Towards a Governance Structure for the Data-Driven Prioritization of Humanitarian Aid

A data ecosystem approach

Master Thesis

By Elise Haak

October 2017

TU Delft

510

PERSONALISED HUMANITARIAN AID THROUGH BIG DATA

AN INITIATIVE OF THE NETHERLANDS RED CROSS

Towards a Governance Structure for the Data-Driven Prioritization of Humanitarian Aid
A data ecosystem approach

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

In **Complex Systems Engineering and Management**

Faculty of Technology, Policy and Management

By

Elise Haak

Student number: 4111672

To be defended in public on November 6th, 2017

Graduation committee:

Chairperson	: Prof. Dr. B.A. (Bartel) Van de Walle, Policy Analysis
First Supervisor	: Drs. J. (Jolien) Ubacht, ICT
Second Supervisor	: Dr. S. (Scott) Cunningham, Policy Analysis
External Supervisor 1	: M. (Marc) van den Homberg, 510 Global
External Supervisor 2	: S. (Stefania) Giodini, 510 Global

PREFACE

In front of you lies the result of my past months of researching the possibilities to enable the data-driven prioritization of humanitarian aid. It has been a great experience to apply the knowledge I gained during the CoSEM Master's program to a research problem with such a societal relevance. I learned a lot about the very interesting and dynamic humanitarian sector, and the impact of the trend of the increasing availability of data on this sector and its operations. Also, I had a great time working with the extremely motivated and inspiring 510 Global team.

I have been very lucky that while carrying out this Master's thesis, I was supported by supervisors who were very helpful and very much involved with my research process and progress.

I want to thank my graduation committee from the TU Delft: Jolien, Bartel and Scott. Jolien, thank you for supporting me from the beginning, for helping me to structure my ideas, reminding me to stay focused, and for always providing me with very extensive feedback. Bartel, thank you for sharing your humanitarian expertise with me, and for helping me to remain critical. And thank you, Scott, for your detailed comments and insights, and for showing me different perspectives on how to approach my research problem.

I would also like to thank the 510 Global team, for the guidance and support I received from them in developing this thesis. Thank you, Marc, for your time, dedication and all the great insights I received from you, both on a practical and on an academic level. Your expertise and experience very much helped me to make this happen. Stefania, thank you for thinking along at points where I needed assistance, and for always being flexible and helpful. Maarten, thank you for giving me the opportunity to write my thesis at 510 Global and for helping me specify my research proposal, and thanks to Jannis for patiently answering all my questions.

And lastly, thanks to the entire 510 Global team, and to everyone else who contributed in one or another way to this thesis.

Elise Haak
October 2017

SUMMARY

The incidence of natural disasters worldwide is increasing. As a result, there is a growing number of people in need, whereas a limited amount of resources to help these people is available. It is therefore important to effectively and efficiently prioritize the most vulnerable people prior to a disaster (preparedness phase), and the most affected people in the response phase of humanitarian action. The rapidly changing information environment that results in an increasing availability of data, offers the potential to make these prioritization processes both more effective and more efficient. A growing number of studies is dedicated to data-analytic models that can help in humanitarian decision-making processes when trying to respond to the consequences of a natural disaster. One of these models is the 'Community Risk Assessment and Prioritization toolbox' of 510 Global, the data team of the Netherlands Red Cross (NLRC). This is a data-driven solution that aims to better prepare vulnerable communities in the preparedness phase, and, once a disaster hits, to be able to provide timely and adequate humanitarian aid. The toolbox is under development for three different countries: Malawi, the Philippines and Nepal. If it can be taken to scale across the humanitarian system by being able to roll it out in many countries prone to natural disasters, it could contribute to significant time and cost savings in the provision of aid.

Currently it is a very challenging, time-consuming and resource-intensive process to collect all the data necessary for the toolbox to function well. Therefore, to realize this scale up, it must become clear how to facilitate, stimulate and coordinate data-sharing between humanitarian actors on a large scale. 510 Global needs a generic, non-country-specific governance structure to determine which approach to adopt in countries where they want to roll out the toolbox. This led to the following research question of this thesis:

WHAT DOES A GENERIC GOVERNANCE STRUCTURE THAT FACILITATES THE PROCESS OF SCALING UP THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX LOOK LIKE?

To answer this question, the theoretical outcomes of a systematic literature review were combined with the empirical findings that resulted from interviews with eight humanitarian data experts representing different countries or regions, with a wide knowledge on the inter-organizational data landscape in their country or region.

The systematic literature review focused on 'data ecosystems', which are networks of actors between whom data-sharing needs to be facilitated. To come to a generic governance structure for scaling up the Community Risk Assessment and Prioritization toolbox, a 'humanitarian data ecosystem' must be established. So far, there is no scientific literature that addresses this concept.

The outcome of the systematic literature review, which was validated with a group of open data researchers from the Delft University of Technology, was a framework of criteria to consider when developing a successful data ecosystem. Three different categories of criteria were distinguished: 1) Data supply, relating to the provision of data as open data, 2) Governance, being the framework of policies, processes and instruments to realize common goals in the interaction between entities (and facilitating data supply), and 3) User characteristics, including their needs and capacities.

Translating this theoretical framework to the process of scaling up the Community Risk Assessment and Prioritization toolbox in an evaluation session with 510 Global team members led to the conclusion that the humanitarian data ecosystem around the toolbox is still very

much in its infancy. As a result, some of the theoretical criteria only become relevant in a later stage of the humanitarian data ecosystem. Currently there is a strong focus on initial data collection. The governance of the humanitarian data ecosystem is very much underdeveloped. There is a lack of central governance that can stimulate the data ecosystem development, and can coordinate and align the different and fragmented individual humanitarian data initiatives. The empirical findings, which were structured based on the theoretical governance criteria found in literature, showed that to successfully govern the process of rolling out the Community Risk Assessment and Prioritization toolbox in a country and thereby contributing to its scale up, there are generically four different roles to be identified:

1. *The initiator/coordinator:* Leading agency that should initiate, coordinate and support the development of the model in a country. The most suitable organization to take up this role is a UN agency, due to their strong and structural position in a country, and their relatively advanced information management capacities.
2. *The local lead:* National body (government agency) responsible for managing and maintaining (e.g. including new features in and updating the data in) the toolbox, to ensure its sustainability and to create local buy-in.
3. *Data providers:* Organizations sharing data that can serve as an input for the toolbox. The identification of data providers is a time-consuming process, and data providers are not always willing to share their data, which is why it is important to make clear 'what is in it for them'.
4. *Toolbox users:* Organizations that use the toolbox and incorporate the output in their operations. There is a wide range of potential users for the toolbox, mostly humanitarian and government agencies.

The initiator/coordinator is responsible for the capacity-building of the other roles. For all roles, a shift in mentality and organizational culture is required to move towards more data-driven operations. To realize this and to incentivize parties to take up their role, it is very important for the initiator/coordinator to involve all relevant stakeholders from the start of the model development, to create ownership and have them realize the added value. 510 Global are advised to either try to directly approach a suitable initiator/coordinator using their current network, or adopt a stepwise approach, in which they start the toolbox development in a country with a less ideal initiator/coordinator and try to attract other parties as the development advances.

The outcomes of this study are twofold. In the first place, the study provided a generic governance structure on how to approach the process of scaling up the Community Risk Assessment and Prioritization toolbox in a country, to contribute to the shift towards more evidence-based humanitarian decision-making (societal contribution). In the second place, the study resulted in the creation of a theoretical framework of data ecosystem success criteria, which was theoretically validated by open data experts, and empirically validated by applying it to the humanitarian context of this study (scientific contribution).

Future research should focus on further validating the theoretical framework by applying it to other case studies. Other case studies can also provide insight in the evolvement of the maturity of other data ecosystems, which is something the immature humanitarian data ecosystem can learn from.

KEYWORDS: Humanitarian data, natural disasters, data-driven aid prioritization, data ecosystem, governance

CONTENTS

Chapter 1 – Introduction 1

1.1. Context description 1
1.1.1. Aid prioritization after a disaster 1
1.1.2. Increasing data availability in the humanitarian sector 1
1.1.3. The 'Community Risk Assessment and Prioritization toolbox' of 510 Global 2
1.2. Problem definition 4
1.2.1. Scaling up the toolbox 4
1.2.2. Uncertainty about data collection 4
1.3. The study 6
1.3.1. Research objective 6
1.3.2. Scope 6
1.3.3. Research questions and approach 11
1.3.4. Societal and scientific relevance 14

Chapter 2 – System Description 16

2.1. Introduction 16
2.2. A socio-technical system 16
2.2.1. Technical complexity 17
2.2.2. Actor complexity 24
2.5. Chapter conclusion 26

Chapter 3 – Theoretical Perspective: Data Ecosystems 28

3.1. Introduction 28
3.2. The ecosystems approach 29
3.2.1. What and why? 29
3.2.2. Different ecosystem applications 30
3.2.3. The context of the Community Risk Assessment and Prioritization toolbox as a data ecosystem 30
3.3. Systematic literature review methodology 31
3.4. Criteria for a successful data ecosystem 35
3.4.1. Category 1 - Data supply 37
3.4.2. Category 2 - Data Governance 41
3.4.3. Category 3 - User characteristics 45
3.4.4. Category A - External context 46
3.4.5. Visualization of findings 46
3.5. Systematic literature review evaluation: expert validation 47
3.5.1. Approach 47
3.5.2. Initial remarks of the researcher 48
3.5.3. Additional comments of participants 50
3.5.4. Conclusions of expert validation 50
3.6. Chapter conclusion: Final framework of criteria for a successful data ecosystem 51

