Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development

Iryna Susha
Örebro University
iryna.susha@oru.se

Marijn Janssen
Delft University of Technology
M.F.W.H.A.Janssen@tudelft.nl

Stefaan Verhulst
The Governance Lab of New York University
stefaan@thegovlab.org

Abstract

Data collaboratives present a new form of cross-sector and public-private partnership to leverage (often corporate) data for addressing a societal challenge. They can be seen as the latest attempt to make data accessible to solve public problems. Although an increasing number of initiatives can be found, there is hardly any analysis of these emerging practices. This paper seeks to develop a taxonomy of forms of data collaboratives. The taxonomy consists of six dimensions related to data sharing and eight dimensions related to data use. Our analysis shows that data collaboratives exist in a variety of models. The taxonomy can help organizations to find a suitable form when shaping their efforts to create public value from corporate and other data. The use of data is not only dependent on the organizational arrangement, but also on aspects like the type of policy problem, incentives for use, and the expected outcome of data collaborative.

1. Introduction

Access to new datasets has the potential to improve people's lives and to support policy making by enabling evidence-based and more agile decisions. However, access to important datasets is often hard to get. The open data movement has led to governments worldwide sharing their data. Yet many datasets that could help solve public problems are proprietary. Accelerating data sharing and collaboration between those who hold valuable data and those able to deliver solutions is key to reaping the public value from data.

A number of initiatives have emerged recently to harness the benefits of (corporate) data sharing for public good. For example, in the Netherlands Liander, an energy provider, shared their data on energy consumption to spur innovation and smarter energy use. Another example is Statistics Netherlands (CBS), who partnered with the mobile phone company Vodafone to analyze mobile call records to better understand mobility patterns and inform urban planning.

Data-driven collaboration between sectors for public good has been termed differently in the community of practitioners, e.g. as "data philanthropy" [1] or "data collaborative" [2]. In this research we adopt the term "data collaboratives" proposed by Verhulst and Sangokoya [2] in previous writings, because it emphasizes the process of collaboration between parties and thus suggests a more encompassing view going beyond data sharing. We define data collaboratives *as cross-sector (and public-private) collaboration initiatives aimed at data collection, sharing, or processing for the purpose of addressing a societal challenge.* In this definition an essential element is that organizations from different sectors collaborate together to create value from data. Both business and government can share data; however data shared by the private sector for public good is of particular interest to us, as much of the data which is critical for addressing societal challenges of today rests in private hands [3].

Understanding the emerging eco-system of data collaboratives is important for a variety of reasons. First, we witness a resurgence of attention towards evidence-based policy making. Unfortunately, many – especially developing – countries have limited access to datasets that can provide for a deeper understanding of their own society to determine what intervention may work best. Second, advances in technology have radically changed how data is collected, stored and analyzed, yet the impact of the big data era has so far been limited as it relates to improving people's lives. Third, there is not only a re-distribution of who collects and has access to data – where governments have traditionally been dominant players – but also a re-distribution of skills and talent to analyze the data with corporations having superior data analytics capabilities as opposed to government. Finally, governments have started to share and open up their own data, yet the real value of open data often comes from integrating government data with non-government data sources and from having a close partnership between the supply and demand of data. Data collaboratives are at the nexus of these four developments and when designed well can radically improve the impact data may have on the public good.

Data collaboratives can be viewed as a new frontier in big and open data research for several reasons. First,

URI: http://hdl.handle.net/10125/41481
ISBN: 978-0-9981331-0-2

they are specifically aimed at helping solve complex societal problems. This addresses one of the main challenges the open data movement has faced to date – achieving high-impact results and solving pressing societal problems with data [4]. Besides, the defining characteristic of a data collaborative is its focus on realizing public benefits, rather than commercial innovation, as was the case in the early days of open data. Second, the data in a data collaborative can come from different sources: private or public sectors, as well as from non-profit or academic stakeholders. A recent study of data for policy initiatives in the EU [5] showed that presently public datasets are the main data source used for policy making. Using sources of data, such as call details records, social media feeds, or sensors is relatively new. These new data may have varying degrees of openness, e.g. provided only to certain users or provided as processed insights. This goes beyond the usual focus on open government data and beyond the definition of open data as free to access, use, modify, share for any purpose by anyone [6]. In addition, collaboration was found to be one of the main challenges which (big) data initiatives for public good currently face [7]. This concerns collaboration between data scientists, domain experts, policy makers, and local experts. Therefore, research on data collaboratives as a new form of cross-sector and public-private collaboration is particularly needed.

