Data collaboratives as “bazaars”?

A review of coordination problems and mechanisms to match demand for data with supply

Susha, Iryna; Janssen, Marijn; Verhulst, Stefaan

DOI
10.1108/TG-01-2017-0007

Publication date
2017

Document Version
Accepted author manuscript

Published in
Transforming Government: people, process and policy (online)

Citation (APA)
Susha, I., Janssen, M., & Verhulst, S. (2017). Data collaboratives as “bazaars”? A review of coordination problems and mechanisms to match demand for data with supply. Transforming Government: people, process and policy (online), 11(1), 157-172. https://doi.org/10.1108/TG-01-2017-0007

1. Introduction

Access to new datasets has the potential to improve people's lives and to support policy making by enabling evidence-based decisions (Jetzek, Avital, & Bjørn-Andersen, 2014). However, gaining access to important datasets is often hard. The open data movement has led to governments worldwide disclosing and sharing their data (Zuiderwijk & Janssen, 2014). Yet many datasets that could help to solve societal problems are proprietary, from which many are not owned by the government. Accelerating data sharing and collaboration between those who hold valuable data and those able to deliver solutions is key to reaping value from data.

Data-driven collaboration between sectors for public good has been termed differently in the community of practitioners, e.g. as "data philanthropy" (Kirkpatrick 2013) or "data collaborative" (Verhulst and Sangokoya 2015). The first term suggests that costly data is given way, whereas the second term stresses the collaboration. In this research we adopt the term "data collaboratives" as proposed by Verhulst and Sangokoya (2015), because it emphasizes the process of collaboration between parties, which goes beyond mere 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"* (Susha, Janssen, and Verhulst, 2017, p. 2691). 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 as this has been given limited attention so far. Whereas much of the data which is critical for addressing societal challenges of today rests in private hands (Noveck, 2015).

A number of initiatives have emerged recently to harness the benefits of (corporate) data sharing for public good. For example, Statistics Netherlands (CBS) partnered with the mobile phone company Vodafone to analyze mobile call records to better understand mobility patterns and inform urban planning (see http://bit.ly/2jUg85C). However, not all data collaboratives achieve their goals as planned. For instance, the partnership between Uber and the City of Boston was compromised due to the fact that the data shared did not correspond to the exact needs of the city (see http://bit.ly/2kEu3Q2). This shows that coordinating activities, goals, and resources of participating actors in a data collaborative is very important yet challenging for achieving desired outcomes and creating value with data collaboratives. In a data collaborative information needs to flow in a coordinated fashion through a multi-organizational and multi-level arrangements. The success of a data collaborative does not only depend on internal interactions, but also on the interactions with other organizations. In fact, collaboration was found to be one of the main challenges which (big) data initiatives for public good currently face (Vaitla, 2014). This concerns collaboration between data stewards, data scientists, domain experts, policy makers, and local experts distributed over the independent organizations in a data collaborative. Therefore, understanding how their activities are coordinated is needed.

In practice there is a need to understand what coordination problems can be expected when initiating a data collaborative and what means are available to address

them. Coordination may require additional efforts or even costs and thus should be accounted for. So far there has not been any studies dedicated to the coordination of data collaboratives. In our previous research (Susha, Janssen, & Verhulst, 2017) we developed a taxonomy which distinguishes among different forms of data collaboratives. This paper reuses this taxonomy to analyze the coordination challenges and mechanisms of data collaboratives from the perspective of coordination theory. The purpose of this study is to identify and conceptualize common coordination problems which can be expected in a data collaborative and to propose a selection of potential coordination mechanisms which can be used to mitigate these problems.

The paper is structured as follows: first, we provide an overview of the data collaboratives concept from the literature; second, we describe the basics of coordination theory; third, we outline our method; then, we discuss our findings and close with concluding remarks about the value and implications of the analysis of coordination.

2. Data collaboratives in the literature

There are only a few works about data collaboratives in the academic literature, considering the definition of data collaboratives proposed in this study. Data collaborative as a new organizational form was described in studies of the MetroGIS initiative in the state of Minnesota dating back to 1996 (Johnson 2005; Masser and Johnson 2006). 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 (Scheich and Bingham, 2015) or the more recent Health Data Collaborative of the World Health Organization (see http://www.healthdatacollaborative.org/). Another report describes a similar data-collection-focused initiative in education in the US – the Education Data Collaborative (Byrd et al., 2011), 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. However, experimentation in practice is growing, as the new resource Data Collaboratives Explorer shows (see http://datacollaboratives.org/). In our previous research we proposed a taxonomy (Susha et al., 2017) which differentiates between the different characteristics of data collaboratives based on how the data is shared from the supply side and on how the data is used on the demand side. The taxonomy was derived from the analysis of real-life cases and classified using the relevant academic literature.

<Insert Table 1 about here>

The taxonomy of Susha et. al (2017) 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 varies 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 one or more aspects of our taxonomy. Each different form might have different benefits and disadvantages. In this study we are interested in how the identified characteristics are related to coordination problems and mechanisms of data collaboratives. Hence we will use this taxonomy as a backdrop for our analysis of coordination.

