Electronic Markets (2025) 35:92
https://doi.org/10.1007/s12525-025-00831-6

RESEARCH PAPER

Understanding data collaboratives ten years after their definition: Distinctive features, impacts and research priorities

Federico Bartolomucci[1] · **Gianluca Bresolin**[1]

Received: 19 March 2024 / Accepted: 29 August 2025

Abstract
The use of data for social good has received increasing attention from institutions, practitioners and academics in recent years. Data collaboratives are cross-sectoral partnerships that aim to foster the use of data for societal purposes. However, the proliferation of initiatives on the topic of data sharing has created confusion regarding their nature and scope. To advance research on the topic, using existing literature, this paper offers a refinement of the concept of data collaboratives ten years after their first definition. This enables the distinction between data collaboratives and other forms of initiatives such as open platforms and data ecosystems. Through the analysis of a dataset of 171 data collaboratives, the paper proposes an enhanced categorisation that identifies five clusters of data collaboratives. Each cluster is described with a focus on its individual characteristics and development challenges. The holistic approach adopted and the maturity of the field allowed us to gain valuable insights into the domains and scopes that these types of partnership may serve and their potential impact. The results highlight the heterogeneity of initiatives falling under the concept of data collaboratives and the necessity to address their development challenges by either concentrating on a specific cluster or conducting comparative and horizontal studies. These findings also enable comparability and improve the identification of benchmarks, which is a valuable resource for the development of the field.

Keywords Data collaboratives · Data sharing · Data ecosystems · Cluster analysis

JEL Classification O33 · O35 · P33

Introduction

The use of big data for public benefit has gained increasing momentum in recent years; however, the practice is still in its preliminary stages, with many technical, organisational and ethical constraints limiting its diffusion. Numerous pioneering applications across diverse fields such as transportation management (Williams, 2020), migration (Rango & Vespe, 2017) and urban development (Farmer et al., 2022) have demonstrated the effectiveness of data-for-good initiatives in designing evidence-based policies, gaining new insights into social phenomena, fostering innovation and increasing citizen participation in the management of public resources (Verhulst & Young, 2016). This newfound awareness has prompted multilateral efforts advocating the establishment of data-sharing standards and the mobilisation of data to tackle particular data-related challenges. However, despite the numerous initiatives put in place internationally, the use of data for social good has so far been limited (Flanagan Anne & Sheila, 2022). Implementing data-for-social-good projects represents a complex socio-technical challenge (Jussen et al., 2024; Liva et al., 2023) that often requires the involvement of diverse actors (Coulton et al., 2015), the implementation of complex technological solutions (Klievink et al., 2018) and the capacity to deal with tangled societal challenges (George et al., 2016).

To overcome these limitations, a viable solution identified is the establishment of close data partnerships (Rasche et al., 2019), which are collaborative arrangements in which data is exchanged among a limited number of actors for pre-defined social purposes. These data-sharing arrangements provide a

Responsible Editor: Cinzia Cappiello

✉ Federico Bartolomucci
Federico.bartolomucci@polimi.it

Gianluca Bresolin
Gianluca.bresolin@mail.polimi.it

[1] Department of Management Engineering, Politecnico Di Milano, Via Lambruschini, 4/B, Milan, Italy

Published online: 10 November 2025

concrete alternative that conceptually sits between the open data movement (Sieber & Johnson, 2015), which advocates the unrestricted accessibility of data for all purposes, and market-based forms of data sharing, such as data ecosystems or data spaces (Oliveira & Lóscio, 2018; Otto & Hompel, 2022), where data sharing is typically driven by economic transactions. The benefits of closed data partnerships, hereafter referred to as Data Collaboratives (DCs) (Verhulst & Sangokoya, 2015), have attracted increasing attention in recent years from academic literature and grey literature (Global Partnership for Sustainable Development Data, 2023; Hoffman et al., 2019; The New Hanse Project, 2023), as well as national and international agencies and institutions (Berreteaga Barbero et al., 2020). A final surge of interest has been sparked by the emergence of the data altruism concept, codified in the European Union's Data Governance Act.

Since its definition in 2015 by Verhulst and Sangokoya as 'new organisational forms in which government agencies, non-profit organisations and private firms share specific datasets, including private datasets, with the purpose of addressing an important societal problem and thereby creating public value', the concept of DCs has struggled to establish itself as a standalone field of research and practice. At the research level, similar concepts like data ecosystems (Oliveira et al., 2019) or data spaces (Otto & Hompel, 2022) have emerged. At the practitioner level, a few domain-specific clusters of activity (e.g. migration, cooperation, education) have been created (see Global Partnership for Sustainable Development Data, 2023; Rango & Vespe, 2017) without a shared knowledge platform among them.

The lack of clarity on the scope of the phenomenon has so far hindered the development of new research on DCs. The lack of a comprehensive and empirically grounded classification offering researchers a clear understanding of its nuances has made it difficult for them to navigate the multiple forms that DCs may take. To stimulate the development of the field ten years after the concept's definition, this research seeks to fill this research gap by addressing the following questions: RQ1: How can data collaboratives be differentiated from other types of cross-sectoral data partnerships? RQ2: In which domains do data collaboratives represent an effective solution? RQ3: What impact categories do data collaboratives encompass? To this end, we first enhance the conceptualisation DCs by introducing a distinction around the economic nature of data, translating these reflections into six operational characteristics that clearly distinguish DCs from other forms of data-sharing initiatives. Secondly, building on previous classifications (Susha et al., 2017; Verhulst & Sangokoya, 2015), the enriched conceptualisation proposed and the analysis of a dataset comprising 171 different data collaboratives through a clustering methodology, we contribute to theory development by identifying five clusters of projects. These share organisational, technological and purpose-related characteristics, illustrating the domains in which data collaboratives have been most effective and the impact categories generated by these DCs. These findings provide policymakers, researchers and practitioners with a more detailed framework to understand the diverse applications and potential impacts of DCs compared to other types of data-sharing initiatives. Additionally, through cross-cluster analysis of the variables, we identify cluster-specific research challenges that could guide more targeted and effective future research within the field.

The paper is structured as follows: Sect. 2 reviews the literature on the DC phenomenon and offers a comparative analysis of DCs and other forms of cross-sectoral data partnerships, concluding with a list of six operational characteristics that distinguish DCs from other types of projects. Section 2 also defines the research gaps, while Sect. 3 details the methodology used for the analysis and describes the dataset. Section 4 presents the results, outlining the five clusters identified. Section 5 discusses the findings, presents the impact categories and proposes future research questions for the development of the field. Sections 6 and 7 address the limitations of the research and present the conclusions, respectively.

Related work

Defining DCs

The definition of DCs provided by Verhulst and Sangokoya (2015) cited in the introduction clearly states three identifying characteristics of this type of partnership: (a) their cross-sectoral nature, involving actors from different sectors; (ii) data centrality, meaning that data are the pivotal asset around which the collaboration is built and through which value is generated; (iii) the presence of a clear social purpose. In 2019, Susha and colleagues conducted a literature review to identify common elements that combine DCs with other similar forms of data-for-purpose partnerships, such as data partnerships, data donations and data philanthropy). In their study, they adopt the term Data Driven Social Partnership, defined as 'a collaboration between actors in one or more sectors to leverage data from different parties, at any stage of its lifecycle, for public benefit in policy or science.' Compared to the previous definition, this one delineates even broader boundaries for the concept by adhering to the second and third characteristics mentioned earlier while not considering the cross-sectoral nature of the partnership as a pre-requisite.

Both definitions are quite broad and allow for considerable variability among the initiatives falling under the concept of DCs. Verhulst and Sangokoya (2015), further

Table 1 Dimensions determining different configurations of data collaboratives

Dimension	Sub-dimension	Source
Actors	Actors involved, rules of engagement, incentives	Susha et al., (2019a, 2019b); van den Broek and van Veenstra (2018)
Data	Type of data, data provider, purpose of use	Verhulst and Sangokoya, (2015), van Loenen (2006) Susha et al., (2019a, 2019b)
Nature	Flexibility of the collaboration, domain, time horizon	Verhulst & Sangokoya, 2015; van den Broek & van Veenstra, 2018
Scope	Outcome of the collaboration, intended impact	Susha et al. (2017)
Business model	Profitability model	Susha et al., 2020, GSMA (2018), Alaimo et al., 2020
Facilitation	Intermediate or distributed	Perkmann and Schildt (2015); Stalla-Bourdillon et al. (2019); Verhulst and Sangokoya (2015)
Context	Socio-economic context	Castelnovo et al., (2016); Liva et al. (2023)

expanded by Verhulst, Young and colleagues (Verhulst et al., 2019a), instead distinguish DC models based on the degree of involvement between parties and the level of data accessibility. Zygmuntowski et al. (2021), differentiate them from other forms of data sharing by considering stakeholder control (i.e. institutional/trust-based) and value allocation (i.e. private-driven). Susha et al., (2019a) proposed a first taxonomy, identifying fourteen dimensions related to data supply and demand aspects, which can be used to determine different DC models.

DCs can be framed as highly heterogeneous and dynamic meta-organisations (Guggenberger et al., 2025), with organisational, technological and data-related differences among them (Alaimo & Kallinikos, 2024, p.151). DCs may involve actors from the private, public or social sectors who may play different roles in the collaboration (Jussen et al., 2024). Fo example, the problem holder owns or is close to the social issues (Niño et al., 2017); the data owner possesses the data; the skills holder has technological, data or social skills that compensate for the others (Oxford Internet Institute, 2014); the resource owner supports the project and manages the relationships that enable collaboration between such diverse players(Varshney and Mojsilovi´c 2019). DCs may also use different types of data, which could be classified based on multiple perspectives, including their nature, and the method by which they were collected (Susha et al., 2017; van Loenen, 2006). The nature of data can be: personal or non-personal, with personal data referring to specific users and non-personal data relating to things like traffic conditions or rainfall; disclosed or observed, with disclosed referring to data that has been intentionally made available by the data owner and observed relating to data that has been collected through observation methods such as web scraping or crawling. Furthermore, data can be classified based on their purpose: primary (i.e. data used for the purpose for which it was collected); secondary (i.e. data used for a similar purpose to the one for which it was collected); tertiary (i.e. data used for a purpose other than the one for which it was collected); or end-use (i.e. processed data whose results are used by end users)[1]. The nature of DCs can also vary depending on the degree of partnership flexibility, the time horizon and the domain of activity. The degree of flexibility is mainly determined by the ability to include actors or change scope over time; the DCs' scope might be narrowly defined or in some cases not precisely specified. In terms of time horizon, some DCs are created on a temporary basis, while others are aimed at long-term stability. The funding models supporting DC activities may also vary, with each model offering a unique value proposition and varying in terms of data value extraction (Alaimo et al., 2020), data ownership and control, bargaining power in decision-making, profitability and thus reliance on other network participants (Susha et al., 2020). DCs may also be distinguished by facilitation modalities. This role may be performed by ad-hoc organisations (Digital Civil Society LAB, 2017; Perkmann, 2016; Stalla-Bourdillon et al., 2021) acting as data stewards (The Gov Lab, 2020; Verhulst, 2021), or it may be distributed among the partners. Finally, DCs are also dependent on the contextual setting in which the partnership takes place. Indeed, socio-economic factors, such as the level of economic development, may influence other dimensions and determine the scope of action available to the collaborative (Castelnovo et al., 2016; Liva et al., 2023). Table 1 below summarises all the elements that may be used to determine different DC configurations.