In particular, there is a need to describe and analyze data collaboratives in a more systematic and structured manner. In this paper we take first steps towards creating systematic knowledge and structuring this emerging field. The purpose of this research is to develop a taxonomy which distinguishes among different forms of data collaboratives. A taxonomy (also sometimes referred to as a typology or classification) is a system for grouping objects of interest in a domain based on common characteristics [8]. We expect the taxonomy to be useful to three user groups: researchers, policy-makers, and companies potentially interested to (learn how to) share data. The taxonomy can be useful to them for determining the most suitable data collaborative form given the circumstances and goals of the different parties involved. Since the target users of our taxonomy are both potential providers and users of data in a data collaborative, we chose to differentiate among the different forms of data collaboratives on the basis of *the relationship between demand and supply of data*. We expect that the taxonomy can be used to answer a number of important questions about data collaboratives. How can data be shared? For what purpose can the data be used? How open are companies in terms of data sharing for good? For researchers such a taxonomy provides insight into the diversity of data collaboratives and can be used to evaluate their effectiveness.

2. Related work

At the time of writing, a search in Google Scholar, Scopus, and Web of Science for the term "data collaborative" in the title returned only 3 relevant results in the academic literature. A search for the term "data philanthropy" in the same databases in the title returned 2 results. Data collaborative as a new organizational form was described in studies of the MetroGIS initiative in the state of Minnesota dating back to 1996 [9, 10]. This initiative was a collaboration between geospatial data producers and user communities to enable more efficient sharing of georeferenced data. In healthcare the initiatives known as "data collaboratives" primarily focus on large scale data collection, such as the Perinatal Staffing Data Collaborative in the US [11] or the more recent Health Data Collaborative of the World Health Organization[1]. Another report describes a similar data-collection-focused initiative in education in the US – the Education Data Collaborative [12], which provided a single database of student and teacher performance for near-real-time monitoring. As one can see from the low number of found publications, the concept of data collaborative has received marginal attention in the academic literature. To date no efforts have been made in the academic literature to build a taxonomy of data collaboratives or systematize what is known about them otherwise. However, valuable insights can be gained from the grey literature in this respect which highlights the importance of this phenomenon.

With the focus on corporate data sharing, Verhulst and Sangokoya [13] proposed six data collaborative forms: research partnerships, prizes and challenges, trusted intermediaries, application programming interfaces (API), intelligence products, and corporate data pooling. This taxonomy was drawn from anecdotal examples of data collaboratives and based on a mix of characteristics, such as with whom the data is shared, for what purpose, and in what way. In this taxonomy the forms overlap, as for example in a challenge competition an API may be provided.

A study commissioned by the OECD [14] examined public-private partnerships to leverage new sources of data for statistics and discussed several models. These included in-house data analysis by data provider, transfer of datasets to the user, transfer of datasets to a trusted third party, and outsourcing of data collection functions. These four models are based on the

[1] http://www.healthdatacollaborative.org/

characteristics of data sharing protocols, i.e. how much is shared, with whom, and at what stage in the data process. This analysis is limited to statistics agencies as the user of data. In our study we aim to take a more encompassing view and consider different user groups. However, we limit our analysis to the cases of data sharing by the private sector with governmental, non-profit, or academic stakeholders, as explained in the Introduction.

Besides grey literature, it is essential to contextualize data collaboratives in related research domains to see what can be learnt from them for taxonomy purposes. Based our proposed definition of data collaboratives, we consider this concept to be founded on two main research domains: cross-sector social partnerships (CSSP) and open and big data. Research on CSSPs, which is rather mature, offers a number of ways to categorize partnerships: there exist taxonomies of CSSPs organized around who the actors are, types of resources exchanged, characteristics of agreement, level of intensity such as commitment and engagement, dynamics and time dimension of CSSPs [15]. On the other hand, research on open and big data has just taken root and offers predominantly exploratory results in terms of taxonomies. Existing taxonomies are organized around how data is collected and opened [16], in which format it is provided [17], and how data can be used [18]. Hilbert [19] classified data sources based on the content and what they capture: words, locations, behavior, transactions, production, nature, or other. Furthermore, a report by Vaitla [7] distinguished between the different tracking technologies to capture these data: data exhaust (e.g. locations captured by call details records); online activity (e.g. data from social networks or web searches); sensing technology (e.g. data captured by satellites or personal sensors).