More specifically, in relation to the theme of coordination, we did not find any prior studies dedicated specifically to that in the body of academic literature on data collaboratives. However, grey literature offers a number of valuable points of departure. Latonero & Gold (2015) discuss the problem of coordination between stakeholders from different spheres of expertise in data philanthropy projects (which they term as "cross disciplinary translation work"). They discuss the need to bridge technological expertise with context-specific knowledge of the problem and of affected populations and propose to do so by using brokers or "translators" as they label them. Another report (Data-Pop Alliance, 2015, p.35) highlights the importance and need for coordination in the context of big data for resilience projects and makes several recommendations, such as creating new avenues and means of cooperation for stakeholders, facilitating exchange with affected communities, promoting common standards governing data format, documentation, access. This shows that the challenge of coordinating activities in a data collaborative is recognized in practice but poorly understood in research. Therefore, our study aims to bridge this gap and contribute to knowledge about how coordination in data collaboratives can be improved.

3. Research approach

In our study we used comparative case analysis to infer and illustrate coordination challenges in the context of data collaboratives. We reused the sample of ten cases from the study of Susha et al. (2017). The ten cases were derived from the Data Collaboratives Explorer (see http://datacollaboratives.org/) which is the first emerging repository of data collaborative initiatives. The cases were selected by the principle of diversity to include two cases per each of the five listed domains: Health, Economic Development, Education, Environment, and Infrastructure. Table 2 in the Appendix contains short descriptions of the cases studied. To identify the coordination mechanisms, we used coordination theory.

3.1 Coordination theory

The choice of coordination theory in this study was motivated by the following considerations. Coordination theory is quite generic and thus applicable to different

domains. It offers a concrete framework for analyzing the components of coordination and a range of potential coordination mechanisms to apply. This theory has been tested and found useful in previous studies, such as to study coordination in the open data process (Zuiderwijk & Janssen, 2013), organizational process design (Crowston, 1997), distributed group decision making (Cao, Burstein, & San Pedro, 2004). The study of Zuiderwijk & Janssen (2013) is particularly useful to us, as it examines coordination in the context of (open) data sharing similar to our study. Data collaboratives involve a different mix of actors and goals, but some similarities in terms of the process may be observed.

Coordination is a broad concept occurring at various levels (Comfort, Dunn et al., 2004). A widely accepted definition of coordination is "the managing of dependences between entities" (Malone and Crowston, 1994). Malone and Crowston (1990) argue that the need for coordination arises from constraints imposed on the performance of tasks by the interdependent nature of these tasks. They view coordination at the task level in which coordination mechanisms manage the interdependencies among tasks. Besides, there exist a lot of coordination mechanisms at the organizational level in which coordination is organized through institutional arrangements that regulate the positions and relations between parties (Koppenjan and Groenewegen, 2005).

At the organizational level, coordination theory draws the attention to the differences between markets, hierarchies and networks (Malone, Yates et al. 1987; Clemons, Reddi et al. 1993). In a hierarchy the flow of data is coordinated by having hierarchical levels in which decisions are made (Malone, Yates et al., 1987). Markets coordinate the exchange of data through supply and demand forces and external transactions between entities (Malone, Yates et al., 1987). Networks or network organizations are positioned in between and can potentially overcome the problems with hierarchies and create greater structural effectiveness and responsiveness (Powell, 1990). A network is often characterized by long-term, relatively stable partnerships to increase resources utilization through greater explicit coordination and short-term relationships for incidental transactions. Clemons, Reddi et al. (1993) use the term explicit coordination to distinguish networks from the implicit coordination of the 'invisible hand' of market competition. They define explicit coordination as the extent to which decisions reflect and are tailored to a specific relationship. These three models should be perceived as 'ideal' types since hybrid forms are common. More recently, Demil & Lecocq (2006) proposed a fourth model - '*bazaar*' - which has distinct features in terms of coordination. These authors characterize coordination in the four models in terms of (1) means of communication (coordination mechanism governing the exchange), (2) intensity of incentives, and (3) intensity of control. The features of the four models – hierarchy, market, network, and bazaar – are captured and contrasted in Table 3 below.

<Insert Table 3 about here>

At the task level, the need for coordination arises when multiple, interdependent activities are performed to achieve goals (Malone and Crowston, 1990). At the task level coordination can be analysed from the perspective of actors who perform certain activities to achieve certain goals; these activities are not independent and are

characterized by interdependencies. Thus, coordination theory aims at identifying what kinds of interdependencies between activities are possible and how different kinds of interdependence can be managed (ibid). The interdependence between activities can be analyzed in terms of "common objects" that are involved in some way in both actions (ibid). In the case of data collaboratives, the common object is the data which is shared by one actor and then used by another.

