CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.22013

Unveiling the Structure and Strategy of the EU Cluster Ecosystem Using Web Data

Marcos Ranchal[1], Ana Debón[1], Josep Domenech[1], Francisco Puig[2]

[1]eSMART investigation center, Universitat Politècnica de Valencia, Spain, [2]Department of Management, Universitat de València, Spain

How to cite: Ranchal, M.; Debón, A.; Domenech, J.; Puig, F. (2026). Unveiling the Structure and Strategy of the EU Cluster Ecosystem Using Web Data. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.* [To be completed by the publisher]

Abstract

How can web data reveal the organisational model of the European cluster ecosystem? This paper presents a computational social science framework to measure the structure and strategy of the EU cluster ecosystem using internet data from the European Cluster Collaboration Platform (ECCP). We implement an end-to-end pipeline with web scraping, semantic normalisation, feature engineering, and multivariate analysis. From ECCP profile pages, we build a structured dataset with 1,673 cluster records and 162 variables covering governance, membership structure, sector specialisation, collaboration interests, and digital and green orientation signals. The methodological contribution combines robust HTMLextraction heuristics, reproducible data cleaning in R, one-hot encoding of multivalue fields, and interpretable variable construction for dimensionality reduction. Results provide descriptive evidence on coverage, territorial distribution, and strategic profiles, and define a PCA-ready analytical space to identify latent patterns of cluster organisation. The workflow is auditable and directly aligned with current internet-data research in economics and social sciences.

Keywords: *EU clusters; cluster organisations, cluster policy, ECCP; web scraping; internet data; computational social science; PCA.*

1. Introduction

The CARMA 2026 call highlights the growing relevance of internet data, web scraping, and computational methods for research in economics and the social sciences. This paper contributes to that agenda by transforming heterogeneous web profiles from the European Cluster Collaboration Platform (ECCP) into a reproducible analytical dataset for studying cluster organisations and their strategic behaviour across Europe. Although cluster platforms contain rich information on governance, sector focus, collaboration dynamics, and digital

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

transformation, these data are typically weakly structured and difficult to use for systematic empirical analysis without robust extraction and normalisation procedures.

This approach builds on the literature on industrial clusters and cluster policy, which emphasizes the role of localized collaboration and collective competitiveness in regional development (Porter, 1998; Puig et al., 2023; Sölvell et al., 2003). In the increasingly institutionalised European cluster ecosystem, cluster organisations (COs) play a key bridging role by coordinating cooperation, managing collaborative projects, and connecting firms with public policies and innovation systems (Morgulis-Yakushev and Sölvell, 2017). In this context, digital platforms such as the ECCP provide a valuable but still underutilised source of organisational data (Kramer et al., 2024). Accordingly, this paper pursues two objectives: first, to develop a transparent and auditable data pipeline that extracts and structures information from ECCP cluster profiles; and second, to propose an analytical framework capable of identifying strategic and organisational patterns across cluster organisations using derived quantitative features.

2. Background

Economic literature has long documented the tendency of firms and activities to agglomerate geographically, forming industrial clusters that generate competitive advantages through knowledge spillovers, collective learning, and dense collaboration networks. Firms located within clusters often outperform non-cluster firms in terms of innovation, internationalisation, and survival, while clusters themselves contribute to regional economic development (Puig et al., 2023). These insights have inspired cluster policies to foster collaboration and competitiveness across firms and territories (Porter, 1998).

Evolving from national manufacturing initiatives to broader European policies across industries and services, such policies expanded geographically and sectorally (Sölvell et al., 2003; Sopoligová and Pavelková, 2017). This process has contributed to the institutionalisation of cluster dynamics and the emergence of a European cluster ecosystem in which cluster organisations coordinate cooperation, manage collaborative projects, and act as intermediaries between firms, knowledge institutions, and public authorities, a role described as bridging capacity (Morgulis-Yakushev and Sölvell, 2017).