Distinguishing DCs from other partnerships

As discussed earlier, the concept of data collaboratives (DCs) has remained relatively broad, encompassing a wide range of initiatives. However, the recent surge in data-sharing initiatives (Susha et al., 2022) has expanded the diversity and scope of initiatives falling under this notion, leading to increasing overlaps with other related concepts.

[1] For a more detailed classification of data, see Susha et al. (2017) and van Loenen (2006).

In particular, the lack of a clear distinction between data collaboratives, open data initiatives and data ecosystems may have hindered researchers' ability to understand the uniqueness of DCs (Rasche et al., 2019).

Open data initiatives are based on the conceptual premise that data is a *public good*, implying that data should be universally accessible to everyone and reusable for any purpose (Corrales-Garay et al., 2019; Janssen et al., 2012; Whitelaw et al., 2020). From this perspective, data are seen as a non-rivalrous and non-excludable asset, which means that many parties can use and reuse the same data a limitless number of times without diminishing its quality, quantity or value (Charles & Tonetti, 2020; Nikander et al., 2020). DCs, on the other hand, may fall within the scope of closed data partnerships, which refer to initiatives where data is shared among a well-defined set of participants, with their usage directed towards a predetermined goal (Rasche et al., 2019). This approach shifts from a view of data as a public good to one of data as a *club good* (Global Partnership for Sustainable Development Data, 2023; Savona, 2020). While acknowledging the non-rivalrous character of data, this alternative conceptualisation proposes that data can be excludable, with access restricted to approved members. Because of this distinction, open data-sharing initiatives and closed data partnerships activate different collaboration dynamics on various collaborative dimensions, such as incentives (Jussen et al., 2023), funding, trust and legitimacy generation dynamics (Rasche et al., 2019).

With respect to data ecosystems, although their definition is still under discussion (Geisler et al., 2022; Liva et al., 2023; Oliveira & Lóscio, 2018; Oliveira et al., 2019), distinctions can be made based on the autonomy and competitiveness of the ecosystem participants and the presence of a shared goal among the actors. In data ecosystems, data are perceived as a market good (Spiekermann, 2019), whose exchange should be regulated by the standard forces of supply and demand. Undoubtedly, both data ecosystems and DCs are complex socio-technical networks that facilitate collaboration for the purpose of data exploitation (Jussen et al., 2024; Oliveira et al., 2019). However, the scholarly discourse surrounding data ecosystems has repeatedly acknowledged the dual nature of competition and cooperation among ecosystem participants (Liva et al., 2023; Oliveira et al., 2019). This does not apply to DCs, where the non-competitiveness of actors is a prerequisite for the analysis developed using the data shared within the collaborative. Additionally, while Oliveira et al. (2019) highlight the independence of data ecosystem actors, and van Donge et al. (2022) note the potential lack of a common goal, these characteristics do not pertain to DCs. In DCs, alignment of the value proposition is essential for the success of the partnership (Ruijer, 2021; Susha, 2020), and the achievement of the social mission is only possible through the collaboration of interdependent actors. Acknowledging these distinctions, and in order to define a clear empirical setting for this study, we will therefore consider only those initiatives showing the characteristics described in Table 2. These can be considered identifying features of DCs and will enable us to build an empirical setting composed of initiatives adhering to this enhanced conceptualisation of DCs, thus excluding projects whose collaboration dynamics differ significantly (e.g. open data initiatives).

Existing categorisations and research gaps

The diversity of forms that DCs can take, combined with the novelty of the topic and the overlap with other similar data-sharing initiatives, has resulted in a research field where many questions remain unresolved. In this regard, Susha et al. (2018) identified nine critical research areas requiring attention to advance the development of DCs. These include evaluating data value, aligning data with relevant problems and partners, developing impact analysis methodologies, exploring incentive models, designing governance frameworks and improving data management and interoperability (Susha et al., 2018). However, addressing these challenges is difficult without a clear definition of the boundaries of DCs. The lack of clarity around its definition often leads to varying interpretations, making the phenomenon somewhat ambiguous and causing researchers to arrive at contradictory assumptions and conclusions.

To address this gap, Verhulst and Sangokoya (2015), expanded by Verhulst et al., (2019a, 2019b), proposed a high-level qualitative categorisation. Their classification identifies six types of DCs, including Public Interfaces, Trusted Intermediary, Data Pooling, Research and Analysis Partnerships, Prizes and Challenges, and Intelligence Generation. While this classification has its merits and has contributed significantly to the field's development, it also has several limitations. Indeed, being based on only two variables –engagement type and data accessibility – the categories identified are not mutually exclusive and do not exhaustively describe the various variables involved. For example, some categories focus on intermediation modalities, while others address data collection methods. As the only available classification, this has led to considerable ambiguity within the field. Moreover, the classification developed in the initial stages of the DC movement includes both open data projects and initiatives led by single organisations, which have since evolved into separate research streams.

Given these limitations, there is a clear need for a comprehensive classification system that provides a more holistic view of the organisational, technological, data- and purpose-related factors shaping DCs. Such a classification, grounded in extensive empirical evidence, would address the current gap in the literature and allow researchers to engage in more

Table 2 Operational characteristics of data collaboratives

Characteristic	Description
Cross-sectoral	Including actors from at least two different sectors among the private, public, not-for-profit and civil society
Purpose-driven	Aiming to pursue general interest purposes and generate public value
Data-centric	Leveraging data as a main value creation asset
Closed data partnership	Promoting the sharing of data among a limited number of approved partners
Non-competitiveness	Data is shared under a non-competitive regime
Interdependence	Actors are interdependent in reaching a shared value proposition

detailed analyses of the various DC models, thereby addressing unresolved questions in the field (Susha et al., 2018). Furthermore, as the field has evolved over the past decade, a new classification would better capture the nuances of this phenomenon, reflecting its maturation since the original framework proposed by Verhulst and Sangokoya (2015).

This research aims to bridge the gap between the initial exploratory studies (Susha et al., 2018; Verhulst & Sangokoya, 2015) and the growing need for clarity on the nature of DCs (i.e. the collaborative structures they adopt), their scope (i.e. the domain they operate in and the impacts they can achieve) and the unique development challenges faced by different DC models. By applying quantitative clustering methodology (see Sect. 3.3) and creating a new, expanded dataset that incorporates multiple additional variables, this study aims to offer a fresh perspective on DCs and open new avenues for research by clarifying the ambiguity that has surrounded this phenomenon to date.

Methodology

Building the dataset

The research started with the dataset available on datacollaboratives.org, which is the most extensive and up-to-date collection of DCs worldwide. The 249 collaboratives listed (as of November 2024) were sorted according to the six characteristics of DCs outlined earlier (see Table 2). Open data projects, single-player initiatives, competitive data-sharing ecosystems and projects without a clearly declared social purpose were excluded. Additionally, projects for which it was not possible to gather data online (e.g. no website or documents available) were excluded. Ultimately, after two rounds of independent reviews performed by the authors, 77 projects were excluded, resulting in a dataset of 171 DCs.[2]

As a second step, we proceeded with the selection of variables. These were selected based on insights gained from the explorative literature review presented in Sect. 2, with the objective of adopting a comprehensive view on the phenomenon. Therefore, the variables selected cover both organisational, technological, data- and purpose-related variables. To select the best mix of variables, we complemented those emerging from previous classifications and taxonomies (Susha et al., 2018; Verhulst et al., 2019a, 2019b) (see Table 1) with variables from the datacollaboratives.org repository, including those related to the impact dimensions of the collaboratives and those aiming to ensure that the collaboratives considered are still active. Multiple rounds of discussions among the authors and feedback received by experts in the field led to the selection of 17 variables, which were populated using secondary data sources. It is worth noting that to ascertain the possible values for the *impact dimension* variable, an inductive content analysis was conducted (Elo & Kyngäs, 2008).While the literature recognises the relevance of this variable (Azzone, 2018; Chui et al., 2018; Verhulst & Young, 2011), efforts to demonstrate the impact categories created by DCs have been limited (Susha & Gil-Garcia, 2019; Susha et al., 2019a, 2019b), and there is a lack of defined impact categories specific to DCs in the dedicated literature.

All identified variables, the source from which they were integrated, their potential values and the rationale for excluding them from the research's analytical phase (further elaborated in Sect. 4.3) are presented in Table 3. For calculation purposes, some have been converted into Boolean variables.

Aggregated dataset description

The resulting dataset is composed of 171 DCs. Among the various types, research projects account for one-third (58), while prizes and challenges account for only 10. The remaining DCs across the other various types of collaboratives is evenly distributed, with around 20 projects each. In terms of domains covered, the Health sector is the most represented with 72 DCs, while all other domains appear in fewer than 20 cases. Regarding the geographical scope, most DCs operate either globally (59) or at the country level (52), with fewer occurrences at continental (33), city (16) or area (11) levels. With respect to the number of partners

[2] The full dataset is available at the following link: https://zenodo.org/records/15092656.

Table 3 Variables in the dataset and details on those used for the analysis

Macro dimension	Variable name	Type	Description	Possible values	Exclusion criteria	Literature
Actors	(Actors involved) Private	Boolean	Whether the private sector participates in the data collaborative	Yes, No	Excluded from the data analysis phase because of a low correlation between the clustering labels (Cramer's V test < 0,2)	Adapted from Susha et al. (2019a)
	(Actors involved) Public	Boolean	Whether the public sector participates in the data collaborative	Yes, No	Included	Adapted from Susha et al. (2019a)
	(Actors involved) Not-for-profit	Boolean	Whether the not-for-profit organisations participate in the data collaborative	Yes, No	Excluded from the data analysis phase because of a low correlation between the clustering labels (Cramer's V test < 0,2)	Adapted from Susha et al. (2019a)
	(Actors involved) Civil society organisations	Boolean	Whether civil society organisations (international organisations, researchers, universities, citizens and, their representative) participate in the data collaborative	Yes, No	Included	Rasche et al. (2019)
Data	Data purpose	Categorical	The difference between the purpose for which data was initially collected and the purpose for which it is used in the collaborative	Primary, secondary, tertiary, multiple data sources	Included	Susha et al (2017)
	Type of data used	Categorical	The type of data used by the collaborative, based on four non-exclusive categories of data	Disclosed personal data, observed personal data, disclosed non-personal data, observed non-personal data	Excluded because of missing data in the dataset	datacollaborative.org
	Continuity	Categorical	How data are released and accessed by partners over time	Continuous, event-based, on demand	Included	Susha et al., (2017, taxonomy) Van de Broek & Van Veenstra (2018)
Nature	Domain	Categorical	The main domain to which the project belongs	Agriculture, criminal justice, crisis response, digital society, economic development, education, environment, health, infrastructure, transportation, telecommunications, humanitarian	Redundant with the 'Data collaborative type' and'Impact dimension' variables; less structured and informative	datacollaborative.org
	Multiple domain	Numerical	If the project spans multiple domains, this attribute accounts for their number	Numbers from 2 to 11 (the total number of possible domains)	Considered to have a low value for clustering analysis, due to both the information it provides and the lack of a properly formalised method used to analyse the domain of a project in the repository	datacollaborative.org
	Status	Categorical	The present status of the project	Ongoing, concluded, failed	Not adding valuable information to the analysis	(Own elaboration for research purpose)

Table 3 (continued)