Malone and Crowston (1994) identified three types of interdependencies regardless of the domain: prerequisite, shared resource, simultaneity. Based on that, they proposed four generic ways to manage the interdependencies between activities performed by different actors: 1 – take it or leave it (low coordination), 2 – negotiation, 3 – transfer of knowledge between the parties, and 4 – third party steps in to coordinate. March and Simon (1958) present two types of coordination processes to achieve organized behavior: coordination by plan and coordination by feedback. Mintzberg (1983) describes three kinds of coordination mechanisms: mutual adjustment, direct supervision and standardization. Mutual adjustment achieves coordination using a process of informal communication. Direct supervision achieves coordination by having one person or organization take responsibility for the work of others by issuing instructions and monitoring. Standardization achieves coordination by standardization of work processes, skills or output. These coordination mechanisms are quite generic and can be easily transferred to different organizational settings. Besides, there can be more than one coordination mechanism which is appropriate for a given coordination problem (Crowston, 1997).

5. Findings: coordinating data collaboratives

In this section, we discuss our proposed taxonomy in Table 1 from the perspective of coordination theory. This will enable us to identify coordination problems associated with data collaboratives and available coordination mechanisms to mitigate these problems. For this, we use the analysis framework of Malone and Crowston (1994) which views coordination from the perspective of coordination components: actors, goals, activities, resources, and dependencies among them. The first step is to identify how the dimensions of our taxonomy relate to these coordination components (see Table 4 below).

<Insert Table 4 about here>

The second step is to identify dependencies with regards to the actors, resources, activities, and goals in the context of data collaboratives. This resulted in Figure 1 which is structured by using the coordination dimensions on the vertical axis (i.e. actors, resources, activities and goals). The numbers in circles show the 5 coordination problems which are derived by looking at the interdependencies between the actors, resources, activities and goals. The numbers in squares indicate the dimensions of the data collaboratives taxonomy which are relevant to the different components of coordination.

<Insert Figure 1 about here>

Hereafter we describe the five coordination problems in more detail and discuss the coordination mechanisms which can be used to deal with them. The identified coordination problems are common and occur in all data collaboratives, although their relevance, impact and how the problem is tackled varies. Moreover, the choice of the appropriate coordination mechanism depends on the characteristics of the data collaborative. Therefore, we will discuss the coordination mechanisms in light of the relevant characteristics of the taxonomy.

Coordination problem 1. Matching potential data providers and data users. There is a need for an arrangement to facilitate the finding of partners in data collaboratives. This dependency occurs between the actors involved in a data collaborative. Data collaboratives as a form of partnership are open for participation to actors from various sectors. The initiative to organize a data collaborative may come from the side of data provider, data user or data intermediary. There is a need for rules on how organizations collaborate. Coordinating different parties involved in a data collaborative becomes more challenging when multiple data providers or multiple data users are involved (dimensions S4 and U7 of the taxonomy). The boundaries of responsibilities between them are often unclear. In such cases the need for an intermediary becomes more pronounced (dimension S5).

Coordination mechanisms. A number of different coordination mechanisms can be identified that are currently used in data collaboratives based on our sample. The first such mechanism is 'take it or leave it', when data collaboratives are initiated by data providers who invite a relatively open audience of interested data users to access the data shared. In the taxonomy the user selection dimension (U2) illustrates this nuance. The innovation challenges – such as Yelp Dataset Challenge, Telecom Italia Big Data Challenge, and Orange D4D Challenge – fall into this category (open user selection). This mechanism favors low coordination costs, but does not provide incentives to stimulate the use of data. The second coordination mechanism to mitigate the problem of finding partners for a data collaboratives is mutual adjustment/negotiation/coordination by plan. In this case data providers and interested data users negotiate and collectively define the terms of engagement (user selection on agreement basis). Typically, such data collaboratives are bilateral partnerships, such as Uber – City of Boston partnership, Twitter – MIT Lab for Social Machines, and Google Flu Trends. The third coordination mechanism we find in our sample of cases is transfer of knowledge between the parties (in combination with third party coordination), when data providers share what they can offer and under what conditions, while data users share their detailed intentions on how they plan to use the data. An intermediary can then be involved to coordinate the matching of organizations and their data provided from the two sides. The corresponding taxonomy characteristic is user selection by application (U2). Examples of data collaboratives using this coordination mechanism to match data providers with data users are DERP and Clinical Study Data Request Program. Both examples employ a third party intermediary to manage the data requests

and the provision of access to the data. Moreover, the process for the data users to request access to the data is fairly standardized, which points to standardization of process as an additional coordination mechanism for matching data providers and users.

Coordination problem 2. Maintaining control over the data and its unforeseen uses once it has been shared. The sharing of data in a data collaborative can be quite complex; the data may contain personal (customer) information or other sensitive information for the data provider. Privacy agreements might have to be signed or data might need to be anonymized and aggregated before it can be shared. If shared improperly or 'leaked' into the public domain, this data may be used in harmful ways. This coordination problem becomes more complex depending on the purpose of use (dimension U8 in the taxonomy). For example, data collected for one purpose but shared and used for a completely different purpose (tertiary purpose of use) may pose a challenge of ensuring proper consent of the individual described in the data. The issue of data ownership is often unclear as the data changes hands from the data provider to the user.