Recent evidence suggests that European programmes such as COSME and Horizon 2020 increasingly favour larger and more consolidated cluster organisations (Kramer et al., 2024). From the perspective of regional diversification and industrial relatedness (Neffke et al., 2011), this trend may generate risks such as path dependency, limited scalability for emerging clusters, organisational homogenisation, and misalignment between institutional cluster structures and territorial productive dynamics.

Article in Press

Against this background, this paper analyses the structure and strategy of organisations shaping the European cluster ecosystem. The empirical analysis draws on 1,673 cluster organisations registered in the European Cluster Collaboration Platform (ECCP), a voluntary and open-access platform that facilitates international visibility, inter-cluster cooperation, and participation in European initiatives. Its public accessibility enables reproducible data collection and the construction of a structured dataset of cluster organisations, which we analyse using web-scraping techniques to identify patterns in their structural characteristics and strategic activities.

3. Methodology

3.1. Data Source

The study uses ECCP cluster profile pages as the primary source. Each observation corresponds to one profile page. The full clean dataset contains 1,673 observations and 162 variables (all with extraction status ok). For the analytical sample used in this paper, we retain only EU27 country profiles, resulting in 1,286 observations.

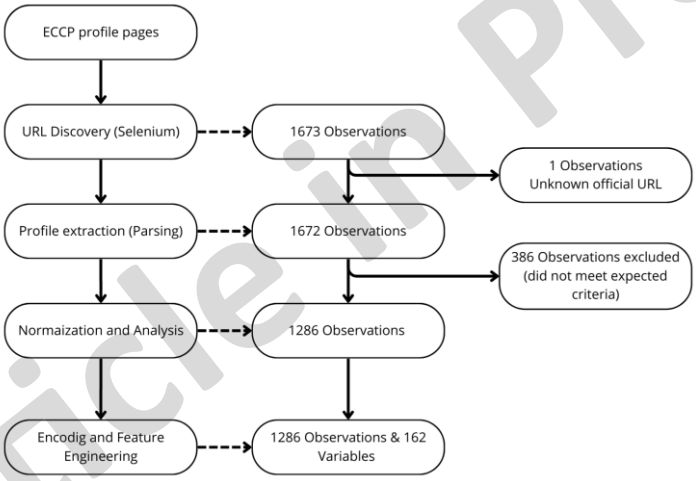

Figure 1. ECCP data pipeline, including EU27 country filtering for the analytical sample.

3.2. Data pipeline

The pipeline was designed for heterogeneous page structures and dynamic content. It follows two sequential phases: (i) URL discovery on ECCP listing pages using Selenium; and (ii) profile-level extraction with a parser that combines multiple HTML heuristics (h1, dt/dd, table cells, Drupal blocks, and label-value patterns). The implementation supports both Selenium and parallelized requests modes with thread pools for scalable retries. After extraction, fields are normalised into canonical variables (e.g., cluster_name, country, year_established,

Article in Press

sector_nace_2, services_tags). A country filter is then applied to retain only EU27 observations for the analytical sample. Multivalue fields are split and one-hot encoded for languages, membership types, themes, collaboration interests, industrial ecosystems, and service tags. Additional indicators are derived for exploratory and dimensionality-reduction analysis, including size proxies, composition shares, collaboration breadth, diversification, and strategic orientation variables. The full workflow is summarised in Figure 1.

3.3. Statistical methods

Principal component analysis (PCA) is applied to a numerical matrix built from cleaned and engineered indicators. To reduce sparsity, the analysis prioritises continuous and aggregated variables instead of full set of one-hot dummies. Preprocessing includes missingness filtering, median imputation, zero-variance removal, and standardisation. Components are retained using explained-variance and scree-plot criteria, then interpreted through loadings and scores (Jolliffe and Cadima, 2016). This preprocessing strategy follows standard feature engineering and applied data science practice (Kuhn and Johnson, 2019; Provost and Fawcett, 2013).