On a more general note, it is also helpful to refer to the literature discussing the data lifecycle from a process perspective. For example, Bizer, et al. [20] identified six steps for dealing with data: data capturing, data storage, data searching, data sharing, data analysis, and data visualization. Chen, et al. [21] used three steps – data handling, data processing, data moving – and Marx [22] proposed five steps – problem definition, data searching, data transformation, data entity resolution, answering the query/solving the problem. In this paper we make a difference between data sharing and usage phases which is similar to data supply and demand. This division is useful as in data collaboratives different parties are involved in the supply and use of data.

3. Research method

To develop a taxonomy of data collaborative forms we used the Taxonomy Development Method formulated by Nickerson, et al. [8]. According to these authors, this is the first comprehensive effort to formalize the process of taxonomy development in Information Systems as a method. Previous studies largely relied on ad hoc approaches. This method has been successfully applied by a number of studies, e.g. to classify crowdsourcing processes [23], web-based inbound open innovation initiatives [24], and health 2.0 collaboration platforms [25].

The first step in this method is identifying a meta-characteristic, the most comprehensive characteristic to serve as a basis for the choice of characteristics in the taxonomy. In our study we view the relationship between demand and supply of data as the meta-characteristic of data collaboratives. This means the taxonomy is expected to convey characteristics of data collaboratives related to data supply (the sharing aspect) and demand for data (the use aspect).

The second step is defining the ending conditions for terminating the iterations in developing the taxonomy (Ibid.). The objective ending conditions for our method are: all cases in the sample have been examined; and there is no duplication of dimensions or characteristics. The subjective ending conditions for finishing the analysis are: the taxonomy is determined to be concise, robust, comprehensive, extendible, and explanatory (Ibid.).

The third step is choosing either the empirical-to-conceptual or conceptual-to-empirical approach. We started our analysis with the empirical-to-conceptual approach, during which we identified a sample of cases to infer the characteristics and dimensions of the taxonomy. Since much of the development in the data collaboratives field takes place in practice, starting taxonomy building with an inductive approach was considered most appropriate. When applying the empirical-to-conceptual approach we used a subset of the sample of data collaboratives cases found in the Data Collaboratives Directory[2]. As of April 2016, this database contained 23 cases in five different domains: Health (10), Economic Development (3), Education (3), Environment (4), and Infrastructure (3).

In the first (A) iteration we used a convenience sample of five cases which were selected from each of the five domains in the Directory (see Table 1 in the Annex). To develop an understanding of each case we produced short summaries based on the official webpages of these initiatives. In the second (B) iteration

[2] Compiled by The Governance Lab

we examined the second subset of cases, five more selected according to the same principle as in the first iteration, to determine whether the existing characteristics and dimensions are sufficient to describe them. By selecting cases from different domains we aimed to increase the representativeness of the sample. Thus, the cases relate to different types of data shared, different users, and different purposes of use.

After two iterations of the empirical-to-conceptual approach, we opted to use the conceptual-to-empirical approach. First, we made sure that we covered the dimensions identified by previous studies classifying data collaboratives mentioned in section 2. Then we conducted a search for additional articles which are not necessarily focused on taxonomies but discuss data-driven collaboration in general. A search in Scopus – using the terms "big data" or "open data" in combination with the terms "collaboration" or "partnership" in the title – returned 24 publications. 3 of them were relevant to us and offered insights as to the additional dimensions and/or characteristics of the data collaboratives taxonomy.

4. Findings

In what follows we present the dimensions of our resultant taxonomy. Using the empirical-to-conceptual approach (analysis of cases in Table 1), we identified ten dimensions of data collaboratives (see Table 2 in the Annex). The following four relate to the first part of the meta-characteristic – *data sharing and supply* in data collaboratives (S stands for Sharing):

S1 **Type of data**: data about natural persons, legal persons (e.g. movement of fishing vessels in the Global Fishing Watch case), or natural phenomena (e.g. data on the amount of sunshine provided in the Orange case). The characteristic of natural persons is further divided into: (a) consumer data – data collected, or "observed" [26], about people's activities without their explicit knowledge (e.g. locations of mobile phone users); (b) user-generated data – data provided by individuals explicitly (e.g. social network data); and (c) volunteered data – data provided by individuals on volunteer basis (e.g. patient data as was in the Clinical Trials case).

S2 **Content of data**: words, locations, behavior, transactions, or nature [19]. Transaction data concern data about people's activities in a commercial setting as a customer (e.g. details of trips by Uber). Behavioral data, on the other hand, concern data about people's actions in a non-commercial situation (e.g. as a patient in the Clinical Trials case).

S3 **Administrative level associated with data**: specific (e.g. call details records in Bangladesh in the MDEEP case) or unspecific (e.g. social networks data without relevance to a particular country in the DERP case).