Coordination mechanisms. Overall, there is a lack of responsible data principles and comprehensive data governance frameworks across different data collaborative initiatives (Berens, Mans, & Verhulst, 2016). However, coordinating this dependency by standardization of norms can provide a much-needed pathway for data collaboratives practice. To mitigate this problem, coordination by plan is often used, whereby the parties agree about the terms of use of the data (as is the case in most of the examples in our sample). In more complex arrangements it is also possible that the data provider retains some control over how the data is used by restricting its use to a secure digital environment (such as in the case of Clinical Data Request Program). This is an example of coordination by supervision to mitigate the lack of control of the data. Also another potential coordination mechanism in this situation is coordination by transfer of knowledge to a designated specialist in the private or public domain responsible for data sharing, or 'data steward' (Verhulst, 2016).

Coordination problem 3. Matching a particular research/policy problem with the specific attributes of the data required. One of the biggest challenges of data collaboratives is finding and accessing the data which has the right attributes to answer a certain research or policy question. It can also be the other way around – formulating the right question which the available data is capable to give a credible answer to. For instance, our taxonomy shows that there is a wide variety of type and content of data collected by the private sector, ranging from customer transaction records to satellite data (dimensions S1, S2, S3 of the taxonomy). To be appropriate for a certain research problem, it may be required that the data is of specific granularity, describes a certain geographic area, that it is of sufficient quality to draw valid conclusions. Data attributes like these just mentioned may have been predefined when the data was collected or may be open for negotiation in terms of how much a company would be willing to share.

Coordination mechanisms. In our sample of cases we can identify a number of coordination mechanisms used to manage this coordination problem. A number of data

collaboratives use mutual adjustment (or negotiation) to match the available data and problem. For instance, in the cases of MDEEP, Uber – City of Boston Partnership, Google Flu Trends, the attributes of the data (to a varying degree in each case) were discussed with the interested data user. Another group of data collaboratives (Yelp Dataset Challenge, Orange D4D, Telecom Italia Big Data Challenge) rely on the most simplistic coordination mechanism of 'take it or leave it' by offering their data to all interested entities in the framework of an event. The attributes of the data (content, type, level, degree of access) are decided upon by the data provider. Thus, there is no explicit coordination prior to the sharing of data as to what particular needs or questions the target user group might have. However, since some of these events are annual there is potential for implementing coordination by feedback. The low coordination approach does not necessarily mean a less successful data collaborative. The contrary can happen: organizing a collaboration around reusing readily available datasets in new unplanned ways, without explicit coordination with the data provider(s), may lead to unexpected value. For instance, the Global Fishing Watch case, among other data sources, uses the data from Automatic Identification Systems transmitted by vessels which is openly available. This required little coordination since the data is readily available and access should not be negotiated. But the value in this case was created by providing a central point of access to end users to search, browse and otherwise explore the data on a global scale through a user-friendly interface.

Coordination problem 4. Making sure the data shared by the data provider is useful and usable by the target user. One of the main challenges of data collaboratives is ensuring that the characteristics of the data shared meet the needs of the data users. In a data collaborative, the activity of sharing data serves as input for the activity of using data, which is a prerequisite coordination dependency.

Coordination mechanisms. Generally, the coordination mechanisms which can be used to address prerequisite dependencies are 1) ensuring usability of the resource, 2) managing transfer properly, and 3) adding a separate activity to remedy conflicts if any (Crowston, 1997). Often the data obtained through a data collaborative is combined by the user with other data sources; this poses the question of interoperability. Good meta data descriptions can smoothen the transfer and use of data. To coordinate the transfer of data in the shape and form useful to the user, the mechanism of the transfer of knowledge can be useful, whereby data providers can offer 'data playbooks', tutorial videos or the like which describe how the data is shared. For instance, in the case of Orange D4D Challenge the details of the datasets and how they were prepared for sharing, including the anonymization techniques used, were provided to the participants.

Coordination problem 5. Aligning incentives for data providers to share proprietary data with the goals of data users. The activity of sharing data in a data collaborative is conditional on the goal of data provider and the incentives to donate data for a societal purpose. In terms of goals (dimensions U3, U4, U6 and U8 in the taxonomy), data collaboratives can have diverse objectives, such as advancing scientific research, spurring data-driven innovation, or informing policy decisions and interventions. The

participating actors can have different objectives and incentives for sharing or using data. Some types of data collaboratives have a more data-driven approach and a looser problem formulation (dimension U3), while others have a more problem-driven approach and a more targeted application of the data.

