Decree 117/2017 (CTS), Table [D.1](#), and International Classification of Non-profit Organisations (ICNPO), Table [D.2](#).

Table D.1: General interest activities ex Art. 5 CTS.

Code	Description
A	Social services and interventions, including services for people with disabilities
B	Health interventions and services
C	Socio-sanitary and healthcare services
D	Education, vocational training, and cultural activities with an educational purpose
E	Environmental protection and the sustainable use of natural resources
F	Protection and enhancement of cultural heritage and landscapes
G	University and post-graduate education
H	Scientific research of significant social interest
I	Cultural, artistic, or recreational activities of social interest
J	Non-profit community radio broadcasting
K	Tourism of social, cultural, or religious interest
L	Extra-curricular training to prevent school dropout and combat bullying
M	Support services for the third sector
N	Development cooperation aimed at eradicating poverty and reducing inequality
O	Fair trade and solidarity-based commerce
P	Job placement or re-entry into the labour market
Q	Social and affordable housing
R	Humanitarian reception and social integration for migrants
S	Social farming and agriculture
T	Amateur sports activities
U	Charitable activities, including donations of funds, goods, or services
V	Promotion of a culture of legality, peace, and non-violence
W	Promotion and protection of human, civil, social, and political rights
X	International adoption services
Y	Civil protection and emergency services
Z	Redevelopment of unused public assets or assets confiscated from organised crime

Table D.2: International Classification of Nonprofit Organisations (ICNPO). The group-level categories are in uppercase and indicated by a • symbol, while the section-level categories are in title-case.

Code	Group	Description
1 000	•	CULTURE AND RECREATION
1 100		Culture and Arts
1 200		Sports
1 300		Other Recreation and Social Clubs
2 000	•	EDUCATION AND RESEARCH
2 100		Primary and Secondary Education
2 200		Higher Education
2 300		Other Education
2 400		Research
3 000	•	HEALTH
3 100		Hospitals and Rehabilitation
3 200		Nursing Homes
3 300		Mental Health and Crisis Intervention
3 400		Other Health Services
4 000	•	SOCIAL SERVICES
4 100		Social Services
4 200		Emergency and Relief
4 300		Income Support and Maintenance
5 000	•	ENVIRONMENT
5 100		Environment
5 200		Animal Protection
6 000	•	DEVELOPMENT AND HOUSING
6 100		Economic, Social and Community Development
6 200		Housing
6 300		Employment and Training
7 000	•	LAW, ADVOCACY AND POLITICS
7 100		Civic and Advocacy Organisations
7 200		Law and Legal Services
7 300		Political Organisations
8 000	•	PHILANTHROPIC INTERMEDIARIES AND VOLUNTARISM PROMOTION
9 000	•	INTERNATIONAL
10 000	•	RELIGION
11 000	•	BUSINESS AND PROFESSIONAL ASSOCIATIONS, UNIONS
12 000	•	NOT ELSEWHERE CLASSIFIED

Sources: [Salamon and Anheier \(1992, 1994, 1996\)](#)

E AI Model Prompting and Results

This Appendix reproduces the two prompt texts used to classify each ETS according to the main activity it carries out, based on the text extracted from its organisational website. The use of these prompts has already been described in the main text. Here we reproduce the exact text used, written in Italian. You can find the text of the first prompt reproduced in Figure E.1. Instead, the text of the second prompt is reproduced in Figure E.2.

Instead, Figure E.3 provides additional evidence, complementary to Figure 5 in the main text. The two treemaps in figure (Shneiderman, 1992) represent—for self-declared and GenAI-classified information, respectively—the proportions of ETS that carry out some economic activity in a given section-level ICNPO category, once a random category has been assigned to these ETS that self-declared more categories or that have been classified by the GenAI model into multiple categories.