Macro dimension	Variable name	Type	Description	Possible values	Exclusion criteria	Literature
Scope	Clear objective	Boolean	Whether the project has a clear and specific purpose, or whether it has a more flexible (or unspecified) objective	Yes, No	Included	Ramon Gil-Garcia et al. (2007); Susha et al., (2017, taxonomy)
	Clearly social	Boolean	Whether the generation of social impact is the primary goal of the project, or whether it coexists with other actors' personal interests	Yes, No	Included	Adaptation of Susha, Rukanova et al. (2019)
	Impact dimension	Categorical	The social impact domain of the project	Improve health, tackle the COVID-19 pandemic, environmental impact, tackle poverty and support underserved economies, provide humanitarian support, promote gender equality, implement city smart management, advance social studies, improve education, multiple social purpose	Included	Inductive content analysis Verhulst, Sangokoya (2015)
Business model	Clear link with profitability	Boolean	Whether the project has the potential to be profitable for the data-sharing organisation outside of image effects	Yes, No	Excluded in the data analysis phase because of a low correlation with the clustering labels (Cramer's V test < 0,2)	Klein & Verhulst (2017)
Facilitation modality	Data collaborative type	Categorical	The type of governance used by the data collaborative, defined by engagement and accessibility	Public interfaces, trusted intermediary, data pooling, research and analysis partnership, prizes and challenges, intelligence generation	Included	Verhulst et al. (2019)
Contextual	Region	Categorical	Geographical location of the project	Either the country in which the project has effect, its continent or worldwide	Excluded because of differing data granularity and ambiguity regarding the subject to which it is refers (promoters, beneficiaries, data owners)	datacollaborative.org
	Extension	Categorical	The geographical extension of the project	City, area, country, continent, worldwide	Included	Adaptation of (Castelnovo et al., 2016; Ooms et al. 2020; Viale Pereira et al. 2017

involved, 106 DCs involve actors from two sectors, with 50 including actors from three sectors) and15 from four sectors. A clear objective – defined asa specific goal to address an identified challenge – is evident in 85 projects.The majority of initiatives (144) declare a clear social purpose. In terms of impact, almost one-third of the projects (54) operate in the Health sector, with an additional 20 focusing on COVID-realted issues. Additional significant groups focus on smart city management (21) and multiple social purposes (24).

Data analysis methodology

Although clustering techniques are more developed for numerical variables, there are many fields in which categorical data are prevalent (Šulc et al., 2018). Asa result, several techniques have been developed to address the unique characteristics of these contexts (Alves et al., 2019; He et al., 2002; Řezanková, 2016).

The first step in conducting an analysis is selecting which variables to include. Nine out of seventeen variables were excluded from the analytical phase for a variety of reasons (described in Table 2), but their distribution across the clusters was later assessed to better interpret the results. The remaining 9 variables are: Actors Involved, which tracks the type of actors engaged; Data Purpose, introduced by Susha and colleagues (2017), which quantifies the degree of divergence between the original purpose for which data was gathered and its purpose in the project; Continuity of Collaboration, a variable developed by Susha et al. (2017), which characterises how data are exchanged over time; Clear Objective, which indicates whether the project is driven by a clearly stated purpose; Clearly Social, a variable designed to simplify the message conveyed by Susha, Rukanova and colleagues (2019).This boolean variable identifies projects where private actors express an interest in participating in the partnership for reasons other than a social one (i.e. economical or reputational); Clear Profitability Model, which indicates whether the DC is designed to sustain its activities via a clear profitability model (Susha et al., 2020); Extension, which indicates whether a project was established to have an influence on a city, an area, a country, a continent or on a global scale;, Impact Dimension, indicating the primary area of social impact in which the collaborative operates; Data Collaborative Type, introduced by Verhulst and colleagues (2019), which primarily refers to the facilitation modality while also incorporating the broader classification proposed by the authors. For calculation purposes, the Actors Involved variable was transformed into four Boolean variables, while other variables (i.e. Clear Objective, Clearly Social and Clear Profitability Model) were converted into Bolean variables with values 0–1. However, to assess the value of each variable, a detailed analysis of their models was conducted using secondary data.

Once the variables to be included in the analysis had been determined, the clustering algorithm followed. Initially, distance measures based on information theory were selected, which relies on the assumption that rarely observed values in a dataset could convey a greater amount of information (Burnaby, 1970). This assumption is especially relevant in rapidly developing fields, where rare values may reveal emerging trends. Five similarity measures proposed by Boriah et al. (2008) were tested: Smirnov, Gower, Goodall1, Goodall3 and Occurrence Frequency. Hierarchical clustering methods, which do not require any prior indication of the number of clusters, were chosen as the preferred option (Řezanková, 2016; Vercellis, 2009). The decision between the four most used linkage methods – linkage, single linkage, average linkage, and Ward's method – (Vercellis, 2009) and whether to perform the clustering using an agglomerative or divisive logic was deferred to the analytical phase, whereeach method was selected to find the best solution.

The Bayesian Information Criterion (BIC) and, more precisely, its first revision (BIC1) suggested by Šulc et al. (2018), was used to evaluate the quality of the different clustering models.

To increase the validity of the analysis, we determined the correlation between variables and the discovered solutions using Cramer's V test (Fraiman et al., 2008). Numerous iterations of hierarchical clustering were performed, using various similarity and distance metrics. Following a qualitative appraisal of the solutions with one additionaland one fewer cluster comparedto the best-performing solution (Šulc et al., 2018), the five-cluster solution was chosen. Table 4 reports the parameters used to generate this output. Annex 1 provides a detailed description of how each methodological choice was taken, including how the parameters reported in Table 4 were selected, the clustering dendrogram, which gives a graphical suggestion on where to operate the optimal tree cut, and the rationale behind selecting the 5-cluster solution.

Table 4 Cluster analysis parameters used for the solution adopted

Type of variable	Distance measure	Clustering methodology	Linkage distance	Goodness of clustering	Variables' relation to clusters	Excluded variables (Cramer's $V < 0.2$)
Categorical or Boolean	Goodall3	Hierarchical	Ward	BIC1	Cramer's V test	CPM, Not-for-profit, Private

Results

The cluster analysis identified five mutually exclusive clusters, meaning that each data collaborative can be assigned to only one group. The interpretation of each cluster was derived from observing the relative distributions of the variables used in the analysis, compared to their average distribution across the entire sample. This quantitative analysis was then complemented by qualitative observations of each project within the clusters, based on the analysis of secondary data sources.

Each author then presented their interpretation of each cluster and engaged in discussions with the others until a consensus was reached. In the next section, we provide a brief interpretation of each cluster following this structure: first, we present our qualitative interpretation, supported by data. We then focus on the strengths of each cluster as well as the inconsistencies revealed by the data that warrant further analysis. To illustrate each cluster, we provide an example of a real DC included in the respective cluster. These observations form the basis for the research questions presented in the discussion (see Table 5). A table comparing all the clusters and the distribution of the analysed variables across them is provided in Annex 2.

Cluster 1: Data-driven initiatives to support innovation

Cluster 1 (see Fig. 1) groups data-driven projects that are initiated by the availability of data rather than a clear social need. These collaborations excel at accelerating data-driven innovation by identifying unexpected social value in data and developing novel, previously unexplored pathways for data usage. Collaboratives in this cluster are not constrained by a specific objective, with only 3% of the projects having a clearly defined goal. There is a strong incentive for individuals with data analysis capabilities to engage in these projects, which is reflected in the highest presence of civil society actors among the clusters (87.1%, compared to the average of 70.7%). Data analysis challenges are often assigned to individual data scientists or informal groups of researchers, resulting in temporary involvement rather than long-term collaborations.

Tertiary data, which is often unrelated to the problems it is addressing, is prevalent in this cluster. Collaboratives here often take the form of single events or temporary calls for solutions and are not confined to a specific domain of application. The majority of these projects present 'multiple social purposes' (64.5%), indicating not only a lack a

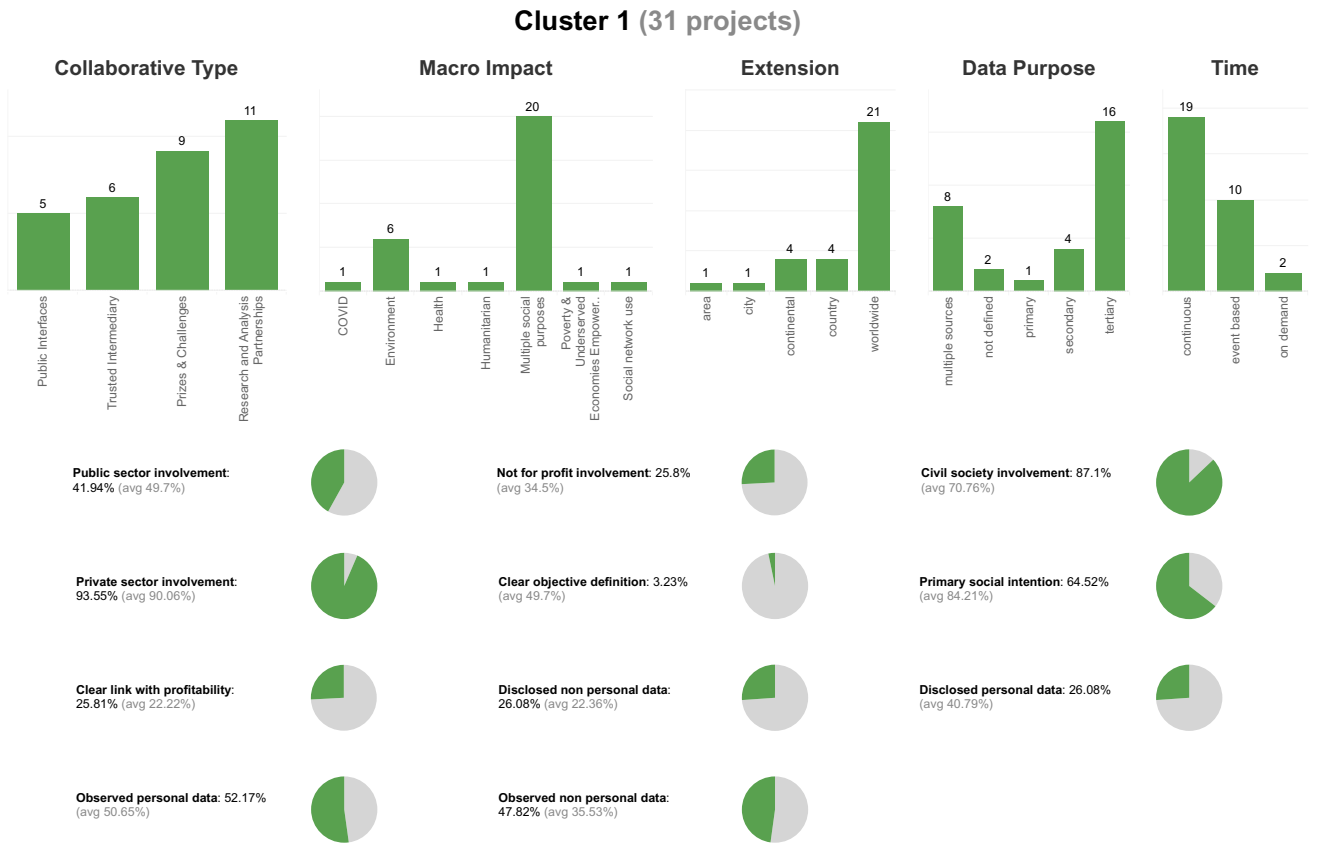

Fig. 1 Variables' distribution in the cluster 'Data-driven initiatives to support innovative studies'

clearly defined aim but also a broad and undefined impact area. Although projects within this cluster have a temporary scope, various types of intermediary and support organisations – acting as conveners (Susha et al., 2022) – have been trialled at the practitioner level. Private companies are often supported by external organisations specialised in promoting such initiatives.