S4 **Diversity of data providers**: one provider (e.g. a solo initiative of one company such as in the Twitter case), several providers from the same industry (e.g. an alliance of companies within one field such as in the Clinical Trials case), or several companies from different industries (e.g. companies offering data in different domains, as was the case in the Telecom Italia case).

The following six dimensions and characteristics relate to the second part of the meta-characteristic – the *demand and data use* in data collaboratives (U stands for Use):

U1 **Target user group of data**: academic, commercial, governmental, non-profit partners, or citizens. In certain cases, such as Global Fishing Watch, participants of a data collaborative do not strictly define the target group and provide data or data insights to all kinds of users, including citizens.

U2 **User selection**: on agreement basis (e.g. users of the data selected based on partnership agreements with their respective institutions), on application basis (e.g. users of the data selected based on individual applications), or open (e.g. not requiring any specific selection procedure).

U3 **Research or policy problem**: specified (e.g. by requesting a research proposal, as in the DERP case) or unspecified (e.g. by opening data for any type of innovative reuse, as in the Yelp case).

U4 **Incentive for data use**: tangible (e.g. monetary reward) or intangible (e.g. to break new ground in science).

U5 **Continuity of collaboration**: on demand (e.g. data shared when it is requested), event-based (e.g. data shared in the framework of a competition or other event), or continuous (e.g.

data shared continuously as it becomes available).

U6 **Outcome of data collaborative**: policy intervention, data science, or data-driven innovation. Policy intervention is further divided into sub-characteristics: prediction and alerts (i.e. using data insights as early warning signals), needs-based planning (i.e. using data to learn about people's needs for aid planning), capacity building (i.e. using data to identify areas to improve government response), and monitoring (i.e. using data to track compliance with policies).

As a result of using the conceptual-to-empirical approach (literature review), we obtained additional insights and identified two more dimensions of data collaboratives. Vale [27] discussed an international collaboration initiative between statistics agencies focused on exploring the use of new data sources for statistics purposes. This case led us to include an additional dimension into the Data Use section of the taxonomy:

U7 **Collaboration among data users**: one user (i.e. data is shared with one organization), self-selected analysis by several users (i.e. several teams use the data for different policy or research issues, as in the Orange case), or collaborative analysis by several users (i.e. several teams use data to analyze one specific policy or research problem, as in the MDEEP case).

Furthermore, the study of open data partnerships between firms and universities by Perkmann and Schildt [28] discussed the role of "boundary organizations" as intermediaries. These are intermediaries who can perform the tasks of "mediated revealing" (i.e. aggregating and anonymizing datasets before transfer to the user) and of enabling multiple goals (i.e. ensuring a win-win situation for both data provider and user) (Ibid). We find this dimension relevant, which was included in the Data Sharing section of the taxonomy as follows:

S5 **Facilitation**: self-facilitated (i.e. direct contact with the user without an intermediary), intermediary with data-related functions (i.e. an intermediary may pre-process data or provide a technology solution for sharing), or intermediary with organizational functions (i.e. an intermediary may play a coordinating role ensuring participation of provider and user).

To assess the usefulness of the taxonomy we held an initial evaluation session with experienced open data researchers (8 persons) at Delft University of Technology on 26 May 2016. The participants were presented with the taxonomy and were asked to fill in an evaluation form. The form included questions about the different aspects of usefulness, as defined by Nickerson, et al. [8] in the discussion of subjective ending conditions: namely, to what extent the researchers found the taxonomy concise, robust, complete, and explanatory. Based on the results of this evaluation, we merged some of the characteristics to make the taxonomy more concise and thus easier to comprehend and use. During the group discussion we identified two additional dimensions:

S6 **Degree of access to data**: real-time direct access to raw data, direct access to a copy of raw data, access to modified or enriched data, access to outcomes of processed data, or data shared as open data [29].

U8 **Purpose of use**: primary (i.e. data is used for the purpose for which it was collected), secondary (i.e. data is used for the purpose which is similar to the one for which it was collected), tertiary (i.e. data is used for a different purpose than for which it was collected), or end use (i.e. data is processed and the result is used by end users) [30]. In our sample we did not have a case which concerns the primary purpose of use, but an example given in section 2 (Health Data Collaborative) fits this category. An example of the secondary purpose of use is the Clinical Trials case, in which clinical trials data, collected for medical research by the funders of trials, was used for medical research but by other researchers. In comparison, in the MDEEP case the call details records, collected to gain insights about customers, was used to infer population movement in relation to a disaster (tertiary use).