A Classification and Regression Tree (CART) model is also applied to the cleaned and engineered indicators to address a separate classification objective. The model predicts the European macro-region of each cluster organisation, using the United Nations M49 classification for Western, Eastern, Northern, and Southern Europe (United Nations Statistics Division, 1999). Its purpose is to characterise the organisational profiles associated with each European area by identifying which structural and strategic variables contribute most to their differentiation. CART is selected because it provides an interpretable supervised method based on explicit decision rules (Breiman et al., 2017), making it suitable for analysing territorial variation in the European cluster ecosystem.

4. Results

The raw scraped dataset covers 64 countries in profile locations. For the analytical sample, the country criterion retains only EU27 profiles. Table 1 summarises the ten most frequent country codes in the EU27 analytical sample. The following coverage rates are computed on raw scraped fields prior to cleaning and EU27 filtering. Coverage is high for core structural fields: about (98.9%), `members_total` (99.3%), `country` (99.4%), and `entity_type` (91.2%). Coverage is moderate for `sector_nace_2` (62.6%), `collaboration_interests` (62.5%), and `services_tags` (61.7%), and lower for `industrial ecosystems` (39.5%), `digital transformation` (24.6%), and `EU initiatives` (17.0%). Strategic orientation signals remain substantial: 58.3% of profiles are tagged as digitally oriented and 44.6% as green oriented.

Article in Press

Table 1. Top 10 country codes in EU27 ECCP profiles

Country code	Count	Country code	Count
ES	207	RO	69
DE	143	SE	47
FR	125	BE	46
IT	117	FI	45
PL	82	CZ	42

For the PCA analysis, Figure 2 presents the scree plot, based on which a model retaining three principal components is selected, explaining approximately 47.6% of the total variance. The scree pattern shows a clear inflection point after the third component, supporting the retention of these three dimensions. Figure 3 displays the variable map on the first two dimensions, while Figure 4 presents dimensions 2 and 3.

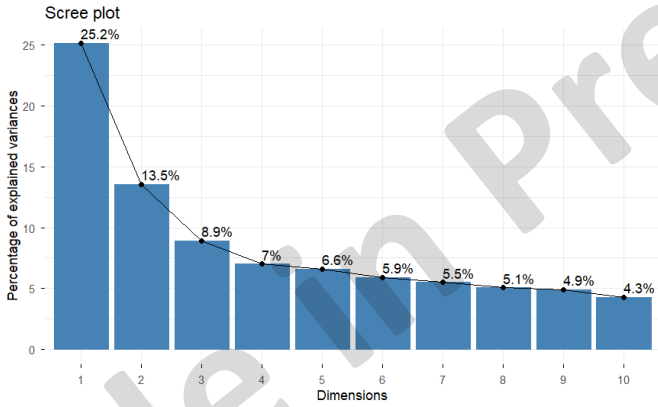

Figure 2. Scree plot for PCA component selection.

The retained components can be interpreted as three underlying dimensions structuring variation across cluster organisations, as reflected in the variable configurations shown in Figure 3 and Figure 4.

Article in Press

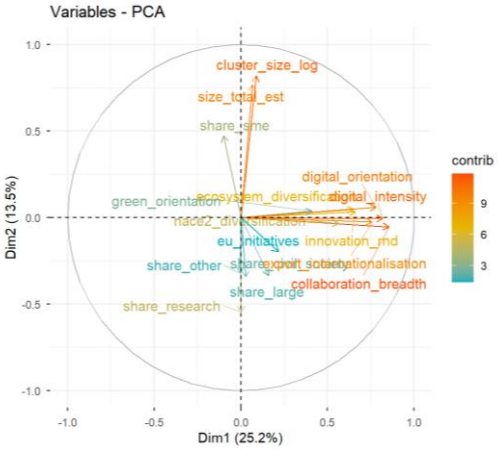

Figure 3. PCA variable map (Dim.1 vs Dim.2).

