CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17830

Augmenting the Italian Third Sector registry using non-profit organisations' websites

Carlo Bottai[1], **Francesco Trentini**[2,3,4], **Anna Velyka**[2]

[1]Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Italy, [2]Department of Statistics and Quantitative Methods (DiSMeQ), University of Milano-Bicocca, Italy, [3]Interuniversity Research Center for Public Services (CRISP), [4]Laboratorio "R. Revelli", Italy.

How to cite: Bottai, C.; Trentini, F.; Velyka, A. 2024. Augmenting the Italian Third Sector registry using non-profit organisations' websites. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17830

Abstract

This paper presents a framework for enriching and complementing administrative data from the Italian Third Sector Single National Register (RUNTS) with textual content extracted from the websites of the non-profit organisations listed in it. Through an automated web-scraping process we associate a website to each organisation and extract from its textual content information to describe the areas of the entity's actual economic activity. We develop a machine learning classifier to allocate each organisation into standardised categories of the International Classification of Non-profit Organisations. We further explore collected web data to identify other dimensions of non-profit operations. Enriching administrative registers with web data can yield trustworthy and detailed insights into the landscape of non-profit economic activities. Obtained results open up opportunities for further research of the labour market and economic development generated by the Third Sector, as well as comparative analysis with the sector of for-profit enterprises.

Keywords: *Third Sector; Administrative data enrichment; Big Data; Web scraping; NLP.*

1. Introduction

The non-profit sector contributes largely to the Italian economy. According to the latest figures, in 2021 more than 363,499 non-profit entities were active in the country, employing 870 thousand workers directly and activating 4.7 million volunteers, vis-à-vis 332,266 active enterprises with more than five employees, almost 23 million employees and 59 million inhabitants in the same period (Italian National Statistical Institute, 2023). They operate in a variety of economic sectors, ranging from social assistance and healthcare to social cohesion, sports and cultural activities.

Nonprofit entities are of great interest since they operate following market principles and an efficient use of production factors, while not pursuing profit. These enterprises embrace a community-oriented model, which aims at keeping the surplus within the entities that generated it, fostering both investment and improvement in social and economic conditions in the context in which they are embedded (Evers & Laville, 2004).

In Italy, the discipline of the non-profit sector has been reviewed in 2017, with the Legislative Decree 117/2017, known as Codice del Terzo Settore (hereon, the Code). The Code institutes seven main categories of Third Sector Entities (Enti del Terzo Settore, ETS hereon): volunteer organisations, social promotion associations, social enterprises, philanthropic entities, networks of entities, mutual aid societies, and other entities. The Code enumerates the set of activities of general interest that every ETS must primarily pursue. The qualification as ETS introduces requirements in terms of governance and transparency of their accounts; at the same time, they benefit from special tax exemptions, they can sign partnerships with public entities and receive public funds. Among the most important innovations introduced by the Code is the establishment of a national electronic register of ETS, the Registro Unico Nazionale del Terzo Settore (RUNTS), owned by the Ministry of Labour and Social Policies, that substituted more than 350 local registries. An entity listed in the RUNTS gains the ETS title and acquires rights and obligations described by the law. The RUNTS is publicly available through a web interface (https://servizi.lavoro.gov.it/runts/) and includes a long list of characteristics of each ETS.

Our work leverages this new administrative source to establish a framework for data enrichment using entities' websites. In recent years a similar framework has been developed and widely applied to for-profit enterprises (see e.g., Daas & van der Doef, 2020; Kinne & Axenbeck, 2020; Crosato et al., 2023; Bottai et al., 2022; Abbasiharofteh et al. 2023). Indeed, nowadays firms commonly use corporate websites to reach customers with information about their products and business activities. Therefore, it is possible to infer the economic activity of a firm by looking at the content of its corporate websites, at least to some extent (Domènech et al., 2012). Non-profit organisations as well largely rely on their own website to communicate with their stakeholders. Consequently, these websites can be leveraged as a—easily accessible and constantly updated—information source to enrich the—already available and trustable—administrative sources. However, such kinds of unstructured sources require some careful clenching and pre-processing to mine useful and meaningful information from them (see e.g., Daas et al. 2015; Rammer & Es-Sadki, 2023). The rest of this work describes this process in detail. Then, some preliminary results and future developments are discussed.

2. Data and methods

The RUNTS (the register, hereafter) is operated by the Italian Ministry of Labour and Social Policies (MLPS).