To summarize, Figure 1 below presents all identified dimensions of the taxonomy based on (a) whether they relate to data or collaboration aspect of a data collaborative and (b) whether they relate to the supply or demand side. Two dimensions – Degree of access and Facilitation – are placed in the middle, as they can be viewed as mechanisms to match the supply and demand in data collaboratives.

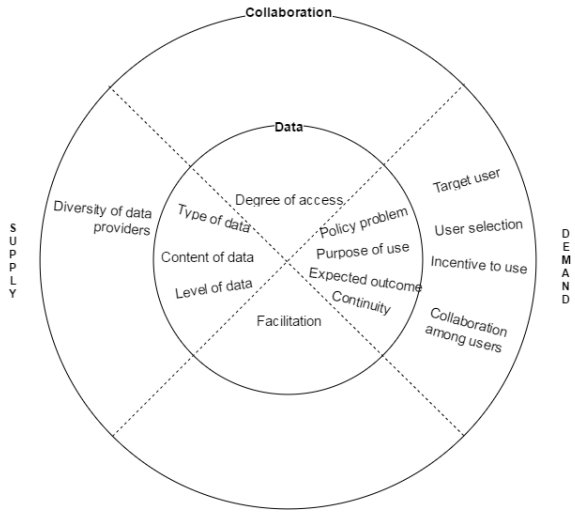

Figure 1. Dimensions of the taxonomy of data collaboratives based on the supply-demand relationship

5. Discussion

The taxonomy shows that data collaboratives are not a homogeneous phenomenon and are characterized by complex interdependencies. It is this complexity that results in value creation, and a taxonomy structuring the different dimensions of data sharing and use is highly needed.

The taxonomy sheds more light on some important questions we raised in the Introduction, namely how data can be shared and for what purpose it can be used, and how open the private sector is when it comes to data collaboratives. We will discuss these three questions hereafter. Besides, we are able to make several observations about the relationships between the different dimensions and characteristics of the taxonomy and the implications of that.

First, we discuss the kinds of data that data collaboratives can address. Dimensions S1 and S2 show that data collaboratives can be distinguished based on what type of data is provided and what it contains. The two dimensions are related, as for example consumer data about natural persons mainly concern details of people's transactions with a certain service (e.g. details of trips in the Uber case) but may also include words (e.g. search terms in the Google case). While user-generated data about natural persons mainly concern textual data consciously published online by people but can also include locations (e.g. check-ins at restaurants in the Yelp case). This means that the same type of data – e.g. locations or words – may be gathered from the public domain or may be gathered as the so-called "data exhaust" as part of consumer analytics. This would have implications on the extent to which such data can be used in terms of privacy issues. As a result, in a data collaborative scenario it is important to differentiate the origins of data and how it was collected. This relates to the dimension U8 Purpose of use which captures to what extent the data is used according to the purpose for which it was collected. From our sample, we can see that initiatives involving tertiary use of data (use for a different purpose than for which it was collected) are more common. We therefore encourage research into data collaboratives which focus on purposeful collaborative data collection (primary purpose of use), in a similar vein with the healthcare cases mentioned in section 2.

Second, the taxonomy shows that there are various ways to organize the sharing of data – in terms of who the recipient is (dimension U1), how they are selected (U2), how much detail is provided to them (S6). The same or similar types of data can be shared using a different data sharing mechanism. For example, both the MDEEP and Telecom Italia cases shared call detail records but differed in terms of target user group, user selection strategy, and incentives for use. The taxonomy also shows that the private sector can provide various degrees of access to their data – ranging from making available select insights from data to select users to making data available to anyone by publishing it as open data. Which degree of access is chosen in a data collaborative depends on several factors, such as the type of data, the purpose of use, and the expected outcome of the data collaborative. We can see that academic users are the most common target user group (in 8 out of 10 cases). There are however examples in which users of data in two or more sectors are targeted, as in the case of Telecom Italia. The user selection procedures vary depending on the context, however selection by application is more common towards academic users (in cases such as DERP or Clinical Trials case) and selection by agreement can involve partners in all sectors.

Third, with respect to the question for what purpose the data can be used in a data collaborative, we can distinguish among data collaboratives based on three characteristics (U6) – policy intervention, data science, or data-driven innovation. The expected outcome of the data collaborative relates to who the participants of the data collaborative are. For examples, initiatives focusing on innovation are more likely to target a broader range of potential users and offer rewards for participation.