Chapter 4 – The Humanitarian Data Ecosystem 53

4.1. Introduction 53
4.2. The data ecosystem around the Community Risk Assessment and Prioritization toolbox 54
4.2.1. An emerging data ecosystem 54
4.2.2. Data supply in the emerging data ecosystem 54
4.2.3. User characteristics in the emerging data ecosystem 55
4.2.4. Governance in the emerging data ecosystem 56
4.2.5. Set of framework elements to be operationalized 56

4.3. Operationalization of relevant ecosystem criteria 57
4.3.1. Conditions to be fulfilled and related metrics 58
4.3.2. Threshold values of conditions to be fulfilled 72
4.4. Chapter conclusion 74
Chapter 5 – A Generic Governance Approach 76
5.1. Introduction 76
5.2. Practical relevance of theoretical governance criteria 77
5.3. Set-up semi-structured interviews 79
5.3.1. Development of interview protocol 80
5.3.2. Selection of interviewees 80
5.3.3. Preparation and carrying out of interviews 83
5.3.4. Processing of interview outcomes 84
5.4. Empirical outcomes: a governance approach 84
5.4.1. 510 Global toolbox vs. INFORM Subnational 84
5.4.2. Genericity of empirical findings 84
5.4.3. Main empirical findings per governance element 84
5.4.4. Additional findings 92
5.5. Empirical validation of governance part of theoretical framework 94
5.5.1. Suggestions for framework adjustments on the governance side 95
5.5.2. Evaluation of the application of the governance part of the framework in the humanitarian data ecosystem 96
5.6. Chapter conclusion 98
Chapter 6 – Conclusion, Discussion and Recommendations 101
6.1. Main conclusion of research 101
6.2. Scientific contribution 104
6.3. Discussion 104
6.3.1. Discussing the data ecosystem approach and limitations 104
6.3.2. Discussing the selection and role of the interviewees 105
6.3.3. Discussing the value of data-driven aid prioritization 106
6.4. Recommendations 107
6.4.1. Recommendations for 510 Global 107
6.4.2. Recommendations for future research 110
6.5. Reflection on link with CoSEM program 111
References 112
Appendices 120
Appendix A – Literature search on the 'humanitarian data ecosystem' 121
Appendix B – Presentation slides expert validation session 124
Appendix C – Interview invitation e-mail 126
Appendix D – Interview protocol and questions 127
Appendix E – Storyline created based on empirical findings 139
Appendix F – Scientific Article 159

LIST OF FIGURES

FIGURE 1 – Municipality prioritization with the Community Risk Assessment and Prioritization toolbox (adapted from Van den Homberg, Van der Veen & Visser, 2016) 3
FIGURE 2 – Disaster Management Cycle (adapted from Stojic, 2013) 9
FIGURE 3 – Timeline of the emergency response phase (own figure). 11
FIGURE 4 – Research steps: methodologies and output (own figure) 14
FIGURE 5 – Example of pre-disaster indicators for Malawi, based on the three INFORM dimensions (adapted from Van den Homberg, Van der Veen & Visser, 2016) 18
FIGURE 6 – Example of Community Risk Assessment dashboard for Malawi. Outputs can be shown for different administrative levels and specified per indicator (510 Global, 2017A). ... 19
FIGURE 7 – Example of Priority Index dashboard for Typhoon Haiyan in the Philippines. The color indicates the number of houses damaged in that area (510 Global, 2017A). 19
FIGURE 8 – Visualization of differences in administrative levels for Malawi (510 Global, 2017A) 23
FIGURE 9 – Main stakeholders around Community Risk Assessment and Prioritization toolbox (own figure) 25
FIGURE 10 – Different operational levels of disaster response, as identified by Vitoriano, Montero & Ruan (2013). (Figure adapted from Van Lint, 2016) 26
FIGURE 11 – Concentric shell model of Backx (2003) 38
FIGURE 12 – Aspects of the governance of a data infrastructure (Kok & Van Loenen, 2005) 42
FIGURE 13 – Open Data Support Mechanisms for Public Bodies (Lee, 2014) 42
FIGURE 14 – Visualization of criteria for a successful data ecosystem found in the systematic literature review 47
FIGURE 15 – Final framework of criteria for a successful data ecosystem – findings from systematic literature review and expert validation combined 52
FIGURE 16 – Development of data ecosystem around the Community Risk Assessment & Prioritization toolbox 55
FIGURE 17 – Visualization of impact of research steps (own figure) 74
FIGURE 18 – INFORM Subnational dashboard for Colombia (left, INFORM, 2017) and Community Risk Assessment Dashboard for Malawi (right, 510 Global, 2016A): illustration of similarities 81
FIGURE 19 – Vicious circle 88
FIGURE 20 – Timeline of the involvement of and relationship between different roles 92
FIGURE 21 – Left: governance elements before empirical validation Right: governance elements after empirical validation 96
FIGURE 22 – Identified roles and the main empirical findings per role 99
FIGURE 23 – Empirically validated theoretical framework 100
FIGURE 24 – Visualization of desired transition from current situation to desired situation ... 107
FIGURE 25 – Timeline of the involvement of and relationship between different roles, adjusted for the recommendation of directly approaching a suitable initiator/coordinator via INFORM 109

LIST OF TABLES

TABLE 1 – List of articles reviewed (objectives and methodologies have been retrieved from the articles) 34
TABLE 2 – Criteria for a successful data ecosystem found in the systematic literature review 37
TABLE 3 – Overview of conditions to be used to assess the eligibility of a given country for the application of the Community Risk Assessment and Prioritization toolbox 59
TABLE 4 – Data maturity assessment tool to assess the eligibility of a given country for the application of the Community Risk Assessment and Prioritization toolbox 60
TABLE 5 – Different ways of acquiring disaster management data “on” and “with” communities (reprinted from Van den Homberg, Van der Veen & Visser, 2016) 64
TABLE 6 – Translation of theoretical governance criteria to practical context of research 79
TABLE 7 – List of interviewees and interview specifications 82
TABLE 8 – Main empirical findings per governance element 92
TABLE 9 – Proposed generic governance structure 103
TABLE 10 – Overview of suitable government agencies proposed by interviewees 143
TABLE 11 – Remarks on data literacy made by the interviewees 151

CHAPTER 1 – INTRODUCTION

1.1. CONTEXT DESCRIPTION

1.1.1. AID PRIORITIZATION AFTER A DISASTER

The humanitarian sector worldwide is under pressure: there is an increasing demand for humanitarian aid and emergency response services, as both the number and length of humanitarian crises grow (510 Global, 2016B; Monaghan & Lycett, 2013; Whipkey & Verity, 2015). One of the causes of this increase is the rising number of disasters triggered by natural hazards, due to environmental degradation and climate change. This results in a fast-growing number of people in need (Hilhorst, 2013).

Due to scarcity of budget and resources for disaster response, humanitarian actors cannot reach all these people. They must identify the geographic areas that are affected most by the disaster and, within those areas, identify the vulnerable families that are most in need of aid. This prioritization process is at the core of all humanitarian interventions. During the process, humanitarian actors quickly need information on the damage in the areas that were hit by the disaster, about for example the affected population, casualties, road blocks and damaged houses (Netherlands Red Cross, 2017). Coordination of these activities is crucial to the disaster response to ensure that humanitarian needs are addressed without duplication of efforts and to efficiently distribute resources (Van de Walle & Comes, 2016).

Currently this prioritization process in which damage assessment and identification of the most vulnerable people takes place is a time-consuming process that can take weeks to complete. Also, it is often unsafe and expensive. Assessment teams need to go into the affected area to interview people and assess damage to houses. Because these teams operate under pressure and severe stress combined with limited information sharing, there is a chance that their assessments are incomplete or unreliable, which affects subsequent accurate decision-making. Additionally, this field-based method is prone to subjectivity, as the selection of priority areas can be influenced by for example political affiliation and media coverage (Van den Homberg, Visser & Van der Veen, 2017; Van de Walle & Comes, 2014).

Obtaining reliable, objective information for decision-making in the initial emergency response phase in the aftermath of a natural disaster is therefore considered to be one of the main challenges of the humanitarian agencies and governments (Vitoriano, Montero & Ruan, 2013).

1.1.2. INCREASING DATA AVAILABILITY IN THE HUMANITARIAN SECTOR

This humanitarian challenge is severely influenced by the fast changing information environment. Paper-based processes get digitized increasingly and trends around crowdsourcing, social media, collaborative digital spaces and mobile services lead to a significant increase in availability of data. This also affects the humanitarian sector. Humanitarian donors for example, push the organizations they sponsor to open and share their data using the International Aid Transparency Initiative (IATI). In addition, major global

agreements that came into effect in 2015, like for example the Paris Climate Agreement, the Sustainable Development Goals, and the Sendai framework for Disaster Risk Reduction, all stimulate countries to collect data on different indicators that are also potentially valuable for humanitarian organizations (Van den Homberg, Visser & Van der Veen, 2017).

This increasing data availability enables humanitarian actors to engage more easily with communities and coordinate collaboration among different stakeholders, and leads to a shift from field-based to data-based prioritization of areas affected by a disaster. For natural hazards, it has become increasingly possible to provide rapid, data-driven impact assessments, such as the predicted damage to houses after a typhoon or a flood. Combining this data on the hazard with socio-economic data on the areas that are prone to the hazard, has a large potential for prioritization (Netherlands Red Cross, 2017).