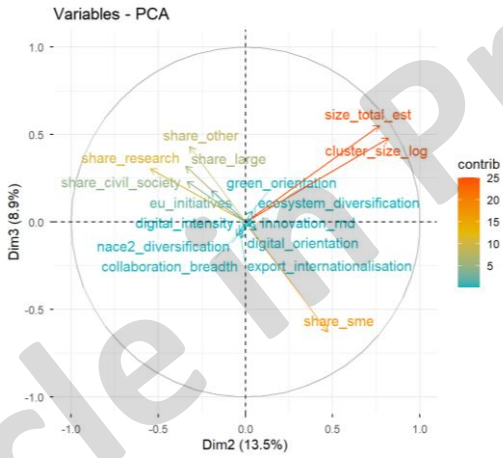

Figure 4. PCA variable map (Dim.2 vs Dim.3).

Dimension 1 (strategic openness and capability breadth). Dim.1 is driven by strong posi tive loadings in collaboration breadth, digital intensity/orientation, export internationalisation, ecosystem diversification, innovation/R&D, and NACE-2 diversification. This dimension sepa rates cluster organisations with broad, outward-oriented and digitally enabled strategies from organisations with narrower strategic portfolios.

Dimension 2 (scale and SME-centered structure). Dim.2 is dominated by size proxies (cluster_size_log, size_total_est) and membership composition. The positive side combines larger scale and stronger SME share, while the opposite side is associated with relatively higher non-SME shares (research, large firms, civil society, and other members).

Article in Press

Dimension 3 (SME specialisation versus mixed composition at scale). Dim.3 shows a marked contrast between share_sme (negative) and scale/mixed-composition variables (positive), especially total size, large-firm share, research share, civil-society share, and other-member share.

Taken together, the first three dimensions provide a coherent typology space: Dim.1 captures strategic sophistication and openness, Dim.2 captures scale and SME intensity, and Dim.3 captures SMEconcentration versus mixed scale. As usual in PCA, sign orientation is arbitrary, but relative distances and variable groupings are stable and interpretable.

The CART classification model was used to predict the United Nations macro-region of each cluster profile within Europe: Western, Eastern, Northern, and Southern Europe. Figure 5 presents the dominant predicted class by NUTS2 region, obtained by aggregating the model predictions territorially.

The map suggests a readable territorial pattern in the organisational profiles of European clusters. Western Europe is associated with larger and more consolidated organisational profiles, while Eastern Europe is associated with smaller and younger structures. Northern and Southern Europe appear in more specific combinations of maturity, services, and collaboration variables. Thus, the CART results indicate that the engineered organisational indicators capture territorial differences in the European cluster ecosystem.

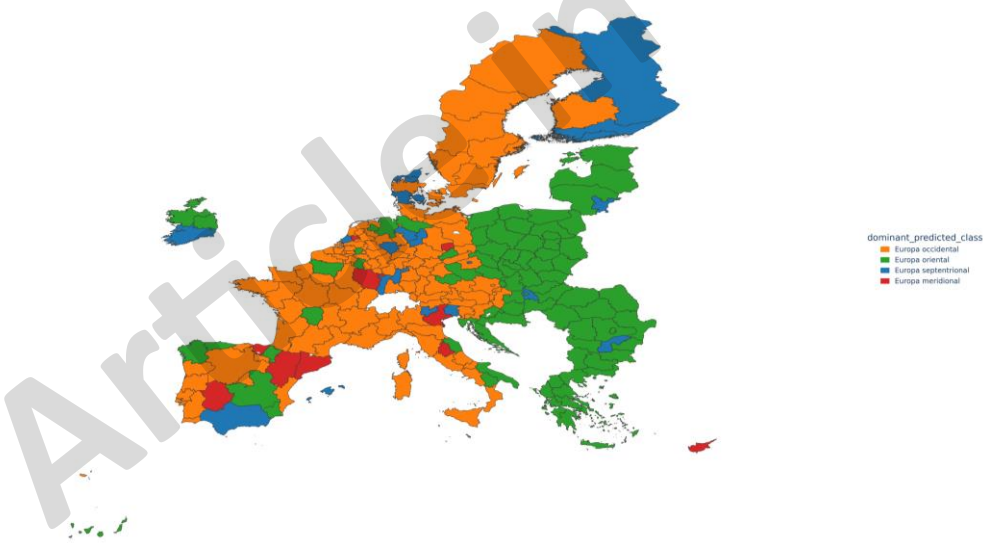

Figure 5. Dominant predicted UN macro-region class by NUTS2 region, based on the supervised multiclass model.