Since in our sample we included cases from five different domains – Health, Economic Development, Education, Environment, and Infrastructure – we were able to evaluate differences across these domains in terms of the features of data collaboratives. Based on

our sample, we can conclude that no specific type of data was relevant only for one particular domain. For instance, the cases in our sample in the domain of Health used volunteered patient data and social networks data for their respective objectives. However, we can observe commonalities across the domains in terms of the outcomes of data collaboratives. For example, in the domain of Infrastructure data collaboratives can be particularly useful for needs-based planning, as can be seen from the Orange and Uber cases.

On a final note, we find our taxonomy to be different from the existing one [13] in several ways: it is more detailed, derived both empirically and theoretically, and developed in a systematic way. We recommend to use our taxonomy for further research in order to test it on a different sample of cases.

6. Conclusions

This paper has argued that data collaboratives are important 21st century experiments in cross-sector and public-private partnerships exchanging data for public good to address complex societal problems. There is a wide diversity of forms of data collaboratives but a few models are emerging around how data is provided and how it is used. Data collaboratives are often created by corporations and third parties across sectors as new ways to signal social responsibility, yet several other incentives come into play such as reciprocity and revenue generation.

In this research we systematically developed a taxonomy of data collaboratives as a new form of collaboration towards addressing societal challenges by leveraging data. The purpose of the taxonomy is to distinguish among different forms of data collaboratives based on how data is shared (supply) and how data is used (demand). Based on the analysis of ten cases and relevant literature, we identified fourteen dimensions which can be used to differentiate data collaboratives.

Our taxonomy shows that data collaborative is a concept encompassing various organizational forms in which data sharing and data use can be organized in a number of ways. The choice of how data is shared in a data collaborative involves considering such aspects, as the type, content, and administrative level of data; degree of access to it, diversity of data providers, and facilitation mode. The use of data, on the other hand, vary depending on the policy or research problem, purpose of use, target user and user selection, incentives for use, expected outcome of data collaborative, and continuity of collaboration. Some data collaboratives might look similar at a first glance, but differ on a few aspects of our taxonomy. Each different form might have different benefits and disadvantages.

The limitations of our study are that, for practical reasons, we focused on the initiatives in which data is shared by the private sector with government, academic, or non-profit partners. Also our sample is not all encompassing, yet it was designed to represent the diversity of practice and data. We plan to test the taxonomy using a larger sample of cases in our future research.

We also anticipate that the rapidly changing technological landscape can affect some of the underlying variables of our taxonomy. Namely, such developments as the Internet of Things, augmented reality apps, or live streaming offer an opportunity to collect hybrid types of content of data about users (e.g. words, behavior, location, nature at the same time in live streaming). If shared in a data collaborative scenario, this hybrid data content will add an extra layer of complexity. The dimension S2 of our taxonomy may be revisited to account for that. In addition, developments in artificial intelligence and other areas of data science may impact the type of analysis data collaboratives seek to conduct. Another issue we anticipate is the developments in data ethics and responsible data sharing. At present in many cases the boundaries between consumer, user-generated, and volunteered data (dimension S1 of the taxonomy) are somewhat blurred. Most often the data subjects are not aware of how their data is used by the service provider and give consent to privacy policies without diving into details. This may change as new policies, practices and standards emerge in the national and international arena around data ownership and data governance.

All in all, data collaboratives have the potential to radically re-distribute power relations as it relates to data in society, and developing a deeper understanding of current practices will be key to inform future directions. Our taxonomy scratched the surface of this emerging eco-system, and future research can provide more understanding with regard to a number of issues. These include, but are not limited to, impact of data collaboratives, influential factors, incentives for sharing, governance processes, risk mitigation strategies, and supply-demand matching infrastructures.

7. Acknowledgements

This research was funded by the Swedish Research Council under the grant agreement 2015-06563 as part of the project “Data collaboratives as a new form of innovation for addressing societal challenges in the age of data”.