Coordination mechanisms. As a general rule, for the private sector it is counter-intuitive to share proprietary information for free for various reasons, one of them being the competitive advantage of the company. On the other hand, sharing such data can contribute to the image of the company and show that the company takes its corporate responsibility. In a data collaborative it is thus important to formulate the value proposition for all participating sides in mutually beneficial terms. In our taxonomy we only considered the incentives for data users to use the data, which is certainly a limitation. However, we observe that to manage the coordination problem of aligning incentives of the parties in a data collaborative, the following coordination mechanisms can be used: coordination by feedback, transfer of knowledge between parties, and third party coordination.

Table 5 below summarizes the coordination problems and mechanisms identified on the basis of the taxonomy and case analysis.

< Insert Table 5 about here>

6. Discussion

Our analysis of coordination problems, based on the taxonomy, shows that data collaboratives are not a homogeneous phenomenon and are characterized by complex interdependencies which result in the creation of value. We identified five coordination problems arising from the dependencies between the activities, resources, and goals of data providers and users in a data collaborative. Although many coordination problems can be common to different 'configurations' of data collaboratives, they may require different solutions (coordination mechanisms) depending on the characteristics of the data collaborative. To this end, we used our previously developed taxonomy of data collaboratives to tap into some of these nuances.

To frame the discussion of our results, we turn to the conceptualization of coordination at a higher (organizational) level. When examining the coordination of data collaboratives at the organizational level, of interest are the institutional arrangements by way of which the positions and relationships between the actors in a data collaborative are regulated. The literature in section 3 shows that the main types of such institutional arrangements are hierarchy, markets, networks, and bazaar. The principal difference between the models is rooted in how the flow of goods or services (data sharing in the case of data collaboratives) in the value-added chain is coordinated (Malone, Yates, Benjamin, 1987). Demil & Lecocq (2006) propose to compare these models in terms of coordination by focusing on three characteristics: means of communication, incentives, and intensity of control. The means of communication relates to several of our identified coordination problems in the previous section.

Namely, problems 1, 2 and 4 – matching the actors, the problems, and attributes of the data in a data collaborative. The remaining two concepts of Demil & Lecocq correspond to the coordination problems 3 and 5, respectively. Thus, discussing these three variables is appropriate here, as they can be seen as an organizing framework for the discussion of the five coordination problems analyzed in the previous section.

With regards to data collaboratives, we find that data collaboratives bear much resemblance to the bazaar model. The main 'means of communication' for the transaction (data sharing) between the parties in a data collaborative is the attributes of the product (data), rather than price or established routines. The incentives for companies to share data for free are low and the intensity of control over the data after it has been transferred are low as well. Identifying data collaboratives as bazaars sheds light on a number of important features and deepens our understanding of data collaboratives. Namely, data collaboratives as bazaars are distinguished by the lack of defined work roles (unlike hierarchies) (Demil & Lecocq, 2006) and by the fuzzy boundaries of responsibilities. A range of intrinsic incentives come into play when companies choose to share, or 'donate', their data for a data collaborative (reciprocity, corporate responsibility). Unlike markets, data collaboratives as bazaars are much more good-driven than profit-driven; and this creates a coordination problem of matching goals and identifying appropriate incentives for the private sector. Finally, data collaboratives show a hybrid of bazaar and network forms of coordination – in the aspect of identity and selection of participants. In current practice of data collaboratives the 'motor' of exchange in data collaboratives as bazaars are the attributes of the data but also the identities of the parties (a number of recognized global players such as UN Global Pulse, World Bank, UNICEF, UNOCHA have emerged). In other words, data providers and users find each other based on their existing network, reputation, but also based on finding new connections by following the data they need. This network is however not open to anyone (as in the pure bazaar form), but different forms of data collaboratives show various levels of entry from lower (e.g. Yelp Dataset Challenge) to higher threshold (e.g. MDEEP case) for joining a data collaborative. Finally, the relationship among the parties in a data collaborative does not have to be a long-term engagement, as is often in networks (see Clemons et al.,1993). However, in practice some data collaboratives are more established and longer running than others. All this points to a number of implications, namely that data collaboratives as bazaar forms of coordination pose a high level of uncertainty in the transaction and the outcome.

Our research helps to further progress the development of data collaboratives in practice. Organizations initiating a data collaborative can be made aware of the coordination problems they will encounter in advance. Subsequently they can tackle the coordination problems by looking at possible coordination mechanisms and how they are used in other initiatives as described in this study. Although some of the coordination problems may be recognized in practice, there is often hardly any explicit mechanism in place to deal with them. Therefore, the practical value of this work lies in breaking down the complexities of coordination of data collaboratives and proposing a number of alternatives to manage common problems. The results of our study can be used to identify appropriate coordination mechanisms for a particular problem given certain characteristics of the project and thus improve coordination.

In terms of theoretical value, this is the first study to investigate coordination problems in the context of data collaboratives in more detail. This is also one of the few studies applying coordination theory to the domain of big/open data (together with Zuiderwijk & Janssen (2013) and Espinosa & Armour (2016)). We find the use of this theory useful and recommend further research to adapt/extend it. For instance, our study found that some coordination mechanisms (such as coordination by direct supervision) do not apply in the context of data collaboratives. Moreover, in our study we make a proposition that data collaboratives fall in the category of bazaars thus opening an arena for future debate. The implications of this proposition are that bazaar forms of coordination require a distinct approach to deal with, or rather capitalize, on the low levels of control and incentives. We invite future research to look into this further by developing design requirements which can be helpful when initiating data collaboratives as bazaars.