This cluster raises two main research questions. The first concerns the business and operational models of these entities, whilethe second focuses on the real impacts generated by this type of initiative. Additionally, these initiatives shed light on how to attract talent and data capacity around social causes, akey limitation preventing social and public sectors from extensively leveraging data. These initiatives demonstrate the potential to overcome this.

An example of the opportunistic and open-ended nature of Cluster 1 data collaboratives is the initiative of Azavea, a B- corporation specialising in creating civic geospatial software, which collaborated with AWS to grant its fellows access to AWS's earth dataset – an extensive collection of satellite data and imagery. The project aimed to mentor and partner with skilled software developers to tackle a range of potential civic, social and environmental issues, from climate indicator analysis to urban planning. This highlights the cluster's emphasis on unlocking latent value in tertiary data, attracting independent data scientists and facilitating short-term, impact-agnostic innovation.

Cluster 2: Collaborative efforts for large-scale research

This cluster consists primarily of international DCs, with 87.5% of operating cross-nationally. Collaboratives in Cluster 2 (see Fig. 2) focus on developing innovative research projects, predominantly in the Health domain, through the systemic pooling and reuse of data collected from various sources. Only 15% of the projects have a clearly defined objective, suggesting that data collaboratives in this cluster are more inclined to develop broad, unconstrained pathways to generate social impact on a global scale through exploratory, high-level research initiatives.

Notably, 92.5% of the DCs in this cluster have a clear social purpose, which aligns with the prominent presence of civil society representatives, particularly universities and research institutions.

Surprisingly, this cluster exhibits relatively low involvement from public bodies at both local and national levels. Another interesting feature is the widespread presence of

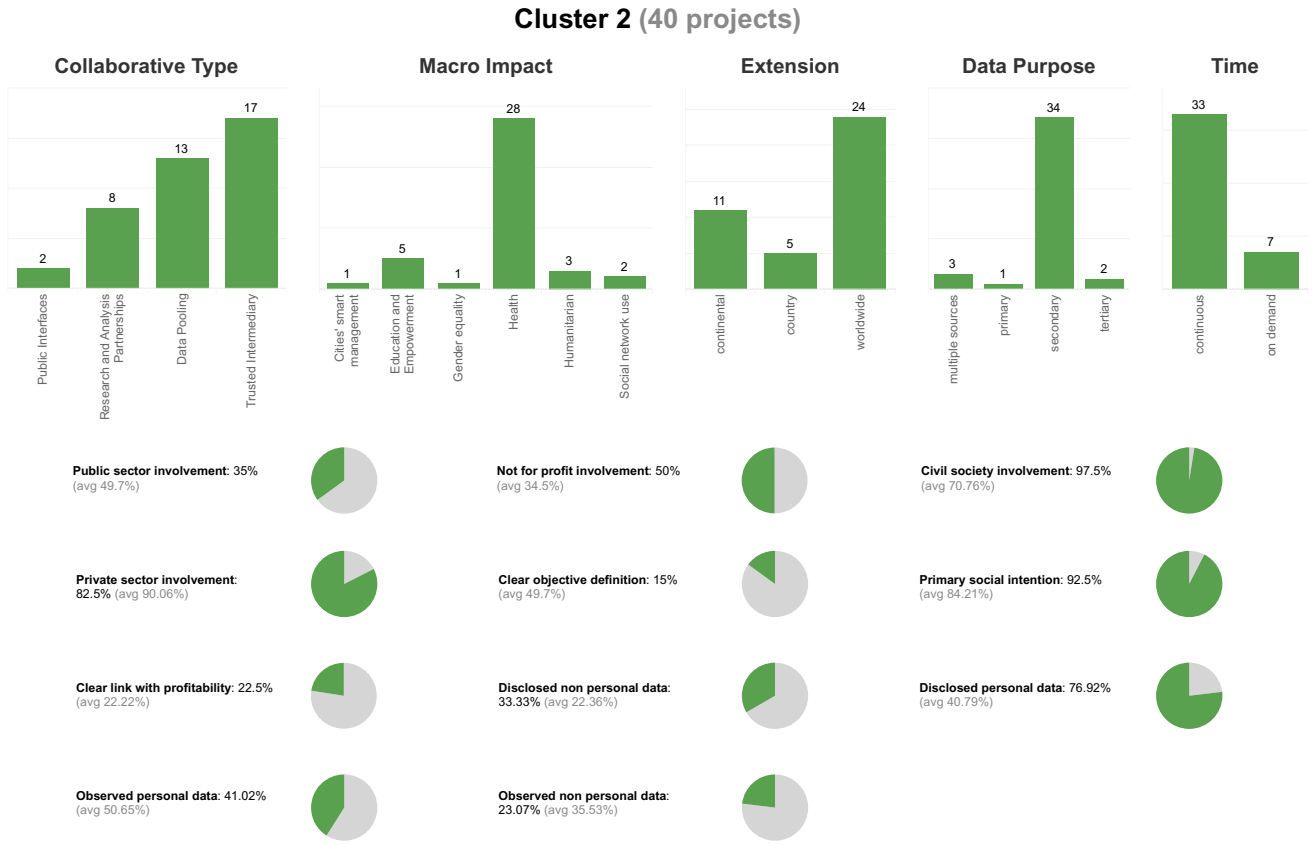

Fig. 2 Variables' distribution in the cluster 'Collaborative effort to support wide-scale research projects'

not-for-profit organisations, which, upon closer inspection, are predominantly research institutions active in the health sector. Additionally, projects within this cluster frequently rely on the on-demand disclosure of large secondary datasets (85%) provided by private companies for research purposes. This highlights the cluster's ability to engage the private sector in data-intensive research initiatives, a capability that surpasses that of other clusters.

The incentives driving private sector participation, along with the property rights, data-sharing conditions and infrastructure they use, warrant further investigation. A deeper examination of these factors could help identify best practices that might be replicated across other clusters. However, a surprising observation is the limited involvement of public sector actors in these projects. This calls for further research to understand the reasons behind this trend and to consider the potential benefits of more extensive public sector participation in these initiatives.

One Mind n.d.) is a leading example among the cross-national, data-intensive research programmes belonging to Cluster 2. By aggregating and anonymising data shared by patients with brain illnesses and injuries, this non-profit platform facilitates large-scale collaboration among researchers, clinicians and healthcare companies, with the aim of advancing mental health. This initiative underscores the cluster's emphasis on pooling secondary sourced data to drive impactful health research, leveraging cross-sector participation while fostering global scientific innovation.

Cluster 3: Continuous effort to improve systemic responses

Collaboratives in Cluster 3 (see Fig. 3) focus on generating systemic improvements to quality of life, promoting a more inclusive and sustainable development of society and urban settlements and enhancing emergency prevention. The heterogeneity of the data within this cluster presents challenges for interpretation. Specifically, the geographical areas covered by these projects vary in their geographic scope. However, upon closer examination, it becomes clear that the geographical and institutional boundaries of these interventions are always clearly defined, whether at the city, national or continental level.

Most collaboratives in this cluster are continuous in nature, with public sector involvement nearly twice that of the average. Additional insights emerge when examining the intended impacts of these projects. Notably, 90% of the projects in the entire dataset that aim to 'implement smart

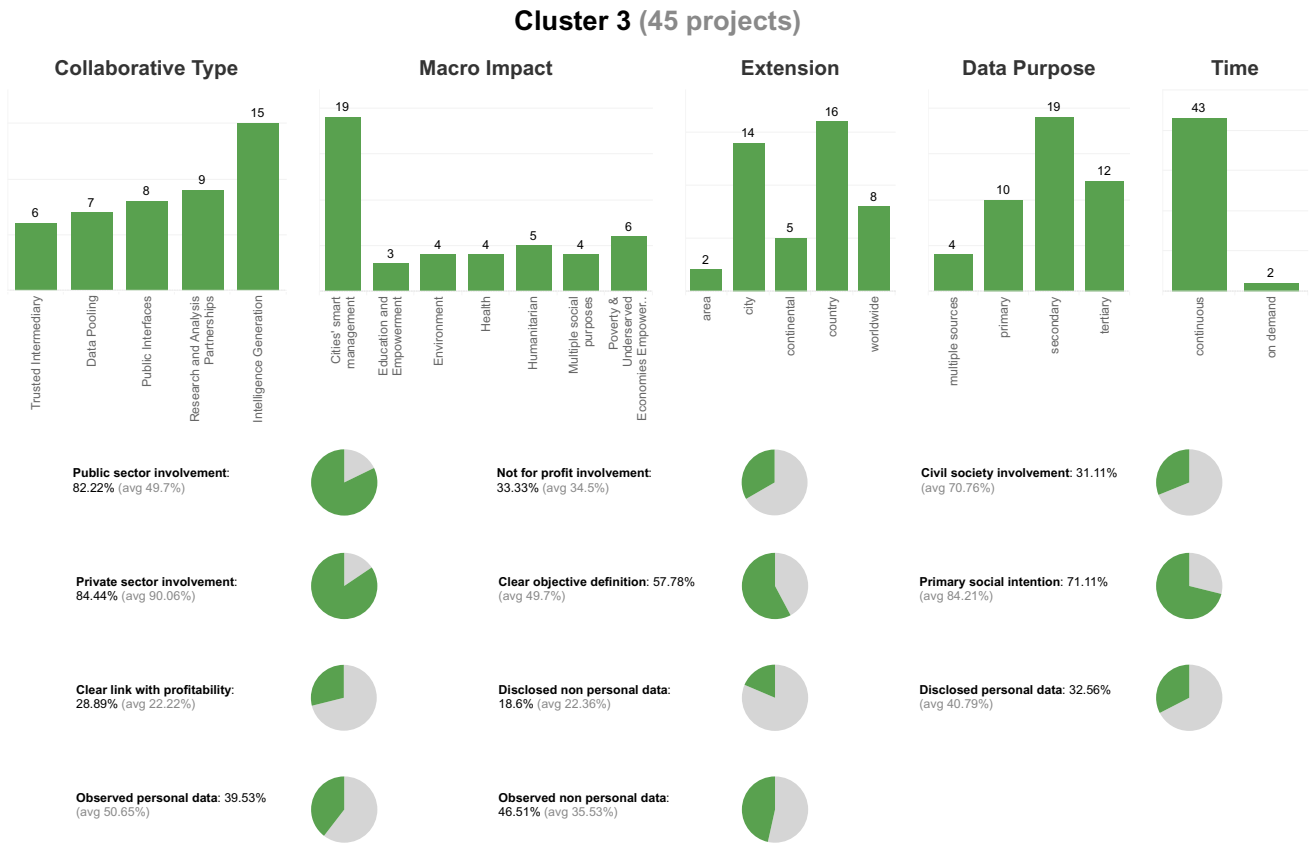

Fig. 3 Variables' distribution in the cluster 'Continuous effort to improve systemic responses'

city management' fall within this cluster, accounting for nearly half of its collaboratives. A cross-sectional analysis of these factors reveals the dual nature of the cluster: while many of these projects focus on implementing data-driven improvements to existing services and infrastructure, others aim to develop sophisticated systems that use data for effective emergency prevention – as opposed to recovery, which is the focus of Cluster 4.