Annex

Table 1. Cases examined in iteration 1(1A-5A) and 2 (6B-10B) of the empirical-to-conceptual approach

No	Cases	Short description	Domain
1A	Google Flu Trends	An initiative by Google to offer real-time search trends data to a number of academic partners for flu and dengue research (re-launched in 2015)	Health
2A	Yelp Dataset Challenge	A challenge competition organized by Yelp offering user-generated data about local businesses to students and researchers for cash rewards (held annually since 2011)	Economic development
3A	Digital Ecologies Research Partnership (DERP)	An initiative offering researchers access to data from a number of online communities for researching social dynamics on the web (launched in 2014)	Education
4A	Mobile Data, Environmental Extremes, and Population (MDEEP) Project	An initiative of a consortium of international partners which uses call details records to understand climate impacts by mapping population flows before and after an extreme weather event (active in 2013-2014)	Environment
5A	Orange Telecom Data for Development Challenge	An innovation challenge organized by Orange, first in the Ivory Coast and thereafter in Senegal, offering anonymized call details records to international research institutions for addressing a range of development-related problems (since 2012)	Infrastructure
6B	Clinical Study Data Request Program	An ongoing initiative to provide interested researchers with clinical trials data from a number of pharmaceutical companies on an application basis	Health
7B	Telecom Italia Big Data Challenge	An innovation challenge hosted by Telecom Italia who, in cooperation with other companies, offered data on mobile calls, energy, local news, and weather to academic and commercial participants in order to advance competitiveness of Italy (held in 2014 and 2015)	Economic development
8B	Twitter-MIT Lab for Social Machines	An ongoing initiative sponsored by Twitter who provide MIT Media Lab scientists with access to Twitter data for studies of public opinion, journalism, governance, and human development	Education
9B	Global Fishing Watch	An ongoing initiative of Google, Oceana, and SkyTruth to visualize satellite data of the movement of commercial fishing vessels around the globe	Environment
10B	Uber – City of Boston Partnership	An initiative of Uber to provide anonymized trip-level data to the City of Boston to support city planning and transportation (active in 2015)	Infrastructure

Table 2. Taxonomy of data collaboratives derived from using the empirical-to-conceptual and conceptual-to-empirical approaches

No	Dimensions	Characteristics	Sub-characteristics	Cases in iteration 1					Cases in iteration 2				
				1A	2A	3A	4A	5A	6B	7B	8B	9B	10B
Data sharing and supply													
S1	Type of data	Natural persons	Consumer data	x			x	x		x			x
			User-generated data		x	x					x		
			Volunteered data						x				
		Legal persons										x	
		Natural phenomena											
S2	Content of data	Words		x	x	x				x	x		
		Locations			x		x	x		x		x	
		Behavior							x				
		Transactions			x					x			x
		Nature						x		x			
S3	Administrative level associated with data	Specific			x		x	x		x			x
		Unspecific		x		x			x		x	x	
S4	Diversity of data providers	One provider		x	x		x	x			x		x

		Several providers from same industry			x			x			x		
		Several providers from different industries							x				
S5	Facilitation	Self-facilitated	x	x		x	x					x	
		Intermediary with data-related functions							x	x	x		
		Intermediary with organizational functions			x			x					
S6	Degree of access to data	Real-time direct access to raw data								x			
		Direct access to a copy of raw data		x	x				x				
		Access to modified or enriched data				x	x	x	x			x	
		Access to outcomes of processed data	x								x		
		Data shared as open data							x				
Data use and demand													
U1	Target user group	Academic	x	x	x	x	x	x	x	x			
		Commercial							x				
		Governmental	x									x	
		Non-profit				x							
		Citizens									x		
U2	User selection	On agreement basis	x			x				x		x	
		On application basis			x		x	x					
		Open		x					x		x		
U3	Research or policy problem	Specified	x		x	x					x	x	
		Unspecified		x			x	x	x	x			
U4	Incentive to use data	Tangible		x			x		x				
		Intangible	x		x	x		x		x	x	x	
U5	Continuity of collaboration	On demand			x			x				x	
		Event-based		x		x	x		x				
		Continuous	x							x	x		
U6	Expected outcome of data collaborative	Policy intervention — Prediction and alerts	x										
		Policy intervention — Needs-based planning					x					x	
		Policy intervention — Capacity building				x							
		Policy intervention — Monitoring									x		
		Data science			x			x		x			
		Data-driven innovation		x					x				
U7	Collaboration among data users	One user								x		x	
		Self-selected analysis by several users	x	x	x		x	x	x				
		Collaborative analysis by several users				x					x		
U8	Purpose of data use	Primary											
		Secondary		x	x			x					
		Tertiary	x			x	x		x	x		x	
		End use									x		

8. References

[1] R. Kirkpatrick, "Big data for development," *Big Data*, vol. 1, pp. 3-4, 2013.

[2] S. Verhulst and D. Sangokoya, "Data Collaboratives: Exchanging Data to Improve People's Lives," in *Medium* vol. 2015, ed, 2015.