As a result, a growing number of studies is dedicated to mathematical models and data-analytic techniques that can help in the humanitarian decision-making processes when responding to the consequences of a disaster (Ortuño et al., 2013). Data-driven priority indices have become popular for identifying communities most affected by disasters and numerous data and machine learning tools have been developed to aid prioritization (Benini, 2015). In the first days after a sudden onset disaster, organizations like the Red Cross and Red Crescent National Societies, governments or UN OCHA can use these indices to identify priority areas for humanitarian aid. It helps them to better understand the impact of a natural disaster and to mobilize humanitarian response faster and more efficiently, without having to wait for the field assessments (Van Lint, 2016).

1.1.3. The 'Community Risk Assessment and Prioritization toolbox' of 510 Global

One of the priority tools developed is the open source 'Community Risk Assessment and Prioritization toolbox' of 510 Global, the data team of the Netherlands Red Cross (NLRC). This team aims to drive smart use of (big) data for faster and more (cost) effective humanitarian aid (510 Global, 2016B).

Their Community Risk Assessment and Prioritization toolbox is an alternative data-driven solution that preemptively gathers and combines relevant data on risk components to provide a detailed risk assessment for areas and communities in a country, that helps to faster identify priority areas for humanitarian intervention related to natural disasters. They aim to retrieve data at the lowest possible granular level, up to community or even household level, as detailed data improves the value of the model predictions: more detailed model predictions result in the ability to target areas and people affected more specifically, which means that field assessment teams need less time for selection and identification of people affected, and hence relief can be provided faster.

The model consists of two components: the 'Community Risk Assessment tool' and the 'Priority Index'.

The Community Risk Assessment tool generates digital community profiles that supports identification of priority areas for disaster preparedness and risk reduction programs, by collating open data on a set of vulnerability, coping capacity and hazard exposure indicators based on the Index for Risk Management (INFORM) (Netherlands Red Cross, 2017). As stated by UN/ISDR & UN/OCHA (2008), "strengthened preparedness for hazard events is mainly concerned with two objectives: increasing capacity to predict, monitor and be prepared to reduce damage or address potential threats, and strengthening preparedness to respond in

an emergency and to assist those who have been adversely affected" (p.1). The application of this Community Risk Assessment tool is hence not dependent on or related to the specific occurrence of disasters, but rather aims to enhance the overall preparation for natural disasters and to reduce the risks.

The 'Priority Index' is a prediction model that predicts high priority areas after a disaster has struck based on a combination of the data from the Community Risk Assessment tool, and hazard and exposure data, like geographical data on the terrain, weather information on the imminent hazard (such as wind speeds and rainfall) and official historical damage data from past disasters. It uses machine learning and data mining algorithms and can identify priority areas within hours after the disaster has happened. For some disasters, like typhoons, the impact can already be predicted before the typhoon makes landfall, based on rainfall and wind speed predictions from local weather services. The predictions serve a timeframe of the first hours up to a week: when official data becomes available it is incorporated in the predictions, to be able to provide the best possible information on priority areas and damage figures. The accuracy of the prediction is verified by comparing it with official impact data from historical hazards of the same nature.

This described prioritization process with the Community Risk Assessment and Prioritization tool is visualized in *Figure 1* below. The different phases of disaster management will be further elaborated on in *Section 1.3.2 – Scope.*

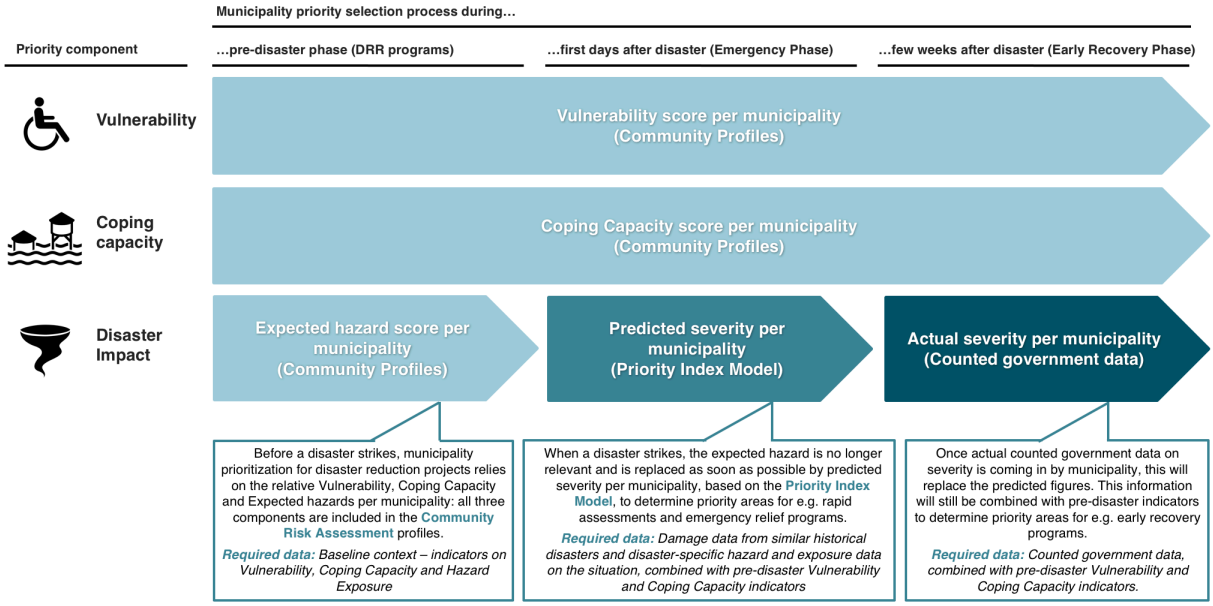


FIGURE 1 – Municipality prioritization with the Community Risk Assessment and Prioritization toolbox (adapted from Van den Homberg, Van der Veen & Visser, 2016)

To summarize, the Priority Index component of the 'Community Risk Assessment and Prioritization toolbox' adds disaster-specific information to the more general pre-disaster community profiling (the Community Risk Assessment component). Thus, it is possible for the Community Risk Assessment component to function without the Priority Index component, but not vice versa, as the Community Risk Assessment component serves as an input for the Priority Index.

The toolbox has so far been tested for typhoons Haima (2016) and Nina (2016) in the Philippines. Applied research is going on for earthquakes in Nepal and floods in Malawi. There

is growing evidence that the toolbox can be used for different natural disasters in different countries.

The intended outcome of the Community Risk Assessment and Prioritization toolbox is for humanitarian decision-makers to use credible and accurate data to identify and prioritize the communities most at-risk during the preparedness phase prior to the disaster, and to prioritize the areas and people that are most affected by the disaster in the response phase. The direct impact is that vulnerable communities at risk are better prepared before a disaster arrives, and that, once a disaster hits, they receive timely and adequate humanitarian aid to help meet their primary needs (Netherlands Red Cross, 2017). This can eventually contribute to an increased efficiency and effectiveness of humanitarian response.

1.2. PROBLEM DEFINITION

1.2.1. SCALING UP THE TOOLBOX

If the Community Risk Assessment and Prioritization toolbox can be taken to scale across the humanitarian system by making it applicable for many countries, it has the potential to lead to improved, more evidence-based humanitarian decision-making and aid provision. Additionally, cost reductions can be realized through an increased efficiency of data collection and data-sharing processes, by focusing data efforts on a limited set of well-defined risk indicators; only the data necessary to generate accurate results with the toolbox (Netherlands Red Cross, 2017). A large-scale realization of data-driven prioritization could hence contribute to faster and more (cost) effective humanitarian aid. Scaling up the Community Risk Assessment and Prioritization toolbox is therefore high on the agenda of 510 Global.

As mentioned, the Community Risk Assessment component can be applied independent of the Prioritization Index, and to a wider range of countries and areas. Where application of the Priority Index is mainly relevant for countries that are prone to natural disasters and have been struck repeatedly by similar disasters (a requirement for the model to be able to run), community profiling with the Community Risk Assessment component can also be valuable in other areas where humanitarian aid is being provided. If humanitarian aid organizations for example aim to reduce extreme poverty, the Community Risk Assessment model can provide insight in the communities for which this non-disaster related aid would be most needed.

Yet, in the first place 510 Global will focus on the implementation and scaling of the toolbox as a whole - so the two components combined. This means that the geographic scope for scaling up the toolbox will be determined by the more limited applicability of the Prioritization Index.

The expectation is that if sufficient baseline and impact data from, for example, historical disasters is available, then machine-learning methodologies can be successfully applied to different contexts, which means that the Community Risk Assessment and Prioritization toolbox could be scaled up to countries and areas prone to natural disasters other than the ones that have been used to test the model so far.

1.2.2. UNCERTAINTY ABOUT DATA COLLECTION

However, there are several data-related challenges to this desired large-scale realization of data-driven humanitarian aid prioritization by means of a wide deployment of the Community Risk Assessment and Prioritization toolbox. For the countries in which the toolbox has already been applied (the Philippines, Nepal and Malawi) an overview exists of how and where to collect the data required for the toolbox to function well, but a more generic governance

structure that can also be used to locate and collect data in other countries, and assess whether they are eligible for the application of the toolbox, is lacking.