2.1. The master database: the RUNTS

The register contains information about 123,221 ETSs, seventy-five per cent of which are either *associazioni di promozione sociale* (APS) or *organizzazioni di volontariato* (ODV), both voluntary organisations—the former carrying out activities primarily for the benefit of third parties, while the activities of the latter may be primarily for the benefit of their own members. For each entity, we know its tax code (*codice fiscale*), business name, legal type, headquarter's municipality, the date of enrolment in the register, and the generalities of the legal representative.

2.2. External data collection

The aim of this work is to enrich the information available on the RUNTS by exploiting textual content extracted from the ETSs' websites. Therefore, we start collecting the business name, tax code, and legal form of each entity as recorded in the register. After some minimal cleaning pre-processing of these strings, we search on Microsoft's Bing web search engine for queries combining these information pieces—like *("amici degli animali" OR 97324990385) APS*—, in a similar fashion to van Delden et al. (2019). For each ETS, we collect up to ten results returned by the search engine. We extract and preserve only the web domain of each result and we exclude those that are linked to multiple ETSs. Then, for each ETS we loop over the remaining web domains, and we scrape a few pages of this domain hoping to find at least one information piece that can verify the matching between the web domain detected and the ETS under scrutiny; see Barcaroli et al. (2016) and Bottai et al. (2022) for a similar approach. Specifically, we search on the home page of the website as well as on pages like "about us" or "contacts", and we consider as positively assigned these websites naming at least one among the tax code, headquarter's county (*provincia*), or full name of the legal person of the ETS linked with that website. While looping over the domains retrieved on Bing, we give priority to those with high textual similarity to the ETS's business name and to these high in the search engine's results. We stop the loop at the first positive match.[1]

To assess the quality of this procedure, we extracted 119 ETSs at random, representative of the population about the geographical area (NUTS 2) and legal type (APS, ODV, social enterprise, etc.). By searching on Bing for this list of entities, we detected at least a web domain for 103 of them (86.6%). Of these, we selected the first result returned by the search engine (we plan to extend this step to include all the first ten results that we got from the search engine) and we scraped a few pages of the website. The scraper has been able to retrieve the website of 81 ETSs (78.6%). Of each of these, we further verified the accuracy of the matching between the ETS in question and the website detected. We developed an automatic classifier that classified each of

[1] The authors are available to provide further information details about these steps.

the 81 cases as either a positive or negative matching. The classifier returned 35 positive and 46 negative matchings. We manually inspected the same matches and verified 31 true positives and 39 true negatives. Therefore, the classifier shows high *sensitivity* (81.6%) and *specificity* (90.7%).

Even though further development is needed, we believe that these preliminary results are encouraging. Most of the improvement must be obtained by improving the quality of the results obtained from the search engine. We believe that, by extending the number of results considered from one to ten (as already mentioned), we will be able to obtain a significant improvement in this respect.

2.3. Classification of economic activity

The RUNTS provides information concerning the main fields of activity of each ETS, according to the International Classification of Non-Profit Organisations (ICNPO) and Article 5 of Legislative Decree 117/2017. Nonetheless, the information is self-declared and not compulsory. Self-declaration may induce losses in precision, as ETSs may declare a smaller number of areas of activity, understating the actual scope of their activities, or be led to declare only those activities that are closer to their statutory mission. For this reason, ETS' web pages are a rich source of information on the actual activities put in place by the ETSs.

A number of classification systems exist at national and international levels to categorise non-profit sector organisations depending on the scope of their activities. At the national level, classifications often coincide with government frameworks and classify ETS based on their primary purpose and economic activities (like the NACE in the EU). The United Nations has developed the International Classification of Non-Profit Organisations (ICNPO) that provides standards and taxonomies specifically for non-profit organisations based on their activities, primary purpose, and mission.

For statistical accounting, the Italian National Institute of Statistics categorises ETS operating in Italy according to ATECO (the Italian version of the NACE). Since the ATECO hasn't been developed specifically for the non-profit sector, it doesn't provide enough categories to account for all the possible range of activities of these organisations (12 macro categories of ICNPO correspond to only 5 macro categories of ATECO). Instead, the ICNPO classification has 12 macro and 39 subcategories. In addition to being recognised as meeting the OECD and UN international standards, this classification provides a brief description of the organisations that fall under each subcategory (United Nations, 2018).