[3] B. Noveck. (2015, 10 May). *Data Collaboratives: Sharing Public Data in Private Hands for Social Good*. Available: http://www.forbes.com/sites/bethsimonenoveck/2015/09/24/private-data-sharing-for-public-good/#28dab08b65bb

[4] I. Susha, "Participation in open government," PhD, Department of Informatics, Örebro University, Örebro, 2015.

[5] M. Poel, R. Schroeder, J. Treperman, M. Rubinstein, E. Meyer, B. Mahieu, *et al.*, "Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking," Data for Policy2015.

[6] Open Definition. (n.d., 27 May). *Open definition*. Available: http://opendefinition.org/

[7] B. Vaitla, "The Landscape of Big Data for Development," 2014.

[8] R. C. Nickerson, U. Varshney, and J. Muntermann, "A method for taxonomy development and its application in information systems," *Eur J Inf Syst*, vol. 22, pp. 336-359, 05//print 2013.

[9] R. Johnson, "Minnesota MetroGIS geospatial data collaborative Minneapolis-St. Paul metropolitan area (2002--Enterprise System)," *URISA Journal*, vol. 17, pp. 41-46, 2005.

[10] I. Masser and R. Johnson, "Implementing SDIs through effective networking: the MetroGIS geospatial data collaborative," *GeoInformatics*, vol. 9, pp. 50-53, 2006.

[11] B. Scheich and D. Bingham, "Key Findings from the AWHONN Perinatal Staffing Data Collaborative," *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, vol. 44, pp. 317-328, 2015.

[12] J. Byrd, "EDUCATION DATA COLLABORATIVE," 2011.

[13] S. Verhulst and D. Sangokoya. (2014, 20 August 2015). Mapping the Next Frontier of Open Data: Corporate Data Sharing. *Internet Monitor 2014: Data and Privacy*. Available: http://bit.ly/1EKIVSq

[14] N. Robin, T. Klein, and J. Jütting, "Public-Private Partnerships for Statistics: Lessons Learned, Future Steps," 2016.

[15] C. Vurro, M. T. Dacin, and F. Perrini, "Institutional Antecedents of Partnering for Social Change: How Institutional Logics Shape Cross-Sector Social Partnerships," *Journal of Business Ethics*, vol. 94, pp. 39-53, 2010.

[16] M. Janssen, R. Matheus, and A. Zuiderwijk, "Big and Open Linked Data (BOLD) to Create Smart Cities and Citizens: Insights from Smart Energy and Mobility Cases," in *International Conference on Electronic Government*, 2015, pp. 79-90.

[17] E. Kalampokis, E. Tambouris, and K. Tarabanis, "A classification scheme for open government data: Towards linking decentralised data," *International Journal of Web Engineering and Technology*, vol. 6, pp. 266-285, // 2011.

[18] T. Davies, "Open data, democracy and public sector reform," Msc, Social Science of the Internet, University of Oxford, 2010.

[19] M. Hilbert, "Big Data for Development: A Review of Promises and Challenges," *Development Policy Review*, vol. 34, pp. 135-174, 2016.

[20] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The meaningful use of big data: four perspectives -- four challenges," *SIGMOD Rec.*, vol. 40, pp. 56-60, 2012.

[21] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, pp. 171-209, 2014.

[22] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, pp. 255-260, 2013.

[23] D. Geiger, S. Seedorf, T. Schulze, R. C. Nickerson, and M. Schader, "Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes," in *AMCIS*, 2011.

[24] F. von Briel and C. Schneider, "A Taxonomy of Web-Based Inbound Open Innovation Initiatives," 2012.

[25] N. Kordzadeh and J. Warren, "Toward a typology of health 2.0 collaboration platforms and websites," *Health and Technology*, vol. 3, pp. 37-50, 2013.

[26] L. Taylor, J. Cowls, R. Schroeder, and E. T. Meyer, "Big data and positive change in the developing world," *Policy and Internet*, vol. 6, pp. 418-444, 2014.

[27] S. Vale, "International collaboration to understand the relevance of Big Data for official statistics," *Statistical Journal of the IAOS*, vol. 31, pp. 159-163, 2015.

[28] M. Perkmann and H. Schildt, "Open data partnerships between firms and universities: The role of boundary organizations," *Research Policy*, vol. 44, pp. 1133-1143, 2015.

[29] M. Janssen, R. Matheus, and A. Zuiderwijk, "Big and Open Linked Data (BOLD) to Create Smart Cities and Citizens: Insights from Smart Energy and Mobility Cases," in *Electronic Government*, ed: Springer, 2015, pp. 79-90.

[30] B. Loenen, *Developing geographic information infrastructures: the role of information policies*: IOS Press, 2006.