It should be stressed here, that locating and collecting humanitarian data often is a challenging process. The data is fragmented; distributed among many different organizations, varying between governmental organizations (e.g. municipalities, national statistic authorities, department of health), international organizations (e.g. United Nations related organizations), joint research centers, open data initiatives, national Red Cross societies and other organizations. There is a lot of diversity in the quality and coverage of the data, and in how the data of these organizations is stored and made available (if at all). There is a humanitarian open data platform, called 'Humanitarian Data Exchange' (HDX), there are national open data platforms of governments, the World Food Program has a more digital data collection and within the Red Cross mobile data collection is on the rise (Payne, Florance & Shain, 2012). But there are also many stakeholders, like local authorities, that only store their data locally, sometimes even paper-based. For example, a field research in Malawi showed that "especially at the lower administrative levels data could be available but not at all easily accessible. The data was not yet uploaded to geospatial sharing platforms, but kept on individual computers. Reasons for this could vary from low priority, lack of time and infrastructure (poor internet for example) to do so" (Van den Homberg, Visser & Van der Veen, 2017, p.11).

The data hence regularly must be collected manually. Whether this succeeds depends on the data owner's willingness to share. Additionally, for geospatial data it is not uncommon to be incomplete (or non-existent), which also leads to the need for field trips to manually map unmapped areas, or the development of crowd-sourcing initiatives. Thus, locating and collecting data in the humanitarian sector is not merely a matter of creating the right IT-architecture.

Due to this complexity, now it is unclear whether sufficient data is available and accessible for a large-scale application of the Community Risk Assessment and Prioritization toolbox to be realized. There is no overview of whether, where and how the required data can be collected. Insight into the minimum set of data that is necessary for the toolbox to generate accurate and realistic priority predictions and insight in the country-specificity or universality of this data set across countries would help to tackle this challenge. Also, more knowledge is required on whether it is possible to collect this minimum data on a large scale for different countries, how these data should be collected (is the data open or closed?) and the related data collection methodologies that are feasible and result in the highest data quality. What are the key gaps between information needs and data availability? How can the process of data collection be arranged effectively and efficiently? What are the guidelines, policies, processes and means required to collect the minimum dataset in a country? To answer these questions, issues like for example data ownership, incentives to share data and the realization of arrangements with data providers should be assessed.

Thus, dealing with the requirements, availability and collection of data is considered a complex issue when assessing the possibilities to scale up the Community Risk Assessment and Prioritization toolbox to other countries. It is very challenging to identify and collect all the data necessary to create priority predictions with the toolbox that are up-to-date, accurate, valid and complete. 510 Global is in need of a generic approach to adopt when rolling out the toolbox in a new country.

1.3. THE STUDY

This research will address the 'knowledge gaps' as described in the problem definition, with the purpose to contribute to the improvement of the efficiency and effectiveness of humanitarian aid by facilitating large-scale data-driven aid prioritization. The research is carried out to support 510 Global, the data team of the Netherlands Red Cross, who are the developers of the Community Risk Assessment and Prioritization toolbox.
This section describes the research objectives and the resulting research questions.

1.3.1. RESEARCH OBJECTIVE

Based on the problem description, the objective of this study is determined as follows:

To enable the data-driven prioritization of humanitarian aid, by developing a generic structure for the governance needed to scale up the Community Risk Assessment and Prioritization toolbox.

510 Global should be able to use this governance structure to determine which approach to adopt in countries prone to natural hazards for which they want to develop the Community Risk Assessment and Prioritization toolbox. In this way, the governance structure can help in the process of scaling up the toolbox, and rolling it out in a new country.
By achieving this objective, this research can contribute to the realization of the vision of 510 Global: "Smart use of (big) data will help towards faster and more (cost) effective humanitarian aid" (510.Global, 2017B).

1.3.2. SCOPE

As the central topic of this study is the Community Risk Assessment and Prioritization toolbox, the scope of the research is determined by the possibilities for the application of this toolbox. In this section, the following boundaries are described:

- Scope within the Red Cross[1]: 510 Global develops the toolbox, but what does the organization around them look like?
- Scope related to (natural) disaster type: for which disasters can the toolbox be applied?
- Country scope: where can the toolbox be applied?
- Scope within the disaster timeline: when can the toolbox be applied?
- Scope within the toolbox: Community Risk Assessment vs. Priority Index

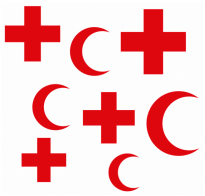
Scope: Within the Red Cross	This study will be carried out for 510 Global, the data team of the Netherlands Red Cross. To be able to place this in the worldwide Red Cross context, a short overview of the organizational structure of the Red Cross will be provided.

[1] Descriptions of the different Red Cross bodies are based on their mission statements, as retrieved from their websites (which have been referred to).

The “International Red Cross and Red Crescent Movement” is a global humanitarian movement of 80 million people that helps people facing disaster, conflict and health and social problems (ICRC, 2017I). The movement consists of different organizations:

The “International Committee of the Red Cross” (ICRC) works neutrally and impartially to ensure humanitarian protection and assistance for victims of armed conflicts and other situations of violence. They also respond to disasters in conflict zones, as the effects of a disaster are intensified if a country is already at war. Examples of ICRC activities are providing water, habitat and health care, protecting civilians, restoring family links and helping detainees. The ICRC works closely with the National Red Cross and Red Crescent Societies (ICRC, 2017II).

The “International Federation of Red Cross and Red Crescent Societies” (IFRC) coordinates and directs international assistance following natural and man-made disasters in non-conflict situations. Its mission is to improve the lives of vulnerable people. Its relief operations (disaster response) are combined with disaster preparedness and capacity building programs, with health and care activities and the promotion of humanitarian values. The IFRC works with National Societies in responding to catastrophes around the world (IFRC, 2017I).



There are 190 National Red Cross and Red Crescent Societies around the world. Each National Society is made up of volunteers and staff, who provide a wide variety of services. The National Societies serve as independent national relief organizations. Their activities include supporting public authorities and mitigating human suffering through prevention, disaster relief, education and health/social welfare problems. They also assist victims of armed conflicts and natural disasters, supporting the ICRC and the IFRC (ICRC, 2013).

510 Global is the data team initiated by the Netherlands Red Cross, the National Red Cross Society of the Netherlands.

Scope: Disaster type

As already made clear with the organizational description, different types of disasters can be distinguished. Broadly, a distinction can be made between *complex disasters* and *natural disasters. 'Complex'* in this context refers to disasters in conflict situations. As explained by Prof. Van de Walle in an interview with the TU Delft Global Initiative (2017), "though natural disasters are complex enough in their own right, there is a clear common enemy, e.g. the earthquake, typhoon or flood that is causing the trouble. In complex disasters, other issues play a part too, such as strategic, military and security concerns". Conflicts or outbreaks of war can cause large-scale medical problems like lack of water, epidemics, mass movements of refugees and displaced persons, and food shortage and hunger. As mentioned, within the Red Cross Red Crescent Movement, the ICRC is the lead agency for response to war and conflict situations.

The IFRC (2017II) distinguishes between the following natural disasters, which they define as "naturally occurring physical phenomena caused either by rapid or slow onset events":

- *Geophysical (earthquakes, landslides, tsunamis and volcanic activity), hydrological (avalanches and floods)*
- *Climatological (extreme temperatures, drought and wildfires)*
- *Meteorological (cyclones/typhoons and storms/wave surges)*
- *Biological (disease epidemics and insect/animal plagues)*

Preparing for and responding to natural disasters in non-conflict situations falls under the responsibility of the IFRC. The Community Risk Assessment & Prioritization toolbox of 510 Global currently focuses on identifying priority areas for humanitarian aid after natural disasters in non-conflict situations. Data collection in non-conflict situations is already a challenging process, and is expected to be even more complex (if possible at all) in conflict areas.

Scope: Natural disaster type

Within natural disasters, *rapid*-onset disasters and *slow*-onset disasters can be distinguished. Rapid-onset disasters are triggered by hazards that arise suddenly, or whose occurrence cannot be predicted far in advance. Examples of rapid-onset disasters are typhoons, floods, earthquakes, volcanic eruptions and wildfires (Twigg, 2004). Their warning time ranges from seconds to minutes in the case of landslides and earthquakes, to several days for most floods and storms. For those disasters, there is a clear moment at which it becomes clear that the Community Risk Assessment & Prioritization toolbox needs to be developed. This is not the case for slow-onset disasters. For droughts, the most discussed slow-onset disaster, it can take months or even years for its results to become disastrous (Twigg, 2004). It is therefore less clear-cut to determine when to develop the toolbox for these kinds of disasters.

The Community Risk Assessment & Prioritization toolbox is currently under development for typhoons, floods and earthquakes, the three most occurring natural disasters between 1995-2015 according to a report of the Centre for Research on the Epidemiology of Disasters (CRED) & The United Nations Office for Disaster Risk Reduction (UNISDR) (2015). These disasters are all usually categorized as rapid-onset disasters. Nevertheless, in the future it should also be possible to generate predictions for other rapid-onset disasters, and even for slow-onset disasters. This research however, only considers the models currently under development and is thus limited to typhoons, floods and earthquakes.

Scope: Countries

The Community Risk Assessment and Prioritization toolbox has in theory the potential to be scaled up to (or developed for) any country that is prone to natural disasters for which historical damage data is available, as this data is necessary to be able to train the model. Other requirements/boundaries will be determined within this research.

Scope: Disaster timeline

The management of natural disasters, necessary to decrease the impact of hazards and the possibility of disaster, consists of different phases. *Figure 2* below visualizes the key tasks within disaster management.