Our work shows that data collaboratives is a heterogeneous concepts and there is no single best form. Coordination problems depend on the situation and can be solved in a variety of ways. Success and value creation depends on how well the coordination problems are solved. We recommend further research in this area to understand value creation of data.

6. Conclusions

Organizations need to coordinate their efforts to create value from data. Data collaboratives are a novel form of partnerships between different sectors to leverage data for social good which encounter five coordination problems. Coordination is a particularly challenging issue for data collaboratives because of the vast amounts of data out there in the private sector, the complexity of societal challenges to be solved, and the diversity of stakeholders potentially involved. In this paper we tapped into the main coordination problems concerning data collaboratives and made a proposal for potential high-level solutions.

To identify coordination problems, we used coordination theory and did a secondary analysis of ten cases and of our previously developed taxonomy of data collaboratives. As a result, five main coordination problems were identified: 1) matching potential data providers and data users, 2) maintaining control over the data and its unforeseen uses, 3) matching a particular problem with the attributes of the data, 4) ensuring the usability and usefulness of the data to the user, and 5) aligning incentives of data providers with the goals of the users.

To tackle these coordination problems, we discussed a variety of coordination mechanisms which can potentially be used. In our discussion we showed that although coordination problems may be common to all forms of data collaboratives, data collaboratives with different characteristics may require different coordination mechanisms to managing these problems. Some of the mechanisms are already in place in different cases of data collaboratives, but some are yet to be further explored in practice, such as e.g. standardization of norms and transfer of knowledge (to data stewards) to tackle lack of control over the data. We find that in many cases the actors

rely on the 'take it or leave it' mechanism of low coordination by adopting a data-driven approach. This may not always be beneficial as coordination requires interaction and adjustment of supply and demand.

Data collaboratives exhibit a bazaar form of coordination. In data collaboratives the matching is often defined by what kind of data is on offer, and the incentives and control are low. The bazaar form of coordination is mostly associated with the open source movement but we see similar patters with regards to data collaboratives. Future research can investigate this proposition in more detail. An important next step is to investigate how data collaboratives should be designed to make the most out of the bazaar form of coordination. Formulating design requirements for data collaboratives is thus recommended to future research.

By analyzing coordination problems and how these are solved, a better grip on value creation mechanism is created. The practical value of our analysis lies in providing an overview of the complexities which can be expected when initiating a certain type of data collaborative. In this way our work contributes to better creation of value from data. We recommend taking a coordination view and studying how coordination mechanisms contribute to the creation of value in future research.

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".

Appendix

<Insert Table 2 about here>

References

Byrd, J. Daggett, W., Silver, D., and Williams, C. (2011), "Education Data Collaborative", available at: http://www.spnlive.org/spn/media/files/articles/research/EDC%20Research%20Report_Final%20Electronic_031811.pdf (accessed 26 January 2017)

Cao, P.P., Burstein, F. and San Pedro, J. (2004), "Extending coordination theory to the field of distributed group multiple criteria decision-making", *Journal of decision systems*, Vol. 13 No.3, pp.287-305.

Clemons, E. K., Reddi, S.P., and Row, M.C. (1993), "The Impact of Information Technology on the Organization of Economic Activity: The "move to the middle" hypothesis", *Journal of Management Information System*, Vol. 10 No. 2, pp.9-35.

Comfort, L., Dunn, M., Johnson, D., Skertich, R., and Zagorecki, A. (2004), "Coordination in Complex Systems: increasing efficiency in disaster mitigation and response", *International Journal of Emergency Management*, Vol. 2 No. 2, pp. 63- 80.

Crowston, K. (1997), "A Coordination Theory Approach to Organizational Process Design", *Organization Science*, Vol. 8 No. 2, pp.157-175.

Data-Pop Alliance (2015), "Big data for climate change and disaster resilience: Realising the benefits for developing countries", available at: http://datapopalliance.org/wp-content/uploads/2015/11/Big-Data-for-Resilience-2015-Report-Final.pdf (accessed 6 February 2017)

Demil, B. and Lecocq, X., (2006), "Neither market nor hierarchy nor network: The emergence of bazaar governance", *Organization studies*, Vol.27 No.10, pp.1447-1466.

Espinosa, J.A., Armour, F. (2016), "The big data analytics gold rush: A research framework for coordination and governance", in *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 1112-1121.

Jetzek, T., Avital, M., and Bjørn-Andersen, N. (2014), "Data-Driven Innovation through Open Government Data", *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 9 No.2, pp.100-120.

Johnson, R. (2005), "Minnesota MetroGIS geospatial data collaborative Minneapolis-St. Paul metropolitan area (2002--Enterprise System)", *URISA Journal,* Vol.17 No.2, pp.41-46.