Article in Press

5. Discussion and Conclusions

The ECCP project demonstrates that internet-native organisational data can be transformed into a reliable empirical asset for social science research when reproducibility and semantic normalisation are treated as first-order methodological constraints. By systematically extracting and structuring information from ECCP cluster profiles, the study shows how web-based sources can support rigorous empirical analysis of organisational ecosystems.

A central contribution of the project is the development of a set of standardised indicators capturing membership composition, sectoral diversification, collaboration breadth, and strategic orientations related to digitalisation, innovation, internationalisation, and sustainability. These indicators allow us to address the guiding research question of this study: how web data can reveal the organisational model of the European cluster ecosystem. In particular, the resulting feature space enables the identification of underlying structural dimensions that differentiate clusters by strategic openness, organisational scale, and membership composition.

The empirical results provide evidence that web data can indeed reveal key features of the organisational model of the European cluster ecosystem. The analysis identifies two main dimensions structuring the observed variation: a structural dimension related to cluster size and membership composition, and a strategic dimension associated with innovation, digitalisation, and internationalisation. The relatively weak association between these dimensions suggests a certain decoupling between organisational structure and strategic orientation, pointing to the existence of heterogeneous organisational models within the European cluster ecosystem.

From a substantive perspective, the pipeline enables comparative analysis of how EU clusters combine membership structures, sector specialisation, and strategic orientations in the context of digital and green transitions. From a methodological perspective, it provides a reproducible template for transforming semi-structured web data into structured datasets suitable for statistical analysis and machine learning.

The approach also closely aligns with the CARMA framework's priorities in five key areas: web scraping infrastructure, data quality control, computational text-to-structure conversion, the construction of machine-learning-compatible feature spaces, and the development of policy relevant measurement tools for economics and social science research.

Overall, the ECCP project can therefore be understood both as an empirical contribution to the study of the European cluster ecosystem and as a methodological contribution to internet data research. Its main outcome is a reproducible pipeline that converts web-based organisational profiles into structured indicators enabling the analysis of cluster organisational strategies.

Additionally, the indicator framework developed in this study will allow the characterisation of regional organisational models within the European cluster ecosystem, enabling comparative

Article in Press

analysis of Northern, Southern, Eastern, and Western European cluster patterns based on differences in membership composition, strategic orientation, and collaboration structures.

References

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.

Kramer, J.-P., Galdiga, L., Schmidt, F., Ginzinger, F., Layher, J., & Vogelsang, V. (2024). *European cluster panorama report 2024* [Technical report]. European Cluster Collaboration Platform (ECCP), EISMEA, European Commission.

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Morgulis-Yakushev, S., & Sölvell, Ö. (2017). Enhancing dynamism in clusters: A model for evaluating cluster organizations' bridge-building activities across cluster gaps. *Competitiveness Review*, 27(2), 98–112.

Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic Geography*, 87(3), 237–265.

Porter, M. E. (1998). Clusters and the new economics of competition. *Harvard Business Review*, 76(6), 77–90.

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.

Puig, F., Debón, A., Cantarero, S., & Marques, H. (2023). Location, profitability, and international trade liberalization in European textile-clothing firms. *Economic Modelling*, 129, 106563.

Sölvell, Ö., Lindqvist, G., & Ketels, C. (2003). *The Cluster Initiative Greenbook*. Ivory Tower.

Sopoligová, M., & Pavelková, D. (2017). Cluster policy in Europe and Asia: A comparison using selected cluster policy characteristics. *Journal of International Studies*, 10(4), 35–52.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Chapman and Hall/CRC.

United Nations Statistics Division. (1999). Standard country or area codes for statistical use (rev. 4). Series M: Miscellaneous Statistical Papers, No. 49, New York, United Nations Division. From https://unstats.un.org/unsd/methodology/m49/

Article in Press