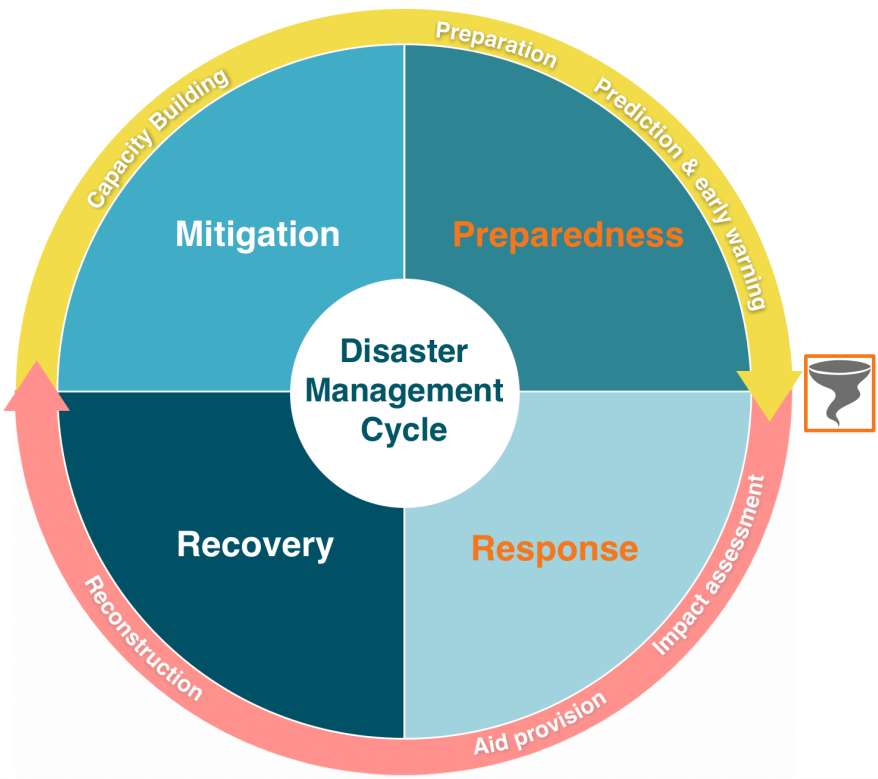


FIGURE 2 – Disaster Management Cycle (adapted from Stojic, 2013)

The United Nations Office for Disaster Risk Reduction (UNISDR) defines these phases as follows (UNISDR, 2017):

Preparedness
The knowledge and capacities developed by governments, professional response and recovery organizations, communities and individuals to effectively anticipate, respond to, and recover from, the impacts of likely, imminent or current hazard events or conditions.

Response
The provision of emergency services and public assistance during or immediately after a disaster in order to save lives, reduce health impacts, ensure public safety and meet the basic subsistence needs of the people affected.

Recovery
The restoration, and improvement where appropriate, of facilities, livelihoods and living conditions of disaster-affected communities, including efforts to reduce disaster risk factors.

Mitigation
The lessening or limitation of the adverse impacts of hazards and related disasters.

The Red Cross is involved in every step of the Disaster Management Cycle. 510 Global, however, focuses its activities only on the right half of the cycle: disaster preparedness and disaster response.
The Community Risk Assessment and Prioritization toolbox mainly aims to make the response phase more efficient, but also plays a role in the preparation phase: the data for the Community Risk Assessment needs to be collected as part of the preparation phase, to be able to apply the toolbox rapidly (start generating predictions on priority areas) immediately after a disaster has struck. This will be further elaborated on in *Section 2.3.1 – Rapid execution.*
Within the response phase, as described earlier, damage and needs need to be assessed and, due to limited resources being available, humanitarian aid needs to be prioritized. Usually, approximately one week after a natural disaster initial field damage and needs assessments come in. Before this time, the predictions of the Community Risk Assessment and Prioritization toolbox can be used to geographically guide field assessment teams (in terms of where they need to go) and to prioritize the initial aid. When more field assessments are coming in, the toolbox dashboard can be updated regularly, showing the damage counts from the field for municipalities of which field reports have already been done, and predictions for the other municipalities. It generally takes two to four weeks of field assessments before a complete overview of the damaged area has been generated. After this, the toolbox predictions can be completely replaced with official damage counts and hence the toolbox stops being of use. However, also after this time, to validate the numbers it could still be useful to compare the predictions with the official damage counts (UNOCHA, 2013). This timeline is visualized in *Figure 3* below.

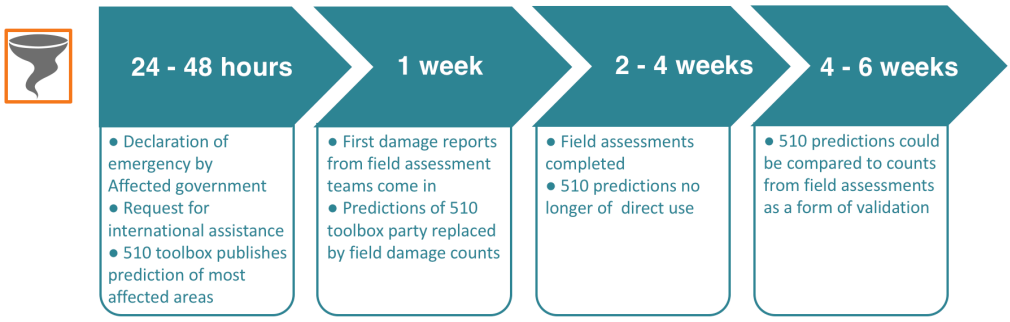


FIGURE 3 – Timeline of the emergency response phase (own figure).

Scope: Within the toolbox

When carrying out this study, is important to realize that the Community Risk Assessment and Prioritization toolbox consists of two parts, as described in the introduction. Both the Community Risk Assessment part and the Prioritization Index are based on different indicators, and hence they might require a different approach when assessing the possibilities for scaling-up. The intent however, is to come to an integrated governance structure that aims at the toolbox as a whole.

1.3.3. RESEARCH QUESTIONS AND APPROACH

Given the stated research objective, the main research question that needs to be answered in this study is the following:

WHAT DOES A GENERIC GOVERNANCE STRUCTURE THAT FACILITATES THE PROCESS OF SCALING UP THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX LOOK LIKE?

This 'generic governance structure' will be a generic (i.e. non-country-specific) approach to be adopted by 510 Global to govern the data-related processes required to roll out the toolbox in a new country.
The following set of sub-questions has been identified to provide additional steering in the research. These sub-questions structure and indicate the issues to be addressed crucial to answering this main question.

1. ***WHAT CHARACTERIZES THE SYSTEM AROUND THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX?***
    a. *What is the data required for the Community Risk Assessment and Prioritization toolbox to generate accurate predictions?*
    b. *Which parties are involved in the application of the Community Risk Assessment and Prioritization toolbox?*

This first sub-question is a descriptive question that will provide insight in the functioning of the Community Risk Assessment and Prioritization toolbox and its context; an analysis of the (complexity of the) system based on a desk research (*Chapter 2*).
A perspective that can be used to deal with complexity and understand interrelationships between actors and entities in complex systems, is the 'ecosystems' approach. Therefore, in *sub-question 2* a systematic literature review around data ecosystems will be performed. By providing an overview of theoretical background in this field, the framework resulting from this

study can serve as a point of departure for the creation of the generic governance structure for scaling up the Community Risk Assessment and Prioritization toolbox. The questions corresponding to this research step are:

2. ***WITHIN THE FIELD OF DATA ECOSYSTEMS, WHICH FRAMEWORK CAN BE USED AS A STARTING POINT FOR THE CREATION OF A GENERIC GOVERNANCE STRUCTURE FOR THE TOOLBOX?***
    a. *What is a data ecosystem?*
    b. *What are the criteria for a successful data ecosystem, according to literature?*

The output of this sub-question will be a framework of important elements (or 'success criteria') to consider when establishing, enabling or improving a data ecosystem (*Chapter 3*).
The next research step will be to determine how this framework can be used in the process of scaling up the Community Risk Assessment and Prioritization toolbox and finding a governance structure to guide this process: the theoretical framework needs to be translated into a practical, applicable tool. The following questions will be answered:

3. ***HOW CAN THIS DATA ECOSYSTEMS FRAMEWORK BE APPLIED TO THE CONTEXT OF SCALING UP THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX?***
    a. *Which criteria for a successful data ecosystem as derived from literature are relevant in the context of the Community Risk Assessment and Prioritization toolbox?*
    b. *How can these criteria be made operational, so they can become of use in the process of scaling up the Community Risk Assessment and Prioritization toolbox?*
    c. *How should these operationalized criteria be arranged in a country, for the country to be eligible for application of the Community Risk Assessment and Prioritization toolbox?*

The outcome of this sub-question will be a list of measurable conditions per success criteria relevant in the context of scaling up the Community Risk Assessment and Prioritization toolbox (*Chapter 4*). The conditions and corresponding metrics will be derived from the functioning of the toolbox in consultation with the head developer of the toolbox, who has experience with developing the model for the Philippines, Malawi and Nepal. It is expected that this will provide the required operational insights.
Assessing the state of these conditions in a country should provide a scan of the situation in a country in terms of its data maturity and hence eligibility for the application of the Community Risk Assessment and Prioritization toolbox.
But, only knowing whether a country is eligible for application of the Community Risk Assessment and Prioritization toolbox, will not lead to the desired goal of scaling up the application of the toolbox. 510 Global needs to know how to increase the amount of countries eligible for the toolbox application to be able to scale it up. They need to know which governance approach to adopt to improve a country's score on the operationalized criteria (the measurable conditions) as identified in *sub-question 3*; which policies, strategies or practices can improve the current conditions in a country. The following sub-question will therefore be answered in *Chapter 5*:

4. ***How should the governance of the data ecosystem around the Community Risk Assessment and Prioritization toolbox to facilitate the process of scaling up the toolbox be approached?***
    a. *How can the governance criteria for a successful data ecosystem as derived from literature be translated to the context of the Community Risk Assessment and Prioritization toolbox?*
    b. *How can these governance criteria be arranged in such a (practical) way that they create an optimal context for the development of the Community Risk Assessment and Prioritization toolbox?*

Sub-question 4a will be answered by translating the theoretical governance concepts derived from literature in such a way that they become of practical relevance in the context of scaling up the Community Risk Assessment and Prioritization toolbox. The answer to this sub-question will form the basis for answering *sub-question 4b*. Namely, to answer *sub-question 4b*, semi-structured interviews with humanitarian data experts that represent five different regions will be held. The findings from these interviews will give an idea of how to meet the criteria for successful data ecosystem governance, and together lead to the definition of a generic governance structure to be adopted when rolling out the Community Risk Assessment and Prioritization toolbox in a new country, in the process of scaling it up.
This research step hence provides practical insights in how to arrange the governance of the humanitarian data ecosystem, but it also serves as an empirical validation of the theoretical framework developed: applying the governance part of the theoretical framework in practice and testing it on the humanitarian case of scaling up the Community Risk Assessment and Prioritization toolbox, provides the opportunity to evaluate it on its suitability for application in practice, and, if needed, further improve it. The outcome of this research question is therefore twofold:

- Practical insights in the arrangement of the governance around scaling up the Community Risk Assessment and Prioritization toolbox through expert interviews.
- Theoretical insights in the validation of the theoretical framework by applying it in practice.

Figure 4 provides an overview of the different research steps, methodologies and output corresponding to the sub-questions.

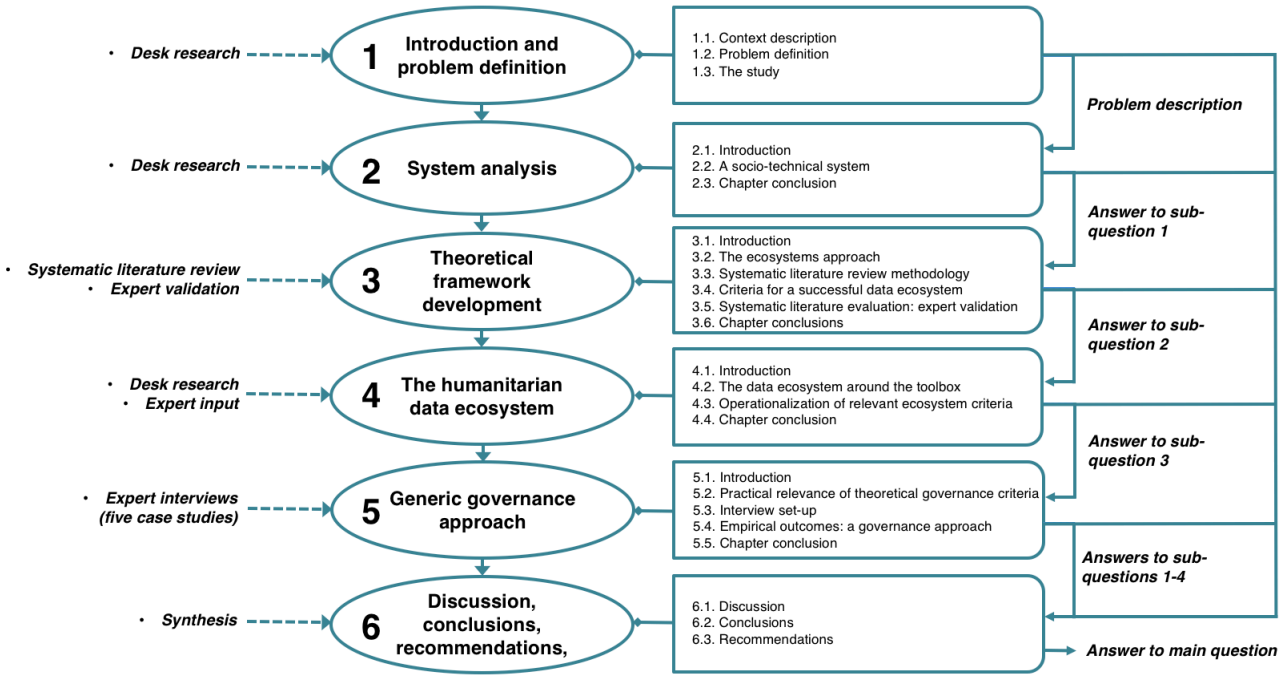


FIGURE 4 – Research steps: methodologies and output (own figure)

1.3.4. SOCIETAL AND SCIENTIFIC RELEVANCE

Societal relevance: The intended outcome of the Community Risk Assessment and Prioritization toolbox is for humanitarian decision-makers to use credible and accurate data to identify and prioritize both the most at-risk communities during the preparedness phase, as well as the most hit areas during the response phase. This would lead to vulnerable communities being better prepared before a disaster arrives, and have them receive timely and adequate humanitarian aid to help meet their primary needs once a disaster strikes. This research will help 510 Global to move towards the realization of this intended outcome.

Scientific relevance: According to Mans, Berens & Shimshon (2015), to ensure that projects around data-driven innovations in humanitarian response "can scale and be sustained in the long run, the humanitarian data ecosystem needs to be better understood" (p.6). This claim applies to the context of this study: The Community Risk Assessment and Prioritization toolbox is a data-driven innovation in humanitarian response, and this research aims to contribute to its scale-up. Additionally, Berens et al. (2016) and Raymond et al. (2016), argue that the data ecosystem perspective can help to address the issue of a lack of centralized governance in the humanitarian sector.

Yet, no scientific studies and very few other studies seem to have addressed the humanitarian data ecosystem and the possibilities for its successful establishment and governance. When searching for 'humanitarian data ecosystem' or 'humanitarian' and 'data ecosystem' in the academic databases Scopus, Web of Science and Google Scholar, this led to zero hits. On Google Search, the term resulted in 50 hits. Within those 50 hits, only six different documents ('grey literature') in which the term popped up were found; the remaining hits were redundant. The six documents that included the term, among which the referred to Berens et al. (2016), Mans, Berens & Shimshon (2015) and Raymond et al. (2016), are briefly described in *Appendix A.* None of them explicitly focuses on the success criteria when establishing and governing a humanitarian data ecosystem. They use the term and acknowledge the need for

the development of central governance processes in the humanitarian sector, but do not further research it or use it in a context in line with this study (see *Appendix A*). Berens et al. (2016) and Raymond et al. (2016) for example use the term to discuss issues around the responsible and ethical use of data.

It could therefore be stated that by taking a data ecosystem approach to analyze the governance around scaling up the Community Risk Assessment and Prioritization toolbox, thereby focusing on the 'humanitarian data ecosystem', this research helps to fill a gap in existing literature on (governance in) the humanitarian data ecosystem.

Furthermore, due to the lack of literature on the humanitarian data ecosystem, a systematic literature review around the criteria for the successful development of a data ecosystem in general is carried out, that serves as a basis for this study. This systematic literature review, which uses the study of Welle Donker & Van Loenen (2017) as a starting point, results in the creation of a framework of data ecosystem success criteria. This framework is both validated and case-tested in practice by applying it to the humanitarian context of the Community Risk Assessment and Prioritization toolbox. The generation and validation of this theoretical framework by means of a combination of a systematic literature review, expert validation and case study-testing can therefore be considered another scientific contribution of this study.

CHAPTER 2 – SYSTEM DESCRIPTION

2.1. INTRODUCTION

This chapter will provide a description of the system around the Community Risk Assessment and Prioritization toolbox. The following sub-question will be answered:

SQ1: WHAT CHARACTERIZES THE SYSTEM AROUND THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX?

a. *What is the data required for the Community Risk Assessment and Prioritization toolbox to generate accurate predictions?*
b. *Which parties are involved in the application of the Community Risk Assessment and Prioritization toolbox?*

This chapter is structured as follows. In paragraph 2.2, the term 'socio-technical systems' will be introduced as a perspective to describe the complex system around the Community Risk Assessment. Following this explanation, the system will be approached by its technical complexity in paragraph 2.3, and its actor complexity in paragraph 2.4. In paragraph 2.5, the findings are summarized.

2.2. A SOCIO-TECHNICAL SYSTEM

The multifaceted problem description (*Section 1.2 – Problem definition*) shows that scaling up the Community Risk Assessment and Prioritization toolbox is a complex and dynamic system to deal with. The system should be considered from multiple perspectives. It is not only important to understand the technical possibilities, but also the social and institutional dynamics of the environment the toolbox is placed in. It is unlikely to come to a sustainable solution if the actors involved in the application of the toolbox, like data providers and data users, are treated as an isolated entity. Instead, the interaction between the actors and the toolbox and their intertwinement need to be understood.
These characteristics are typical for so-called 'socio-technical systems'. The socio-technical systems approach helps to better understand the mechanisms and features of complex phenomena (Wang, Gu & Zhou, 2014). According to De Bruijn & Herder (2009), socio-technical systems are "systems that involve both complex physical-technical systems and networks of interdependent actors" (p.981). Both the system and the actor perspective contribute to the overall behavior of the system as a whole, and both are important to get to understand the complexity of the problems, and to be able to design relevant solutions (de Bruijn & Herder, 2009). This chapter will therefore describe both the technical and actor complexity around the Community Risk Assessment and Prioritization toolbox.