These projects are not created in response to emergency conditions but rather as efforts to improve the systemic ability of cities, governments and continents to respond to them. Field experiments, such as urban transportation management and social service allocation, are frequently supported by public administrations at the local, state and continental levels. The presence of numerous 'intelligence generation' projects (33%) within this cluster supports this interpretation. As defined by Verhulst et al., (2019a, 2019b), these projects involve businesses and organisations analysing data to produce and share new knowledge that 'informs policy-making and service delivery'. Interestingly, the data reveals that fewer than one-third of the collaboratives in this cluster (29%) appear to have developed an economically sustainable model. Although this is above the average, it stands in contrast to the long-term aspirations of these projects. This incongruence highlights the field's immaturity and underscores the need for further research into potential business models that organisations promoting these projects could adopt (GSMA, 2018; Susha et al., 2022). Additionally, the data suggests that further investigation into the relationship between these collaboratives and their capacity to engage public bodies would be valuable. This evidence could contribute to the systematisation of other clusters, which often struggle to involve the public sector.

A fitting example for this cluster is the Metro project launched by the fitness app Strava n.d., which shares anonymised user-generated data – such as cycling and running activity records – with city governments and urban planners. The initiative aims to support evidence-based infrastructure improvements and urban development, such as optimising bike-sharing locations and increasing road safety. This reflects the cluster's emphasis on leveraging data for systemic urban development, smart city management and proactive mobility planning in collaboration with public authorities.

Cluster 4: Prompt response to emergencies

Initiatives in the fourth cluster (see Fig. 4) are designed to provide immediate responses to humanitarian crises triggered by

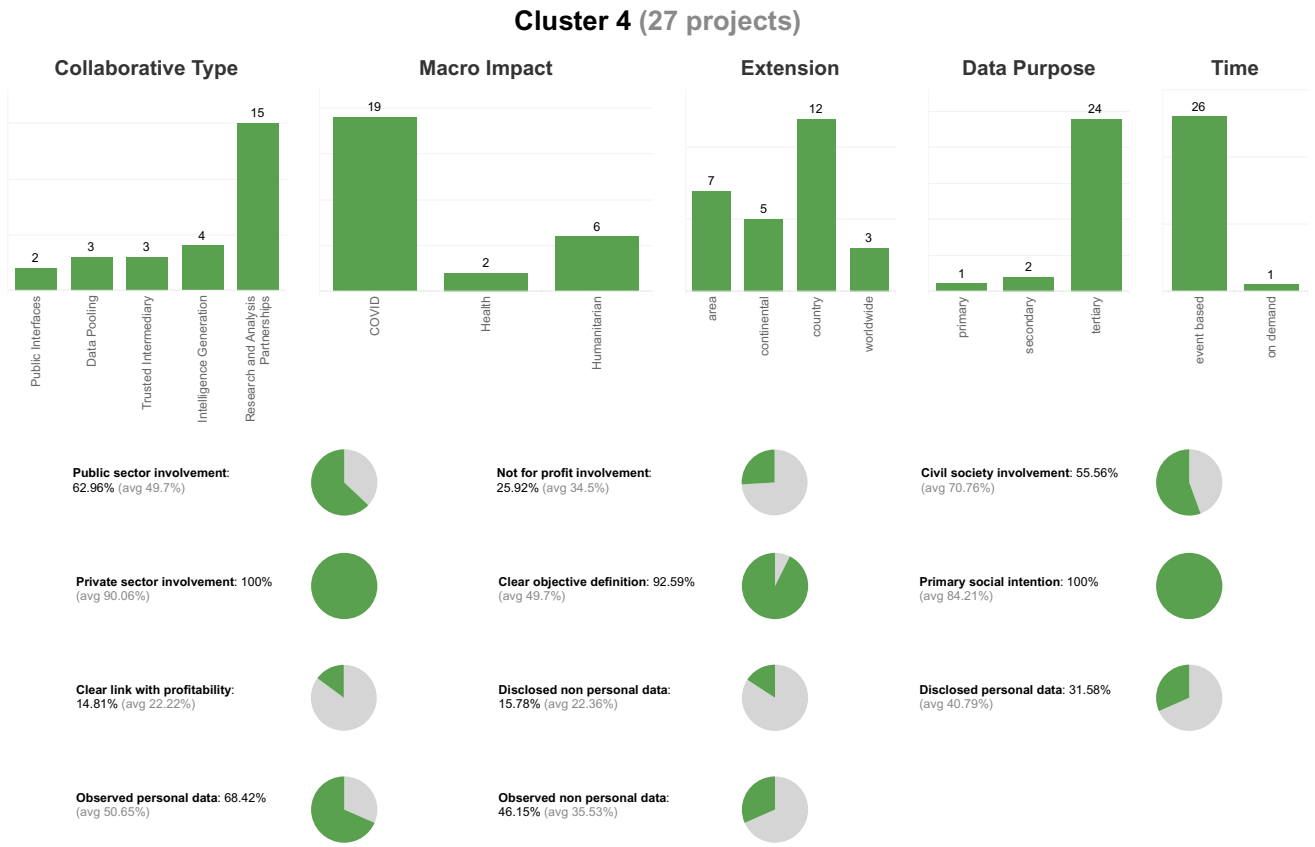

Fig. 4 Variables' distribution in the cluster 'Prompt response to emergencies'

natural disasters or, more recently, the COVID-19 pandemic. This context provides a clear interpretation of this group of collaboratives. Most of the variables analysed support this assessment, confirming that all projects have clearly defined social objectives and are predominantly temporary in nature. This suggests that the initiatives in this cluster are contingent on exceptional circumstances, with both their conception and execution dependent on these conditions.

Most projects (55%) adopt the collaborative framework 'Research and Analysis Partnerships', also known as 'Intelligence Generation', where one partner analyses data and provides the analysis to other partners to enhance their activities (e.g. aid distribution). Eighty-eight percent of the collaboratives in this cluster rely on tertiary data sources, which is consistent with the difficulty of collecting new data during the immediate contingencies of an emergency. Surprisingly, three initiatives in the cluster that had been classified as Trusted Intermediaries in previous studies were, upon closer inspection, revealed to be companies with data analysis functions. Regarding the actors involved, Cluster 4 is unique in that all the collaboratives include the business sector. This suggests that emergencies play a significant role in unlocking data and capacities across different sectors. Although much research has been conducted on the limited effectiveness of these solutions in mitigating pandemic contagion (Curioso & Carrasco-Escobar, 2020; Kretzschmar et al., 2020; Munzert et al., 2021; Whitelaw et al., 2020), insights into why they are able to unlock privately held data could be valuable. While private sector involvement is justified by the emergency conditions that prompt their mobilisation of resources, the minimal participation of the non-profit sector is surprising and warrants further investigation. A closer analysis of the data reveals that while programmes related to natural disasters rely heavily on international collaboration, those focused on pandemics are often established at the regional or country level. Finally, the temporary nature of these partnerships underscores the need to explore mechanisms to preserve the knowledge, infrastructure and relationships developed during emergency conditions beyond the crisis period. This aspect should be further explored to prevent the complete dissipation of the initial investments made to establish these initiatives.

The partnership between the telecom provider Digicel Haiti and researchers from Karolinska Institutet and Columbia University (Karolinska Institute, 2011) following the 2010 earthquake and ensuing cholera outbreak in Haiti exemplifies the core characteristics of Cluster 4, rooted in the nature of its response to natural disasters. By analysing anonymised mobile phone data from two million devices, the initiative provided real-time insights into population movements, helping to optimise aid distribution and prevent the risk of further outbreaks. This aligns with the cluster's emphasis on temporary, data-driven emergency responses that mobilise private-sector resources in crisis situations.

Cluster 5: International mobilization for development

The fifth cluster (see Fig. 5) comprises global initiatives aimed at solving structural problems, primarily in developing economies, through international cooperation schemes. Data show that more than half of these initiatives impact low- and middle-income countries; 71.4% involve international partners and all are designed to generate a predeclared positive social impact, with almost all having a single objective. More than half of the DCs in this cluster aim to create an impact at the country level. Surprisingly, this cluster shows low participation from public institutions, which is offset by the involvement of civil society institutions in 92.8%of projects. A closer look at the individual initiatives reveals a high concentration of international NGOs and local universities. Collaboratives in this cluster are often driven by foreign institutions or enterprises aiming to address systemic problems through data in countries lacking the structural capacity and knowledge to do so. This includes projects with the aim of improving people's living conditions by ensuring access to basic services such as water, food provision and adequate medical care. They also address systemic and long-term issues that affect liveability, such as infectious diseases and environmental migration. Interestingly, the majority of DCs in this cluster are structured as ongoing efforts, although only a minority of projects include a clear profitability model. An examination of all the initiatives reveals that many are sustained through external grant funding, ensuring their long-term development, with NGOs and universities representing the local support structure. The absence of local public authorities raises immediate questions about the capacity of these solutions to scale at the institutional and regulatory levels, highlighting the latent potential that could be unlocked through their involvement.

An example of the international engagements within this cluster is the MINE consortium, which includes, among others, Microsoft and the L V Prasad Eye Institute in Hyderabad, India (Microsoft, 2016). As part of this collaborative, Microsoft shares technical expertise and data with the goal of eliminating avoidable blindness in India and worldwide. Multiple international partners are involved in the initiative, employing advanced data-driven techniques to address structural healthcare challenges and aiming to provide sustainable social benefits, especially in regions with limited local capacity.