Kirkpatrick, R. (2013), "Big Data for Development", available at: http://online.liebertpub.com/doi/pdfplus/10.1089/big.2012.1502 (accessed 27 January 2016)

Koppenjan, J. and Groenewegen, J. (2005), "Institutional design for complex technological systems", *Int. J. Technology, Policy and Management,* Vol.5 No.3, pp.240-257.

Latonero, M. and Gold, Z. (2015), "Data, Human Rights & Human Security", available at: https://ssrn.com/abstract=2643728 (accessed 6 February 2017)

Malone, T. and Crowston, K. (1990), "What is coordination theory and how can it help design cooperative work systems?", in *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, ACM.

Malone, T. W. and K. Crowston (1994), "The Interdisciplinary Study of Coordination", *ACM Computing Surveys (CSUR),* Vol.26 No. 1, pp. 87 - 119.

Malone, T. W., Yates, Y., and Benjamin, R.I. (1987), "Electronic Markets and Electronic Hierarchies", *Communications of the ACM,* Vol.30 No.6, pp.484-497.

March, J. G. and Simon, H.A. (1958), *Organizations*, John Wiley & Sons Inc.

Masser, I. and Johnson, R. (2006), "Implementing SDIs through effective networking: the MetroGIS geospatial data collaborative", *GeoInformatics*, Vol.9 No.6, pp.50-53.

Mintzberg, H. (1983), *Structure in fives: designing effective organizations*, Englewood Cliffs, NJ, Prentice-Hall.

Noveck, B. (2015), "Data Collaboratives: Sharing Public Data in Private Hands for Social Good", available at: http://www.forbes.com/sites/bethsimonenoveck/2015/09/24/private-data-sharing-for-public-good/ - 28dab08b65bb (accessed 10 May 2016)

Powell, W. W. (1990), "Neither Market Nor Hierarchy: Network Forms of Organization", *Research in Organizational Behavior,* Vol.12, pp.295-336.

Scheich, B. and Bingham, D. (2015), "Key Findings from the AWHONN Perinatal Staffing Data Collaborative", *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, Vol.44 No.2, pp.317-328.

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 in Waikoloa, Hawaii*.

Vaitla, B. (2014), "The Landscape of Big Data for Development", available at: http://data2x.org/wp-content/uploads/2014/08/Data2X_LandscapeOfBigDataForDevelopment.pdf (accessed 26 January 2017)

Verhulst, S. (2016), “Data Responsibility: a new social good for the information age”, available at: http://theconversation.com/data-responsibility-a-new-social-good-for-the-information-age-67417 (accessed 27 January 2017)

Verhulst, S. and Sangokoya, D. (2015), "Data Collaboratives: Exchanging Data to Improve People’s Lives", available at: https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a 2015 (accessed 26 January 2017)

Zuiderwijk, A., & Janssen, M. (2013), “A coordination theory perspective to improve the use of open data in policy-making”, in *Proceedings of the 12th Conference on Electronic Government*, Koblenz, Germany.
Zuiderwijk, A., & Janssen, M. (2014), “Open data policies, their implementation and impact: A framework for comparison”, *Government Information Quarterly*, Vol.31 No.1, pp.17-29. doi:http://dx.doi.org/10.1016/j.giq.2013.04.003