2.2.1. TECHNICAL COMPLEXITY

The technical complexity around the development of the Community Risk Assessment and Prioritization toolbox stems from a combination of different main factors:

- *Different types of data:* There is a variety of data that needs to be collected and subsequently combined in the model to generate predictions. The specific content of this data must be determined per country, due to the differences in context, practical needs and data sources available.
- *Timely collection of data:*
    - Part of the data ought to be collected in the 'preparedness phase' prior to a disaster: it must be ensured that countries are 'data prepared'.
    - Part of the data can only be collected during the 'response phase' under severe timing constraints in the complex and dynamic post-disaster environment, due to the need for rapid execution of the model.
- *Data collection at low administrative levels:* All data should be collected at the lowest possible administrative level, to be able to generate detailed predictions. This is challenging as in those countries for which the toolbox development is considered relevant, at lower administrative levels data is often not stored digitally, nor in a proper format, or it is not available at all.

In this section, these different factors will be elaborated on.

Different types of data

As described in *Section 1.1.3,* to predict the geographical spread of damage and thus priority areas for humanitarian intervention to provide decision-support for emergency response in an information poor situation, the Community Risk Assessment and Prioritization toolbox combines different types of data:

Pre-disaster secondary data on affected areas ('preparedness phase')

Input for the 'Community Risk Assessment' model:

- Indicators that together determine the 'Risk Score' of a country on three dimensions: Vulnerability, Coping Capacity and Hazard Exposure (Conform INFORM risk index)
- Dimensions the same for each country, but specific indicators per dimension are determined per country based on availability and determined in a consultative process with key stakeholders and data providers, (example for Malawi in *Figure 5* below)
- Output visualized in the Community Risk Assessment dashboard (example in *Figure 6* below)

Input for the 'Priority Index' model to be collected in the preparedness phase:

- Relevant geographic features of an area (e.g. slope, ruggedness)

- Hazard data on similar historical disasters, needed to train the model

Disaster-specific hazard and exposure data ('response phase')

- Weather and hazard related data
- To be collected just before and at the start of a disaster
- For typhoons, e.g.: wind speed, typhoon path, distance to typhoon
- For floods, e.g.: rainfall, flood maps / areas flooded, expected flood duration
- Data combined with indicators from the Community Risk Assessment, analyzed in 'Priority Index' model and visualized in the corresponding dashboard (example in *Figure 7* below)

Damage data ('response phase')

- Once actual counted government data on severity is coming in per municipality, these numbers can replace the model predictions (for those municipalities for which data is available)
- E.g. number of houses/buildings damaged, number of casualties
- If these data are reliable and become available quickly, they improve the value of the model output

Combining these data in a 'random forest regression', estimations of the spatial distribution of post-disaster human priorities are produced. The empirical model quantifies the relationship between on the one hand the explaining pre-disaster and disaster-specific data, and on the other hand aid neediness, the output. The relationship between those factors is derived from the analyses of pre- and post-event data of past events, hence the term 'empirical' (Bulte, 2017).
The (un)availability of data in a country steers the model-decisions.

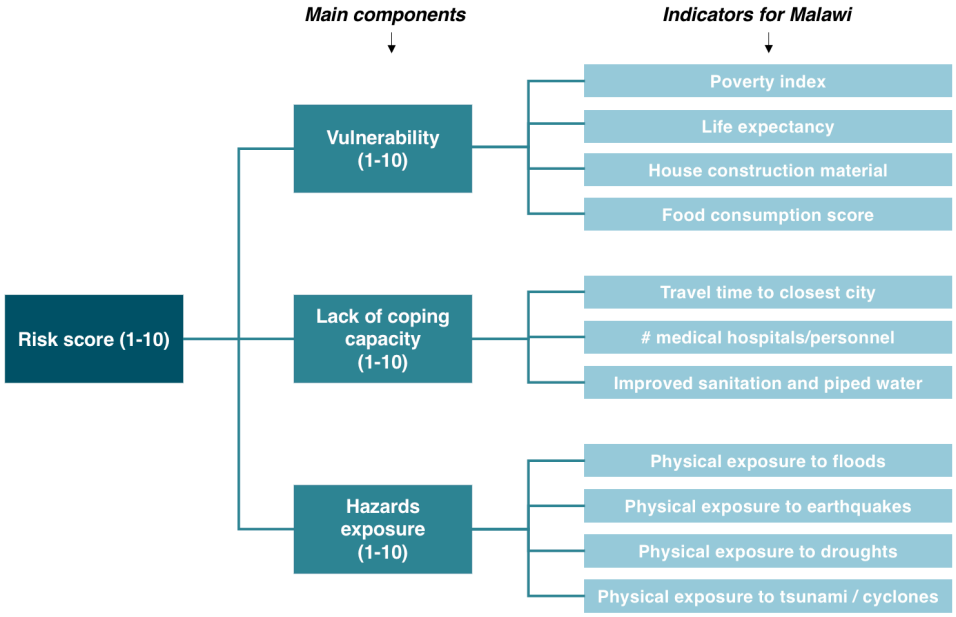


FIGURE 5 – Example of pre-disaster indicators for Malawi, based on the three INFORM dimensions (adapted from Van den Homberg, Van der Veen & Visser, 2016)

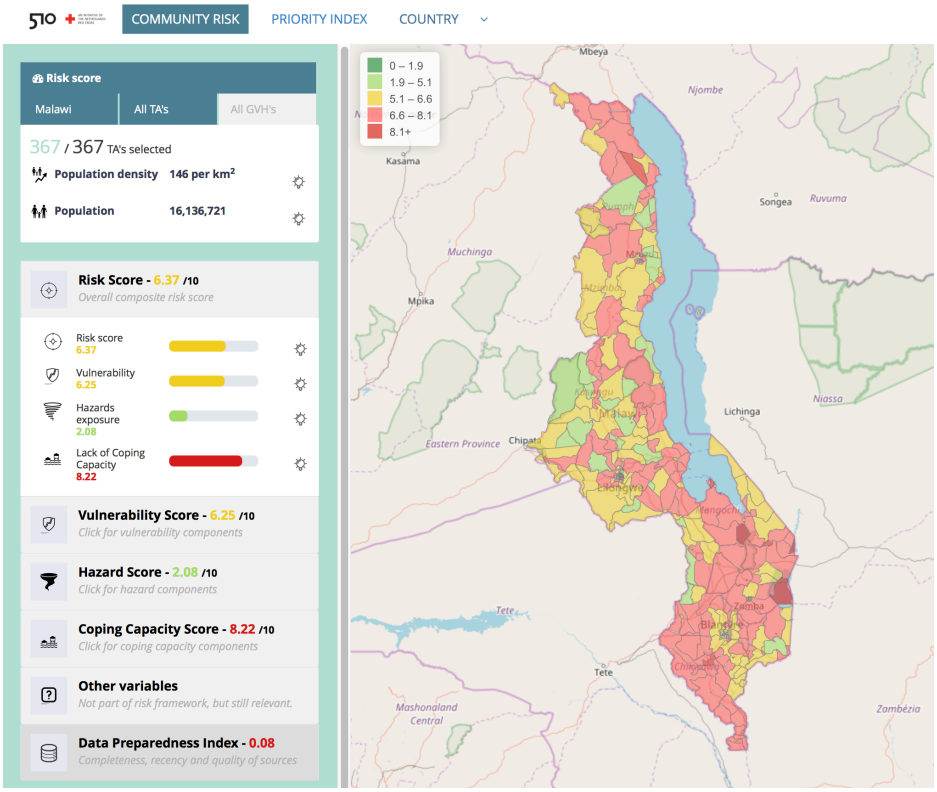


FIGURE 6 – Example of Community Risk Assessment dashboard for Malawi. Outputs can be shown for different administrative levels and specified per indicator (510 Global, 2017A).[2]

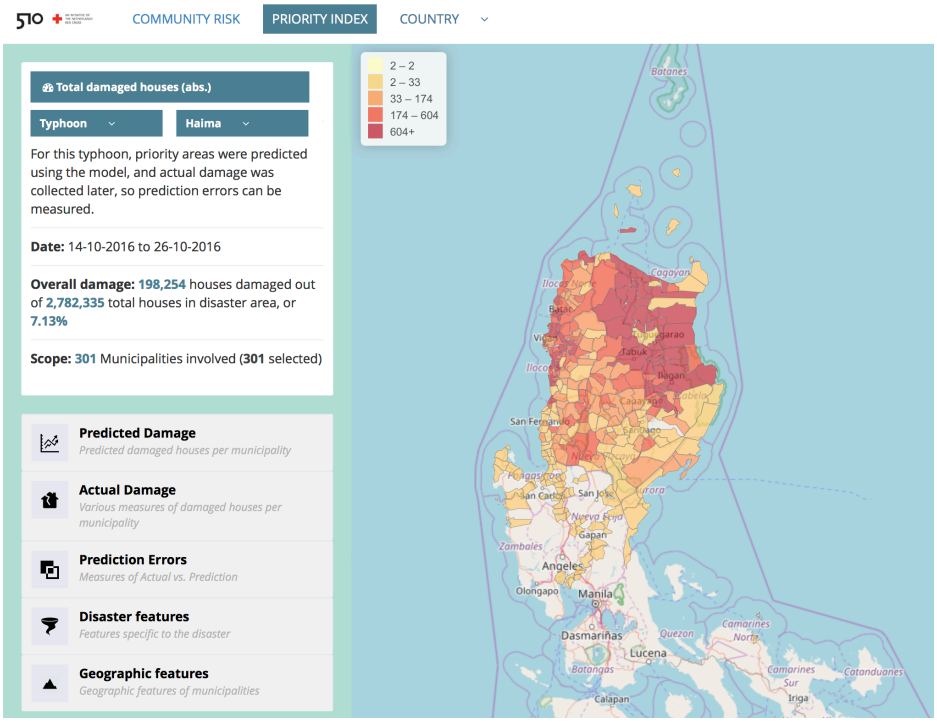


FIGURE 7 – Example of Priority Index dashboard for Typhoon Haiyan in the Philippines. The color indicates the number of houses damaged in that area (510 Global, 2017A).