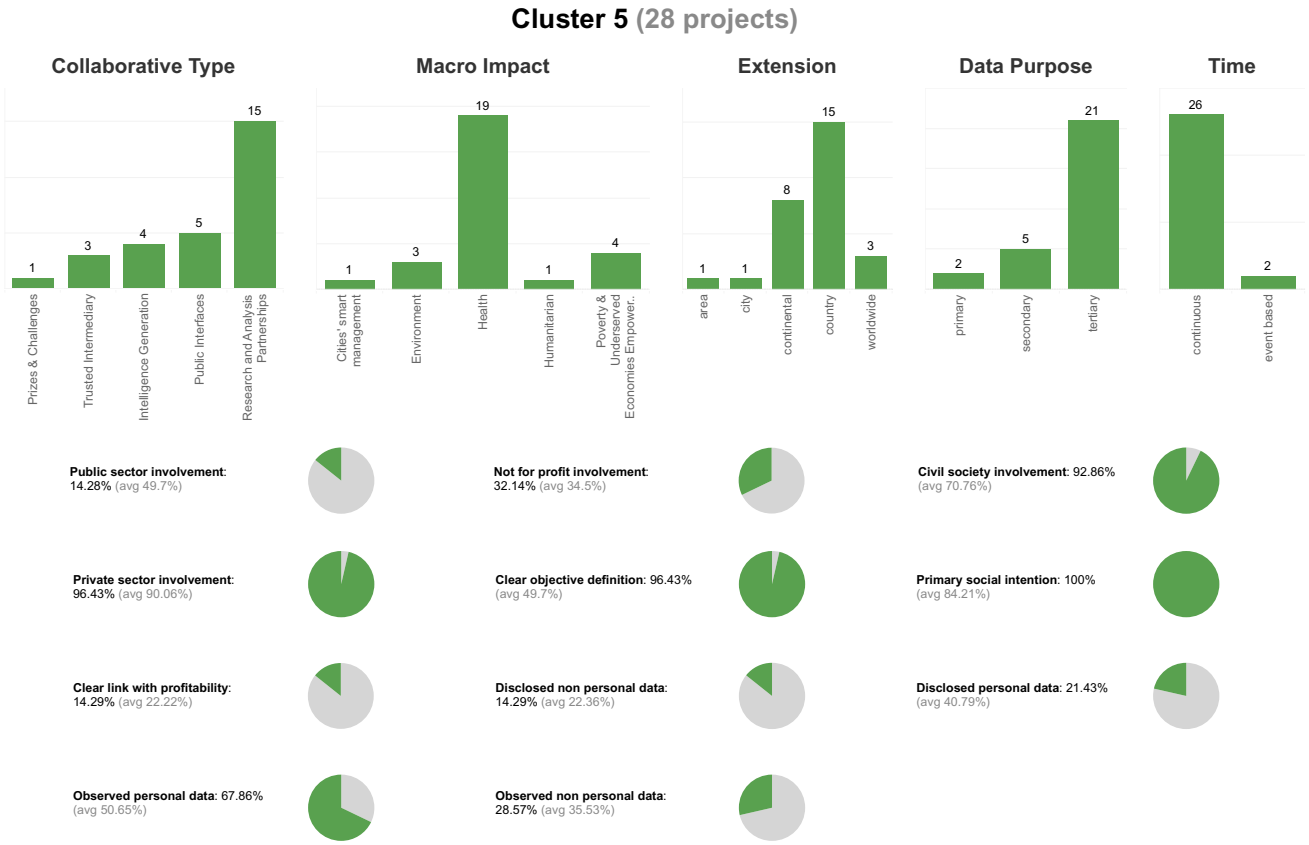

Fig. 5 Variables' distribution in the cluster 'International mobilisation for development'

Discussion

A decade after its introduction by Verhulst and Sangokoya (2015), the proliferation of initiatives at both the institutional and practitioner levels, as well as academic publications on similar concepts (Liva et al., 2023; Macdonald et al. 2023; Oliveira et al., 2019; Otto & Hompel, 2022; Sieber & Johnson, 2015), made it possible and necessary to establish clear conceptual and operational boundaries between DCs and other forms of data-sharing initiatives. We therefore started our research with a comparative analysis of the definitions of DCs provided by Verhulst and Sangokoya (2015) and Susha et al., (2019a, 2019b), as well as the characteristics of similar initiatives such as open data (Corrales-Garay et al., 2019; Zuiderwijk, 2017) and data ecosystems (Oliveira et al., 2019).

Based on this analysis and in response to RQ1, we offer a first theoretical contribution by establishing a clear conceptual distinction among these different phenomena. We adopt the theoretical approach of viewing data as a club good, which serves as a conceptual boundary distinguishing DCs from open data, while the non-competitiveness of actors and their interdependence to reach a clear value proposition differentiate DCs from data ecosystems. Empirically, these conceptual distinctions allowed us to identify six characteristics, defining for the first time a clear perimeter for the phenomenon. By linking the theoretical debate on the economic nature of data and the competitiveness of data-sharing environments to the operational characteristics of different data-sharing configurations, we can offer a new perspective that helps both researchers and practitioners to align the best data-sharing configurations with their value proposition and expectations.

In response to RQ2, the research identifies domains in which DCs have demonstrated effectiveness (e.g. research projects in the health sector and the design of solutions for urban infrastructural challenges). The results highlight five distinct DCs clusters, each distinguished by organisational, technological or purpose-related characteristics. Compared to prior classifications, our framework provides a more structured, enriched and comprehensive understanding of DCs. Previous research (Susha et al., 2020) has often addressed the phenomenon of DCs without acknowledging their differences with respect to other phenomena (e.g. data collaboratives versus open data), as well as their internal heterogeneity (e.g. temporary initiatives versus long-term efforts) and the implications these may have on research results. Different partnerships respond to different logics,

which makes them difficult to compare. Furthermore, due to confusion over terminology, other research that may have benefitted from engaging with DC literature overlooked it (e.g., Haak et al., 2018). Our findings may thus support further research, allowing better comparative studies aimed at understanding how different data-sharing configurations work, how different DC models function, or facilitating model-specific studies focused on a single cluster, thereby advancing knowledge on its dynamics. In this context, our findings constitute an analytical and descriptive theory contribution (Gregor, 2006). Such contributions are particularly valuable when addressing emerging phenomena that remain poorly understood (Fawcett & Downs, 1986). To achieve this, we have followed the key parameters for building analytical theories in information systems (Gregor, 2006). We first identified the scope of validity for our analytical theory (Table 2). Subsequently, we described the category labels used for classification (Table 3), detailed the methodology employed to develop the classification (Table 4) and outlined the defined characteristics of each category (Figs. 1, 2, 3, 4, and 5). In line with Foucault (1994), this led us to establish natural and exhaustive categories, which also serve as a revision and update of existing typologies (Susha et al., 2018; Verhulst et al., 2019a, 2019b).

With reference to the multitude of scopes that DCs may serve and addressing RQ3, the research identifies categories of impacts generated by DCs. The identified impact categories (Fig. 6) highlight the wide range of purposes that these types of collaborations fulfil, establishing them as a significant tool for achieving several objectives, including those in health, gender equality, urban management and education. The identification of impact categories, inductively derived from the content analysis of secondary sources (Elo & Kyngäs, 2008), further facilitates our understanding

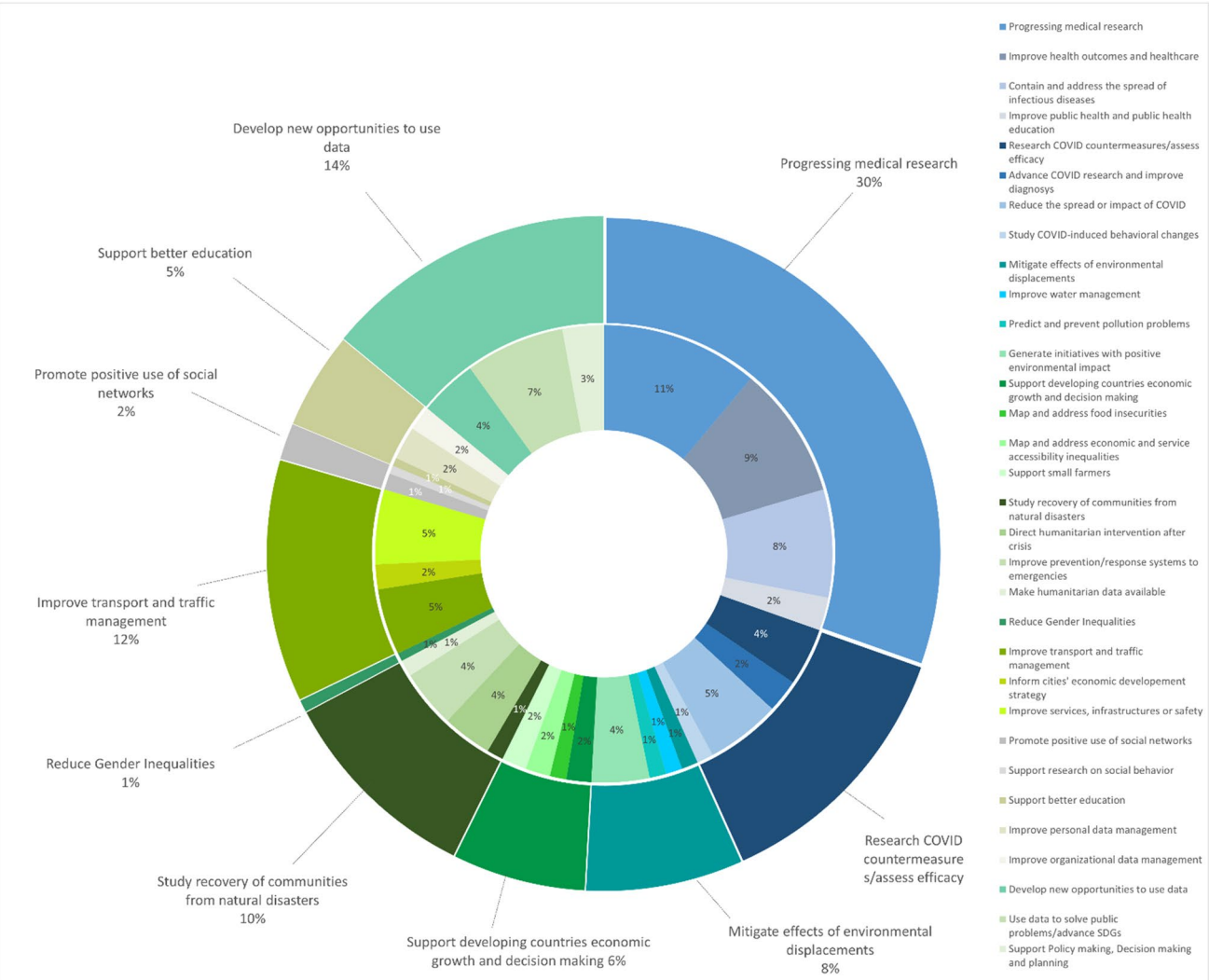

Fig. 6 Impact categories of data collaboratives

of the phenomenon and its evolution over the past decade. Additionally, it allows us to compare the impact dimensions addressed by DCs and those covered by other forms of partnerships, such as open data (Verhulst & Young, 2011).

The cross-cluster interpretation of the results, along with the impact analysis, highlights key differences among the various DC clusters, underscoring the need for dedicated research that addresses the unique characteristics of each cluster. This is particularly relevant across three dimensions: long-term versus short-term projects; projects that either succeed or fail in involving the public sector and civil society; clearly defined projects versus ambiguous objectives – all of which call for separate analysis. Keeping these differences in mind may help future research better address the multiple open questions (Susha et al., 2018) that continue to limit the field's development by better linking them to specific subsets of DCs facing these challenges. This is already the case, for instance, for business-to-government data sharing (Praditya et al., 2017; Signorelli et al., 2024) but remains less widespread for other forms of partnerships.

The cross-cluster interpretation of the results also helps to identify the specific strengths of each cluster, which can be further examined to understand their underlying logic and potentially replicated in other clusters. This is evident, for example, in clusters that demonstrate the ability to engage the public sector or attract a data-skilled workforce (e.g. Cluster 3), in contrast to those that do not. To support this process of mutual learning among clusters, Table 5 presents a set of cluster-specific research questions intended both to stimulate the development of individual clusters and to foster positive cross-cluster exchange. These questions are not supposed to replace more general inquiries, such as those proposed by Susha et al. (2018), but rather to complement them, enabling further research on the topic.

Finally, our findings, together with the detailed dataset provided by the research, may also assist practitioners in identifying benchmarks for their initiatives by selecting them based on their similarities across multiple variables. Compared to previous classifications, the wide range of variables used in this analysis enhances the ability to make comparisons and supports the use of multiple variables.