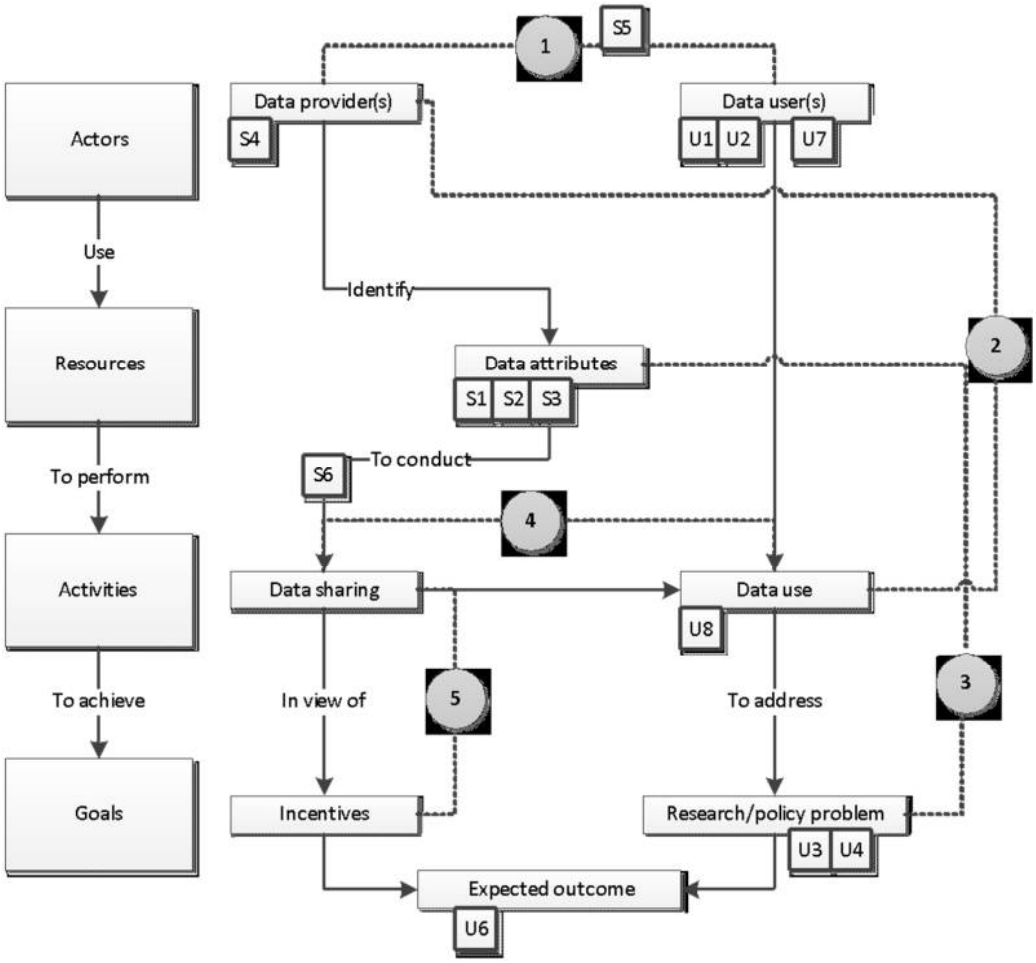

Figure 1. Dependencies among coordination components of data collaboratives

Table 1. Taxonomy of data collaboratives based on the characteristics of data sharing and data use (taken from Susha et al., 2017)

No	Dimensions	Characteristics	Sub-characteristics
Data sharing and supply side			
S1	Type of data	Natural persons	Consumer data
			User-generated data
			Volunteered data
		Legal persons	
		Natural phenomena	
S2	Content of data	Words	
		Locations	
		Behavior	
		Transactions	
		Nature	
S3	Administrative level associated with data	Specific	
		Unspecific	
S4	Diversity of data providers	One provider	
		Several providers from same industry	
		Several providers from different industries	
S5	Facilitation	Self-facilitated	
		Intermediary with data-related functions	
		Intermediary with organizational functions	
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	
		Data shared as open data	
Data use and demand side			
U1	Target user group	Academic	
		Commercial	
		Governmental	
		Non-profit	
		Citizens	
U2	User selection	On agreement basis	
		On application basis	
		Open	
U3	Research or policy problem	Specified	
		Unspecified	
U4	Incentive to use data	Tangible	
		Intangible	
U5	Continuity of collaboration	On demand	
		Event-based	
		Continuous	

U6	Expected outcome of data collaborative	Policy intervention	Prediction and alerts
			Needs-based planning
			Capacity building
			Monitoring
		Data science	
		Data-driven innovation	
U7	Collaboration among data users	One user	
		Self-selected analysis by several users	
		Collaborative analysis by several users	
U8	Purpose of data use	Primary	
		Secondary	
		Tertiary	
		End use	

Table 2. Sample of cases of data collaboratives

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 3. Comparison of models of coordination at the organizational level: hierarchy, market, network, and bazaar (adapted from Demil and Lecocq, 2006)

	Hierarchy	**Market**	**Network**	**Bazaar**
Coordination mechanism governing exchange	Hierarchical authority	Price	Long-term relations	Common goods
Incentives intensity	Low	High	Intermediate	Low
Control intensity	High	Low	Intermediate	Low

Table 4. Components of coordination found in the data collaboratives taxonomy

Components of coordination (Malone and Crowston, 1990)	**Relevant dimensions of taxonomy (Susha et al., 2017)**
Actors	S4 Diversity of data providers
	S5 Facilitation (by intermediary)
	U1 Target user
	U2 User selection
	U7 Collaboration among data users
Activities	Data sharing
	Data use
Goals	U3 Research or policy problem
	U4 Incentive to use data
	U6 Expected outcome
	U8 Purpose of data use
Resources	S1 Type of data
	S2 Content of data
	S3 Administrative level of data
	S6 Degree of access to data

Table 5. Coordination problems and mechanisms found in data collaboratives

	Coordination problem	Relevant coordination mechanisms
1	Matching potential data providers and data users	Low coordination Coordination by negotiation Coordination by transfer of knowledge Coordination by third parties Coordination by process standardization
2	Maintaining control over the data and its unforeseen uses	Coordination by plan Coordination by supervision Coordination by standardization Coordination by transfer of knowledge
3	Matching a particular research/policy problem with the attributes of the data	Low coordination Coordination by negotiation Coordination by feedback
4	Ensuring the usability and usefulness of the data shared to the data user	Coordination by transfer of knowledge
5	Aligning incentives of data providers to share data with the goals of data users	Coordination by feedback Coordination by transfer of knowledge Coordination by third party