Several studies in the past have investigated the possibility of automatically identifying an appropriate category of non-governmental organisations. These studies used both international classifications (e.g., for identifying appropriate ICNPO category for 5,000 Austrian non-profit organisations; see Litofcenko et al., 2020) and national classifications (e.g., remapping the US nonprofit sector by reassigning multiple NTEE codes to organisations with purposes across various domains; see Ma, 2021). The results of this research have been subsequently used to develop a national classification of charity organisations in the United Kingdom (UK Charitable Activity Tags) and an automated system for determining the appropriate category for 4,200 organisations (Damm & Kane, 2022).

Methods that were previously applied to analyse and classify textual data on ETS can be summarised into three categories: rule-based dictionary approach, supervised learning, and unsupervised learning. Data enrichment techniques proposed in this paper allow us to overcome the limitations of previous studies, providing more information about the ETSs activities. To deploy a trained classifier to automatically assign ICNPO category to new ETS based on the text information extracted from their websites we will apply Natural Language Processing (NLP) and test several supervised learning techniques in the following steps:

- ➢ **Pre-processing** of text data collected from websites of non-profit organisations: clean the text data by removing punctuation, stopwords, and HTML tags; tokenise the text into individual words or phrases.
- ➢ **Feature Extraction** from the pre-processed data: convert text into numerical features to be used by machine learning algorithms: bag-of-words and word embedding.
- ➢ **Classification Category Description Parsing**: parse the short descriptions of each ICNPO category to extract keywords that are indicative of the category's focus area.
- ➢ **Model Training**: the information about ICNPO classification available in RUNTS for some ETSs will serve as a basis for training the model. We will train a classifier to map the numerical representations of textual data to predefined ICNPO categories based on labelled training data (Naive Bayes, Support Vector Machines, Decision Trees, and Convolutional Neural Networks, Bidirectional Encoder Representations from Transformers).
- ➢ **Model Evaluation**: some ETS already classified into one of the ICNPO classes will be used as a test set to evaluate the models' performance. A stratified split will help to ensure the proportional representation of the data. Consequently, several metrics will be calculated to demonstrate the effectiveness of the suggested classification model. The accuracy metric will reveal the overall model's performance by denoting the proportion of correctly classified ETS out of the total number of ETS from the test set. Since we expect that not all the ICNPO classes are equally represented, we will also employ other performance metrics. Precision will demonstrate the accuracy of the model in correctly assigning ICNPO classes to companies based on their web text data. Recall metric will indicate the model's ability to capture all ETS of a particular ICNPO class. Finally, a harmonic mean of precision and recall metric (F-1 score) will denote the overall reliability of our model.

2.4. Classification of other dimensions

The information concerning the sectors of activity in which each ETS is involved is the primary dimension of interest to integrate the information available through the administrative register.

In fact, other dimensions can be approached using website data. Other dimensions are the beneficiaries of their activities, the type of projects developed by each ETS, the stakeholders whom they are in relation with, and the funding they receive. We are testing the use of the websites to address these dimensions and extract useful information. The use of web data has the additional advantage of storing pages to run offline analysis, so that they can be queried in subsequent moments over dimensions that were not foreseen in the initial stages of the work.

3. Discussion and future work

The use of web data for the production of both fresh statistics or for the enrichment of administrative sources has reached a stage of maturity and applied in a large set of cases, including the production of official statistics, introducing changes in the production process, in term of theoretical frameworks and required competences (Pfeffermann, 2015; Ricciato, Wirthmann & Hahn, 2020). It is especially important to consider the data generating process of web data is complex. Agents produce data with a multitude of schemes and formats, usually not standardised. Moreover, the processes generating these data are variable, too. Therefore, a lot of attention is needed to understand the veracity of these data, that is, which phenomena these data represent. The opportunity presented by these unique data characteristics is associated with the central challenge of interpreting the content of the recorded data and the correct way to treat and process them. Administrative registers are therefore a valuable starting point, given that they provide the universe of observations concerning specific social facts, generated by law provisions. The RUNTS collects the universe of Third Sector Entities in Italy and enriching it by means of web data can provide trustable, timely and fine-grained information on their operations. The ETSs are of great interest given that they provide essential services to a large part of the population and complement publicly provided services with a diverse business model.