Our findings underscore the significant potential of DCs to enhance data availability for sustainable development. As Schoormann (2023) points out, many existing AI applications in this domain rely on a combination of open and proprietary data, highlighting the need to integrate diverse data sources to fully leverage artificial intelligence capabilities. Nevertheless, restricted data accessibility remains a major challenge. In this regard, the formation of DCs emerges as a promising organisational mechanism to promote broader data sharing by activating data-unlocking mechanisms and enabling extensive data recombination – an essential process for generating data value (Alaimo et al., 2024) and thus amplifying AI's contribution to sustainable development. Our study serves as a reference point regarding the impact domains in which DCs have demonstrated their capacity to unlock privately held data. At the same time, it invites further reflection on the role of DCs as elements of the data value chain (Alaimo et al., 2020). While studies in this field have so far focused on an intra-organisational dimension, it may be important to study the data value chain at an inter-organisational level. Our findings also highlight the effectiveness of DCs as a governance mechanism for data commons, (Yakowitz, 2011) demonstrating their capacity to

Table 5 Cluster-specific research questions

Cluster	Cluster-specific research questions
1. Data-driven initiatives to support innovation	- What long-term impacts do these types of collaborations generate? - How do they incentivise data talent to participate, and how could these incentives be transferred to other contexts?
2. Collaborative efforts for large-scale research	- Why are public actors rarely involved in these collaborations? - What incentives drive private actors to participate in these collaborations, and how could they be replicated in different contexts? - What data-sharing agreements and infrastructures are used in these projects, and how could they be replicated in other contexts?
3. Continuous effort to improve systemic responses	- What business models could be adopted by organisations promoting these types of projects? - What role do impact measurements play in developing these collaborations?
4. Prompt response to emergencies	- What mechanisms may help preserve the knowledge, infrastructure and relationships developed during emergency situations beyond the emergency period? - What benefits could broader involvement of the not-for-profit sector bring to these projects?
5. International mobilisation for development	- What benefits could arise from greater engagement of local public administrations? - How can the capacity to manage these types of projects be maintained in the contexts in which they take place? - How can dependencies on foreign actors be reduced?

overcome existing barriers (Grossman, 2023) to the spread of the phenomenon when their creation is aimed at generating a positive social impact.

Limitations

The predominant use of quantitative methodology in data analysis and the exclusive reliance on secondary data sources are the main constraints of this study. Adopting a quantitative methodology to examine the configurations of DCs may have limited our understanding of the nuances inherent in various models, restricting our analysis to more generalised observations. Nonetheless, we have attempted to mitigate these limitations by incorporating insights from prior exploratory studies into our framework, as exemplified by Verhulst et al., (2019a, 2019b),, and by conducting multiple rounds of qualitative interpretation of the results. This interpretation involved assessing the meaning of different clustering options by comparing the identified solutions with the nature and operations of each DC, searching for qualitative explanations of the quantitative results. Relying exclusively on secondary data sources may have also affected the accuracy of the analysis and, consequently, our results. To mitigate this bias, we chose to eliminate from the analysis those projects for which information accessible online was insufficient to confidently populate the variables considered. The decision to exclude these DCs from the analysis reduced the number of cases considered, while simultaneously minimising the risk of populating the dataset with incorrect information, thereby enhancing the validity of the analysis. Given that previous studies on this subject have predominantly relied on a limited selection of case studies for in-depth analysis, we posit that our research serves as a complementary contribution to the existing body of knowledge.

Conclusions

Despite the growing attention given to the data-for-social-good phenomenon and the increasing number of initiatives around the world, the literature on DCs has struggled to develop empirical knowledge in the last few years. This is reflected in an empirical field where DCs often struggle to progress beyond the pilot stage (World Economic Forum 2021). To advance research on this topic ten years after the concept was first defined, this paper began by outlining six operational characteristics distinguishing DCs from open data projects and data ecosystems. Subsequently, the paper focused on analysing differences among projects within the DCs' perimeter. This analysis led, for the first time, to the identification of several impact categories. The results revealed the existence of five distinct DC clusters. By analysing these clusters across different organisational, technological, contextual and outcome variables – as well as conducting a cross-cluster analysis of the results – we were able to identify their unique characteristics and development challenges. To support this contribution, the research provides a list of cluster-specific research questions that could help researchers develop more focused empirical studies by selecting specific empirical settings aligned with their research objectives. The results could also assist practitioners and policymakers in identifying precise benchmarks for their projects, thereby supporting them in the design of their interventions. This research aims to serve as a starting point for further empirical and theoretical exploration of the phenomenon.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-025-00831-6.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Data availability The dataset used for the analysis can be accessed at the following link: https://zenodo.org/records/15092656.

Declarations

Conflict of interest I have no conflict of interest nor competing interest.

References

Alaimo, C., & Kallinikos, J. (2024). Data rules data rules. *MIT Press - Journals*. https://doi.org/10.7551/mitpress/11751.001.0001

Alaimo, C., Kallinikos, J., & Aaltonen, A. (2020). *Data and Value Handbook of Digital Innovation.* https://doi.org/10.4337/9781788119986.00022

Alves, G., Couceiro, M., & Napoli, A. (2019). Similarity measure selection for categorical data clustering. *HAL Archives-Ouverts. Fr*. hal-02399640.

Azzone, G. (2018). Big data and public policies: Opportunities and challenges. *Statistics and Probability Letters, 136*, 116–120. https://doi.org/10.1016/j.spl.2018.02.022

Berreteaga Barbero, A., Duisberg, A., Hoffmann, A., Duicuing, C., & Hommen, D. (2020). IDSA Rule book. (December), 1–60.

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceeding of the 2008 SIAM international conference on data mining* (pp.

243–254). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972788.22

Burnaby, T. P. (1970). On a method for character weighting a similarity coefficient, employing the concept of information. *Journal of the International Association for Mathematical Geology, 2*(1), 25–38. https://doi.org/10.1007/BF02332078

Castelnovo, W., Misuraca, G., & Savoldelli, A. (2016). Smart cities governance: The need for a holistic approach to assessing urban participatory policy making. *Social Science Computer Review, 34*(6), 724–739. https://doi.org/10.1177/0894439315611103

Charles, J., & Tonetti, C. (2020). Nonrivalry and the economics of data. *Jurnal Online Internasional & Nasional Vol. 7 No.1, Januari – Juni 2019 Universitas 17 Agustus 1945 Jakarta*, *53*(9), 1689–1699. Retrieved from https://www.journal.uta45jakarta.ac.id

Chui, M., Harryson, M., Manyika, J., Roberts, R., Chung, R., Nel, P., & van Heteren, A. (2018). *Applying AI for social-good: Discussion paper*. McKinsey Global Institute, 52. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

Corrales-Garay, D., Mora-Valentín, E. M., & Ortiz-de-Urbina-Criado, M. (2019). Open data for open innovation: An analysis of literature characteristics. *Future Internet*. https://doi.org/10.3390/fi11030077

Coulton, C. J., George, R., Putnam-Hornstein, E., & De Haan, B. (2015). *Harnessing big data for social good: A grand challenge for social work harnessing big data for social good: A grand challenge for social work*. American Academy of Social Work and Social Welfare.

Curioso, W. H., & Carrasco-Escobar, G. (2020). Collaboration in times of COVID-19: The urgent need for open-data sharing in Latin America. *BMJ Health and Care Informatics, 27*(1), 1–2. https://doi.org/10.1136/bmjhci-2020-100159

Digital Civil Society LAB. (2017). *Workshop summary: Trusted data intermediaries*. Medium. Retrieved from https://medium.com/the-digital-civil-society-lab/trusted-data-intermediaries-c58426c2c994

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Farmer, J., Mccosker, A., Albury, K., & Aryani, A. (2022). *Data for social good non-profit sector data projects*.

Fawcett, J., & Downs, F. S. (1986). *The relationship of theory and research*. Appleton-Century-Crofts. Retrieved from https://books.google.com/books?id=NqdqAAAAMAAJ

Flanagan Anne, J., & Sheila, W. (2022). *Advancing digital agency: The power of data intermediaries*.

Foucault, M. (1994). *The order of things: An archeology of the human sciences*. Routledge.

Fraiman, R., Justel, A., & Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association, 103*(483), 1294–1303. https://doi.org/10.1198/016214508000000544

Geisler, S., Vidal, M. E., Cappiello, C., Lóscio, B. F., Gal, A., Jarke, M., Lenzerini, M., Missier, P., Otto, B., Paja, E., Pernici, B., & Rehof, J. (2022). Knowledge-driven data ecosystems toward data transparency. *Journal of Data and Information Quality, 14*(1), 1–12. https://doi.org/10.1145/3467022

George, G., Howard-Grenville, J., Joshi, A., & Tihanyi, L. (2016). Understanding and tackling societal grand challenges through management research. *Academy of Management Journal, 59*(6), 1880–1895. https://doi.org/10.5465/amj.2016.4007

Global Partnership for Sustainable Development Data. (2023). *Effective and ethical data sharing at scale*.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642.

Grossman, R. L. (2023). Ten lessons for data sharing with a data commons. *Scientific Data, 10*(1), Article 120. https://doi.org/10.1038/s41597-023-02029-x

GSMA. (2018). *Scaling big data for social good: The need for sustainable business models*. Retrieved from https://www.gsma.com/betterfuture/bd4sg

Guggenberger, T. M., Schlueter Langdon, C., & Otto, B. (2025). Data spaces as meta-organisations. *European Journal of Information Systems, 00*(00), 1–21. https://doi.org/10.1080/0960085X.2025.2451250

Haak, E., Ubacht, J., Van Den Homberg, M., Cunningham, S., & Van Den Walle, B. (2018). A framework for strengthening data ecosystems to serve humanitarian purposes. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3209281.3209326

He, Z., Xu, X., & Deng, S. (2002). Squeezer: An efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology, 17*(5), 611–624. https://doi.org/10.1007/BF02948829

Hoffman, W., Bick, R., Boral, A., Henke, N., Olukoya, D., Rifai, K., Roth, M., & Youldon, T. (2019) *Collaborating for the common good: Navigating public- private data partnerships*.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management, 29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Jussen, I., Schweihoff, J., & Moller, F. (2023). Tensions in inter-organizational data sharing: Findings from literature and practice. In *Proceedings - 2023 IEEE 25th Conference on Business Informatics, CBI 2023*. https://doi.org/10.1109/CBI58679.2023.10187530

Jussen, I., Möller, F., Schweihoff, J., Gieß, A., Giussani, G., & Otto, B. (2024). *Issues in inter-organizational data sharing: Findings from practice and research challenges*. Data and Knowledge Engineering, 150(December 2023). https://doi.org/10.1016/j.datak.2024.102280

Karolinska Institutet. (2011, July 9). Mobile phone data in Haiti improves emergency aid. https://news.ki.se/mobile-phone-data-in-haiti-improves-emergency-aid

Klein, T., & Verhulst, S. (2017). Access to new data sources for statistics: Business models and incentives for the corporate sector.

Klievink, B., Van Der Voort, H., & Veeneman, W. (2018). Creating value through data collaboratives: Balancing innovation and control. *Information Polity, 23*(4), 379–397. https://doi.org/10.3233/IP-180070

Kretzschmar, M. E., Rozhnova, G., Bootsma, M. C. J., van Boven, M., van de Wijgert, J. H. H. M., & Bonten, M. J. M. (2020). Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study. *The Lancet Public Health, 5*(8), e452–e459. https://doi.org/10.1016/S2468-2667(20)30157-2

Liva, G., Micheli, M., Schade, S., Kotsev, A., Gori, M., & Codagnone, C. (2023). City data ecosystems between theory and practice: A qualitative exploratory study in seven European cities. *Data and Policy*. https://doi.org/10.1017/dap.2023.13

Macdonald, J. L., Green, M. A., Gibin, M., Leech, S., Singleton, A., & Longely, P. (2023). Local data spaces: Leveraging trusted research environments for secure location-based policy research in the age of coronavirus disease-2019. https://doi.org/10.1017/dap.2023.14

Microsoft. (2016). Microsoft, L V Prasad eye institute and global experts collaborate to launch microsoft intelligent network for eyecare.