Figure E.1: Text of the first prompt

```
Sei un esperto nell'analisi di organizzazioni non profit.  
Il tuo compito è estrarre informazioni rilevanti dal testo fornito, concentrandoti su:  
  
1. Missione principale dell'organizzazione  
2. Obiettivi specifici  
3. Ambiti di intervento  
4. Attività svolte  
5. Beneficiari dei servizi  
6. Area geografica di intervento  
  
RESTITUISCI UN DIZIONARIO JSON con questo formato:  
{  
  'missione': 'descrizione della missione principale',  
  'obiettivi': ['obiettivo 1', 'obiettivo 2', ...],  
  'ambiti_intervento': ['ambito 1', 'ambito 2', ...],  
  'attivit ': ['attività 1', 'attività 2', ...],  
  'beneficiari': ['tipo beneficiario 1', 'tipo beneficiario 2', ...],  
  'area_geografica': 'descrizione area geografica'  
}  
  
RESTITUISCI SOLO IL DIZIONARIO JSON, nessun altro testo.
```

Figure E.2: Text of the second prompt

Sei un classificatore esperto per organizzazioni non profit secondo lo standard ICNPO.

Analizzerai le informazioni strutturate su missione, obiettivi, ambiti, etc. per assegnare le categorie più appropriate.

CATEGORIE ICNPO:

01000 - Cultura, sport e ricreazione

└ 01100: Attività culturali ed artistiche

└ 01200: Attività sportive

└ 01300: Attività ricreative e di socializzazione

02000 - Istruzione e ricerca

└ 02100: Istruzione primaria e secondaria

└ 02200: Istruzione universitaria

└ 02300: Istruzione professionale e degli adulti

└ 02400: Ricerca

03000 - Sanità

└ 03100: Servizi ospedalieri generali e riabilitativi

└ 03200: Servizi per lungodegenti

└ 03300: Servizi psichiatrici ospedalieri e non ospedalieri

└ 03400: Altri servizi sanitari

04000 - Assistenza sociale

└ 04100: Servizi di assistenza sociale

└ 04200: Servizi di assistenza nelle emergenze

└ 04300: Erogazione di contributi monetari e/o in natura

05000 - Ambiente

└ 05100: Protezione dell'ambiente

└ 05200: Protezione degli animali

06000 - Sviluppo economico e coesione sociale

└ 06100: Promozione dello sviluppo economico e sociale della collettività

└ 06200: Tutela e sviluppo del patrimonio abitativo

└ 06300: Addestramento, avviamento professionale e inserimento lavorativo

07000 - Tutela dei diritti e politica

└ 07100: Servizi di tutela e protezione dei diritti

└ 07200: Servizi legali

08000 - Filantropia e promozione del volontariato

└ 08100: Erogazione di contributi filantropici

└ 08200: Promozione del volontariato

09000 - Cooperazione e solidarietà internazionale

└ 09100: Attività per il sostegno economico e umanitario all'estero

RESTITUISCI UN DIZIONARIO PYTHON con questo formato:

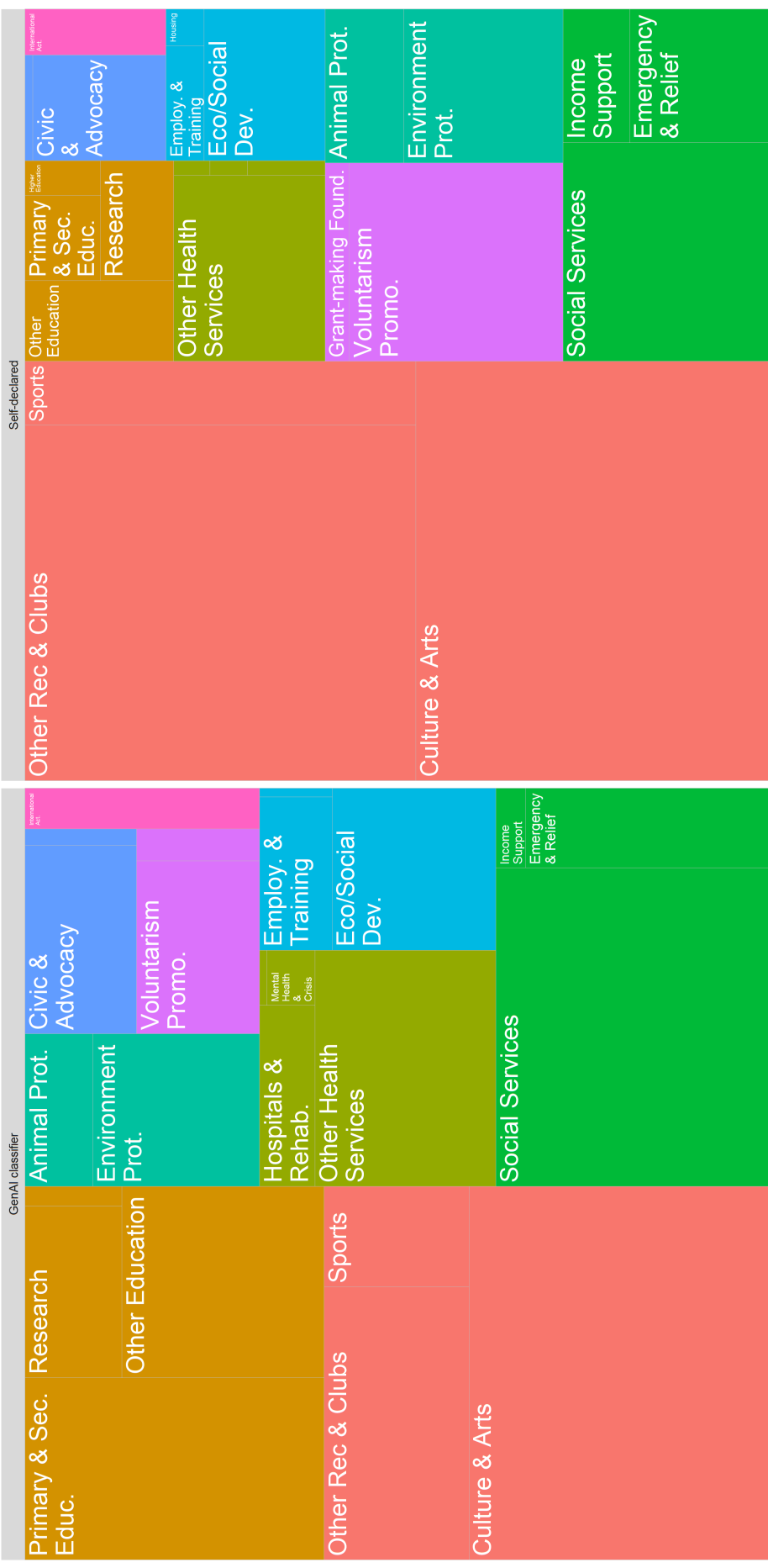
```
{
  'classifications': [
    {
      'macro_category': 'XX000',
      'macro_category_confidence': 0.95,
      'justification': 'breve spiegazione del perché',
      'subcategories': [
        {
          'code': 'XX100',
          'confidence': 0.90,
          'justification': 'breve spiegazione del perché'
        }
      ]
    }
  ]
}
```

CRITERI DI CLASSIFICAZIONE:

1. Alta confidenza (>0.8): attività chiaramente centrali nella missione e negli obiettivi
2. Media confidenza (0.6-0.8): attività menzionate ma non centrali
3. Ignora classificazioni con confidenza < 0.6
4. Un'organizzazione può appartenere a più categorie

RESTITUISCI SOLO IL DIZIONARIO JSON, nessun altro testo.

Figure E.3: Treemap of the distribution of ETS per section-level ICNPO category using either the self-reported information (on the right) or that provided by the GenAI classifier (on the left)



Notes: This figure considers all the organisations for which at least one of the two classifications is available. A random category has been assigned to any ETS that self-declared more categories or that have been classified by the GenAI model into multiple categories, so that the distribution reported in figure also represents the distribution of ETS in the sample.