Having a timely and precise knowledge of the ETSs' area of activity and types of interventions is pivotal to coordinate interventions in areas of general interest through partnerships or agreements with the public administration, which is also one of the pillars of the 2017 reform. It also allows us to analyse the social capital of local communities and study in depth the relationship of this last with economic development. Furthermore, the abundance of individual information in the RUNTS, such as the tax code and the names of the legal representatives, makes the database virtually linkable to a huge variety of data sources. Of particular interest are the enterprise characteristics and performance, on the business demography side, and, on the labour market side, the volume and quality of employment generated by the Third Sector. Possible extensions would also include the analysis of the hyperlinks connecting the ETSs' websites to each other (*webometrics*), e.g., to analyse the *networks of ETSs*, a kind of ETS regulated by the RUNTS, or to study the business relationship among ETSs (see e.g., Vaughan et al. 2006 and Abbasiharofteh et al. 2023). At the same time, it allows running a comparative analysis with the Second Sector, i.e., for-profit enterprises, in terms of generated employment and their business demography and potentially filling a gap in the national statistics, that have a hard time at properly representing such an informal sector by only relying on conventional information sources.

Acknowledgements

Carlo Bottai and Francesco Trentini thank the University of Milano-Bicocca for sponsoring this work under the 'Bicocca Starting Grants 2023–2024' funding schema.

References

Abbasiharofteh, M., Kinne, J., & Krüger, M. (2023). Leveraging the digital layer: the strength of weak and strong ties in bridging geographic and cognitive distances. *Journal of Economic Geography*, lbad037. doi:10.1093/jeg/lbad037

Barcaroli, G., Scannapieco, M., & 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.

Bottai, C., Crosato, L., Domènech, J., Guerzoni, M., & 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* (pp. 338–344). New York, NY: Association for Computing Machinery. doi:10.1145/3524458.3547246.

Crosato, L., Domènech, J., Liberati, C. (2023). Websites' data: a new asset for enhancing credit risk modeling. *Annals of Operations Research*, 1–16. doi:10.1007/s10479-023-05306-5

Daas, P.J.H., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, 36(4), 1239–1251. doi:10.3233/SJI-200627

Daas, P.J.H., Puts, M.J., Buelens, B., & van den Hurk, P.A.M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, 31(2), 249–262. doi:10.1515/jos-2015-0016

Damm, C., & Kane, D. (2022). Classifying UK charities' activities by charitable cause: a new classification system. *Voluntary Sector Review*, 1–27.

van Delden, A., Windmeijer, D., ten Bosch, O. (2019). Searching for business websites. Discussion Paper. Statistics Netherlands (CBS).

Domènech, J., de la Ossa, B., Pont, A., Gil, J.A., Martinez, M., & 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* (vol. 1, pp. 573–578). Washington, DC: IEEE Computer Society. doi:10.1109/WI-IAT.2012.92

Evers, A., & Laville, J.-L. (Eds.) (2004). *The Third Sector in Europe*. Edward Elgar. doi:10.4337/9781843769774

Italian Ministry of Labour and Social Policies (2024). *Registro Unico Nazionale del Terzo Settore* [Data set]. Retrieved from https://servizi.lavoro.gov.it/runts/ on 2024-03-14.

Italian National Institute of Statistics (2023) Censimento Permanente delle Istituzioni Non Profit.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, 125(3), 2011–2041. doi:10.1007/s11192-020-03726-9

Litofcenko, J., Karner, D., & 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.

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.

Pfeffermann, D. (2015) Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture, *Journal of Survey Statistics and Methodology,* 3(4), 425–483. doi:10.1093/jssam/smv035

Rammer, C., & Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators - a literature review. *Technological Forecasting and Social Change*, 197, 122874. doi:10.1016/j.techfore.2023.122874

Ricciato F., Wirthmann A., & Hahn M. (2020) Trusted Smart Statistics: How new data will change official statistics. *Data & Policy*, 2, e7. doi:10.1017/dap.2020.7

United Nations (2018). *Satellite Account on Non-profit and Related Institutions and Volunteer Work* [Data set]. Retrieved from https://unstats.un.org/unsd/nationalaccount/docs/UN_TSE_HB_FNL_web.pdf

Vaughan L., Gao Y., & Kipp M. (2006) Why are Hyperlinks to Business Websites Created: A Content Analysis. *Scientometrics*, 67(2), 291–300. doi:10.1007/s11192-006-0100-6