One Mind. One Mind. https://onemind.org/

Munzert, S., Selb, P., Gohdes, A., Stoetzer, L. F., & Lowe, W. (2021). Tracking and promoting the usage of a COVID-19 contact tracing app. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-020-01044-x

Nikander, P., Eloranta, V., Karhu, K., & Hiekkanen, K. (2020). Digitalisation, anti-rival compensation and governance: Need for

experiments. In *Abstract from Nordic workshop on digital foundations of business, operations, and strategy* (pp. 1–6).
Niño, M., Zicari, R. V., Ivanov, T., Hee, K., Mushtaq, N., Rosselli, M., Sánchez-Ocaña, C., Tolle, K., Blanco, J. M., Illarramendi, A., Besier, J., & Underwood, H. (2017). Data projects for " social good ": Challenges and opportunities. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 11*(5), 896–906.
Oliveira, M. I., de Barros Lima, G. F., & Farias Lóscio, B. (2019). Investigations into data ecosystems: A systematic mapping study Knowledge and information systems (Vol. 61). Springer. https://doi.org/10.1007/s10115-018-1323-6
Oliveira, M. I. S., & Lóscio, B. F. (2018). What is a data ecosystem? *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3209281.3209335
Ooms, W., Caniëls, M. C. J., Roijakkers, N., & Cobben, D. (2020). Ecosystems for smart cities: Tracing the evolution of governance structures in a dutch smart city initiative. *International Entrepreneurship and Management Journal, 16*(4), 1225–1258. https://doi.org/10.1007/s11365-020-00640-7
Otto, B., & Hompel, M. (2022). *Designing Data Spaces Designing Data Spaces*. https://doi.org/10.1007/978-3-030-93975-5
Oxford Internet Institute. (2014). Bellagio big data workshop participants (2014). Big data and positive social change in the developing world: A white paper for practitioners and researchers. In *Rockefeller foundation Bellagio centre conference* (pp. 1–35).
Perkmann, M. (2016). *How Boundary Organizations Facilitate Collaboration across Diverse Communities*. https://doi.org/10.1093/acprof
Perkmann, M., & Schildt, H. (2015). Open data partnerships between firms and universities: The role of boundary organizations. *Research Policy, 44*(5), 1133–1143. https://doi.org/10.1016/j.respol.2014.12.006
Praditya, D., Janssen, M., & Sulastri, R. (2017). Determinants of business-to-government information sharing arrangements. *The Electronic Journal of E-Government, 15*(1), 44–56. Retrieved from https://www.ejeg.com
Ramon Gil-Garcia, J., Chengalur-Smith, I., & Duchessi, P. (2007). Collaborative e-Government: Impediments and benefits of information-sharing projects in the public sector. *European Journal of Information Systems, 16*(2), 121–133.
Rango, M., & Vespe, M. (2017). Big Data and alternative data sources on migration: From case-studies to policy support. *European Commission-Joint Research Centre (JRC)*. Retrieved from https://bluehub.jrc.ec.europa.eu/bigdata4migration/workshop-outcome
Rasche, A., Morsing, M., & Wetter, E. (2019). Assessing the legitimacy of "open" and "closed" data partnerships for sustainable development. *Business and Society*. https://doi.org/10.1177/0007650319825876
Řezanková, H. (2016). Cluster analysis and categorical data., (January 2009).
Ruijer, E. (2021). Designing and implementing data collaboratives : A governance perspective. *Government Information Quarterly*, 101612. https://doi.org/10.1016/j.giq.2021.101612
Savona, M. (2020). Governance of data value. *Policy Brief, Policy@Sussex, 1*, 14.
Schoormann, T., Strobel, G., Möller, F., Petrik, D., & Zschech, P. (2023). Artificial intelligence for sustainability—A systematic review of information systems literature. *Communications of the Association for Information Systems*. https://doi.org/10.17705/1CAIS.05209
Sieber, R. E., & Johnson, P. A. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly, 32*(3), 308–315. https://doi.org/10.1016/j.giq.2015.05.003
Signorelli, S., Fontana, M., Vespe, M., Gabrielli, L., & Bertoni, E. (2024). Towards a taxonomy for business-to-government data sharing. *Statistical Journal of the IAOS, 40*(3), 713–726. https://doi.org/10.3233/SJI-230122
Spiekermann, M. (2019). Data marketplaces: Trends and monetisation of data goods. *Intereconomics, 54*(4), 208–216. https://doi.org/10.1007/s10272-019-0826-z
Stalla-Bourdillon, S., Wintour, A., & Carmichael, L. (2019). Building trust through data foundations: A call for a data governance model to support trustworthy data sharing, 33. Retrieved from https://cdn.southampton.ac.uk/assets/imported/transforms/content-block/UsefulDownloads_Download/E2360AAB5D274223BFDB863BAFC20F34/White%20Paper%202.pdf
Stalla-Bourdillon, S., Carmichael, L., & Wintour, A. (2021). Fostering trustworthy data sharing: Establishing data foundations in practice. *Data & Policy, 3*, e4. https://doi.org/10.1017/dap.2020.24
Strava. Metro. https://metro.strava.com/
Šulc, Z., Cibulková, J., Procházka, J., & Řezanková, H. (2018). Internal evaluation criteria for categorical data in hierarchical clustering: Optimal number of clusters determination. *Metodološki Zvezki, 15*(2), 1–20.
Susha, I., Janssen, M., & Verhulst, S. (2017). Data collaboratives as a new frontier of cross-sector partnerships in the age of open data: Taxonomy development. In *Proceedings of the 50th Hawaii International Conference on System Sciences* (pp. 2691–2700). https://doi.org/10.24251/hicss.2017.325
Susha, I., Pardo, T. A., Janssen, M., Adler, N., Verhulst, S. G., & Harbour, T. (2018). A research roadmap to advance data collaboratives practice as a novel research direction. *International Journal of Electronic Government Research, 14*(3), 1–11. https://doi.org/10.4018/IJEGR.2018070101
Susha, I., & Gil-Garcia, J. R. (2019). A collaborative governance approach to partnerships addressing public problems with private data. In *Proceedings of the 52nd Hawaii international conference on system sciences* (pp. 2892–2901). https://doi.org/10.24251/hicss.2019.350
Susha, I., Grönlund, Å., & Van Tulder, R. (2019a). Data driven social partnerships: Exploring an emergent trend in search of research challenges and questions. *Government Information Quarterly, 36*(1), 112–128. https://doi.org/10.1016/j.giq.2018.11.002
Susha, I., Rukanova, B., Ramon Gil-Garcia, J., Tan, Y.-H., & Gasco, M. (2019b). *Identifying mechanisms for achieving voluntary data sharing in cross-sector partnerships for public good**. Association for Computing Machinery. https://doi.org/10.1145/3325112.3325265
Susha, I., Flipsen, M., Agahari, W., & De Reuve, M. (2020). Towards generic business models of intermediaries in data collaboratives. In *IFIP international federation for information processing 2020*.
Susha, I., van den Broek, T., van Veenstra, A.-F., & Linåker, J. (2022). An ecosystem perspective on developing data collaboratives for addressing societal issues: The role of conveners. *Government Information Quarterly, 40*(1), Article 101763. https://doi.org/10.1016/j.giq.2022.101763
The Gov Lab. (2020). Wanted: Data stewards.
The New Hanse Project. (2023). Governing urban data for the public interest.
van Loenen, B. (2006). *Developing geographic information infrastructures* (pp. 1–404). DUP Science, Delft University Press. Retrieved from http://www.library.tudelft.nl/dup/
van den Broek, T., & van Veenstra, A. F. (2018). Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation. *Technological Forecasting and Social Change, 129*, 330–338. https://doi.org/10.1016/j.techfore.2017.09.040
van Donge, W., Bharosa, N., & Janssen, M. F. W. H. A. (2022). Data-driven government: Cross-case comparison of data stewardship in data ecosystems. *Government Information Quarterly, 39*(2), 101642. https://doi.org/10.1016/j.giq.2021.101642
Varshney, K. R., & Mojsilovic, A. (2019). Open platforms for artificial intelligence for social good: Common patterns as a pathway to true impact. In *International conference on machine learning AI for social good workshop*.

Vercellis, C. (2009). *Business intelligence: Data minining and optimization for decision making*. John Wiley & Sons, Ltd.

Verhulst, S., & Young, A. (2011). *Open data impact when demand and supply meet* (pp. 1–13). Govlab. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

Verhulst, S., & Sangokoya, D. (2015). Mapping the next frontier of open data: Corporate data sharing | by Stefaan G. Verhulst | Internet Monitor 2014: Data and privacy | Medium. Retrieved March 12, 2021, from https://medium.com/internet-monitor-2014-data-and-privacy/mapping-the-next-frontier-of-open-data-corporate-data-sharing-73b2143878d2

Verhulst, S. G., & Young, A. (2016). Open data impact: When data and supply meet. Retrieved from https://thegovlab.org/static/files/publications/open-data-impact-key-findings.pdf

Verhulst, S. G., Engin, Z., & Crowcroft, J. (2019a). Data & policy : A new venue to study and explore policy–data interaction. *Data & Policy, 1*, 1–5. https://doi.org/10.1017/dap.2019.2

Verhulst, S. G., Young, A., Winowatan, M., & Zahuranec, A. J. (2019b). *Leveraging private data for public good: A descriptive analysis and typology of existing practices*. Retrieved from http://thegovlab.org/new-report-leveraging-private-data-for-publicgood/

Verhulst, S. G. (2021). Reimagining data responsibility: 10 new approaches toward a culture of trust in re-using data to address critical public needs. *Data & Policy*. https://doi.org/10.1017/dap.2021.4

Viale Pereira, G., Cunha, M. A., Lampoltshammer, T. J., Parycek, P., & Testa, M. G. (2017). Increasing collaboration and participation in smart city governance: A cross-case analysis of smart city initiatives. *Information Technology for Development, 23*(3), 526–553. https://doi.org/10.1080/02681102.2017.1353946

Whitelaw, S., Mamas, M. A., Topol, E., & Van Spall, H. G. C. (2020). Applications of digital technology in COVID-19 pandemic planning and response. *The Lancet Digital Health, 2*(8), e435–e440. https://doi.org/10.1016/S2589-7500(20)30142-4

Williams, S. (2020). Data action using data for public good paper knowledge. In *Toward a media history of documents* (Vol. 3).

World Economic Forum. (2021). Empowered data societies: A human-centric approach to data relationships. White Paper, (September).

Yakowitz, J. (2011). Tragedy of the data commons. In *Harvard Journal of Law & Technology* (Vol. 25). http://www.rand.org/publications/randreview/issues/rr-12-02/

Zuiderwijk, A. (2017). Analysing open data in virtual research environments: New collaboration opportunities to improve policy making. *International Journal of Electronic Government Research, 13*(4), 76–92. https://doi.org/10.4018/IJEGR.2017100105

Zygmuntowski, J. J., Zoboli, L., & Nemitz, P. F. (2021). Embedding European values in data governance: A case for public data commons. *Internet Policy Review, 10*(3). https://doi.org/10.14763/2021.3.1572