[2] Both the Community Risk Assessment and the Prioritization dashboard can be accessed through https://profiles.510.global/#!/. The different features and functions are explained on https://profiles.510.global/#!/#walkthrough.

Pre-disaster secondary data on affected areas

As described and visualized in *Figure 5*, the pre-disaster indicators are based on the three dimensions of the INFORM risk index. INFORM is a global, open-source risk assessment for humanitarian crises and disasters. An INFORM risk index is comparable across countries, and shows a detailed picture of risk and its components (INFORM, 2016).

About the rationale behind the three dimensions they use to determine overall risk in a country, they state the following:

> *"The INFORM model is based on risk concepts published in scientific literature and envisages three dimensions of risk: hazards & exposure, vulnerability and lack of coping capacity dimensions. They are conceptualized in a counterbalancing relationship: the risk of what (natural and human hazard), and the risk to what (population).*
>
> *The INFORM model balances two major forces: the hazard & exposure dimension on one side, and the vulnerability and the lack of coping capacity dimensions on the other side. Hazard dependent factors are treated in the hazard & exposure dimension, while hazard independent factors are divided into two dimensions: the vulnerability dimension that considers the strength of the individuals and households relative to a crisis, and the lack of coping capacity dimension that considers factors of institutional strength."* (INFORM, 2016)

510 Global uses those same dimensions to comply to this global 'standard' and to be able to join forces in the future. The reason 510 Global cannot make direct use of the risk index, is that INFORM has so far mainly calculated risks on a country level, whereas 510 Global wants to map the risk scores on lower levels of granularity. INFORM has recently started to also publish subnational data, but only for a limited number of regions and mostly covering higher administrative levels (by 'subnational' they also mean different countries within one region).

That is why 510 Global needs to determine the indicators for each dimension per country individually; when extending the toolbox application to other countries, it must be considered that each subnational Community Risk Assessment model needs different priority indicators based on which the toolbox can generate predictions, given the differences in context, practical needs and data sources available. This means that the subnational Community Risk Assessment models are not numerically comparable across countries: one cannot compare a municipality in country *x* with one in country *y*.

Disaster-specific hazard and exposure data

Apart from the pre-disaster secondary data based on the INFORM index, also the disaster-specific hazard and exposure data is essential to generate predictions on aid priority areas. This data is most valuable when integrated with the pre-disaster baseline data on the affected areas (the 'Community Risk Assessment'), as during a disaster the pre-disaster risks materialize into crisis impact. For example, the number of people affected (post-disaster) has a direct relation with the number of people that are exposed (pre-disaster), and the hospitals and doctors available in the affected area can be estimated based on the hospitals present in the area prior to the disaster, in combination with how they were impacted (Van den Homberg, Visser & Van der Veen, 2017).

Damage data

As to the damage data; the model first generates predictions without including the damage data, which comes in later than the disaster-specific data (that usually becomes available within a couple of hours after the disaster has struck). Comprehensive data on the impact, which can replace the model predictions with actual counts, becomes only available well into the response phase and data is scattered across organizations (Van den Homberg, Monné & Spruit, 2017). This damage data includes for example counts of field assessment teams and damage reports of local authorities. Currently, it is unclear how these counts are structured and organized; it is not a very transparent process. There is strong suspicion that the data resulting from the field assessment counts and local authority reports are not always completely reliable. In the first place, because field assessment teams do not have sufficient resources to physically go to all affected municipalities and count the damage. They must make choices on where to go and estimate damage numbers for those municipalities they could not visit, based on the damage they assessed in other municipalities. On top of that, these teams are not always well trained, and they have different interpretations of when they should report damage to houses as 'destroyed' and when as 'partly damaged': there is no centralized methodology, which leads to wide variations in counts between municipalities (510 Global, 2016A).

This methodology can result in incorrect estimates, but also in more drastic mistakes being made. A recent mapping by 510 Global of these damage counts revealed for example that it regularly occurred that in heavily affected areas, in all municipalities but one a lot of damage was assessed. This is very unrealistic; this one municipality is probably as heavily affected as the others, even though the counts show otherwise.

Additionally, there are no centralized guidelines on how detailed the counts should be reported. An analysis from Typhoon Haiyan showed that many damage reports only included very rough estimates of the population percentages counted as affected (either 0%, 50% or 100%). When applying a learning algorithm, needed in the Community Risk Assessment and Prioritization model, using estimates this rough will not result in a reliable prediction (510 Global, 2016A).

Another issue is that these field assessments are prone to biases. It is for example in the best interest of local authorities who report to assessment teams to exaggerate the amount of damage in their area: more damage reported means that it is more likely that they will receive more resources. It is therefore important to get more insight in these processes, and to find a way to validate these field assessments. A structurally higher quality of the damage data can lead to a structural increase of the Community Risk Assessment and Prioritization toolbox output accuracy.

Timely collection of data

As Harke & de Leeuw describe, "humanitarian aid organizations face a complex and highly dynamic environment where high-impact disaster events occur with low frequency" (p.228). Decision-makers are confronted with compressed timeliness and high levels of uncertainty in this chaotic and disruptive situation (Van den Homberg, 2016; Vitoriano, Montero & Ruan, 2013). In situations of immediate crisis, for example when a natural disaster strikes, there is a need to act quickly. To provide adequate aid, information on the disaster and the affected areas, evidence needed to inform response, must become available quickly enough to matter: there are timing constraints associated to the information needs of disaster responders (Van den Homberg, Monné & Spruit, 2017). This information must be quickly converted into products that are useful in dynamic and uncertain contexts (Van de Walle & Comes, 2015).

This has implications for running the Community Risk Assessment and Prioritization Index toolbox. A field study of Van Lint (2016) in the Philippines during the aftermath of the Typhoon Haiyan confirmed this: the outcomes of the toolbox are only perceived as useful by humanitarian decision-makers when they are published within 12 to 24 hours after a disaster has struck.
This means that it must be possible to run the Community Risk Assessment and Prioritization toolbox rapidly after a disaster has struck. To run the toolbox, the required pre-disaster data must be readily available (it is important that countries are 'data prepared'), and all the necessary post-disaster data must be collected as quick as possible after a disaster has struck.

Pre-disaster: data preparedness
Van den Homberg, Visser & Van der Veen (2017) define data preparedness as "all activities, that can be done before a disaster hits, to pre-stage data with sufficiently high data quality (that matches the prospective information needs of responders) and to develop capacities to collect data on affected communities and areas once a disaster hits to ensure a timely, efficient, and effective response" (p.3). They developed and tested a 'Data Preparedness Framework' that can be used to assess and quantify the level of data preparedness in an organization or in the entire humanitarian network in a country.
According to Raymond & Al Achkar (2017), who proposed a 'data preparedness cycle', data preparedness is "the ability of organizations to be ready to responsibly and effectively deploy data tools before a disaster strikes" (p.3).
Translating this concept to the Community Risk Assessment and Prioritization toolbox, it implies that for the toolbox to be able to run in the dynamic humanitarian context, it is important that a country is data prepared. Having all the required pre-disaster indicators readily available, accessible and up to data already before a disaster strikes saves valuable time during the emergency response.

Post-disaster: quick action
Besides the need for a country to be data prepared, in the direct aftermath of a disaster, specific hazard and exposure data, and later also damage data, must be collected rapidly. For the disaster-specific hazard and exposure data, mainly open data sources are consulted. After a disaster has struck, an online search must reveal which sources quickly opened datasets on the disaster that provide the most useful data that results in the highest prediction accuracy. For almost any type of sudden onset natural disasters there are institutions active that rapidly produce datasets on geophysical characteristics of the hazard itself after its occurrence.

Data collection at low administrative levels

Another challenge to the collection of data in general (in both the preparedness as the response phase) is to reach a high level of administrative granularity. The level of geographical aggregation on which the output of the Community Risk Assessment and Prioritization toolbox *(Figure 7)* is presented is very important. A distinction in data levels is made between regional (administrative level 1), provincial or district (administrative level 2), municipal (administrative level 3) up to community level (administrative level 4). The specific names of these levels vary per country. To increase the effectiveness of the Community Risk Assessment and Prioritization toolbox and improve the value of the model predictions, it is important to have data available from the lowest administrative levels (municipal, community or even household)

so that detailed prioritization can take place. This will enable decision-makers to better target relief resources to the right locations. However, it is especially this data that is difficult to find and collect. Many open datasets are available on province or district level, but much less on municipal and community level. A 'best' data collection to gather data at these lower administrative levels is yet to be found. An important aspect to be considered with this regard is the trade-off between cost and time involved in the data collection and the resulting quality and (ICCO et al., 2016).

Currently, the Community Risk Assessment dashboard (*Figure 6*) for the Philippines shows complete visualizations up to the municipality level, whereas in Malawi the highest level of detail is the district level, due to a lower level of 'data preparedness' in this country (Van den Homberg, Visser & Van der Veen, 2017). The Community Risk Assessment dashboard does also show outcomes for lower administrative levels for Malawi, but these risk scores are determined based on less indicators than desired, and hence less accurate.

To illustrate the difference in level of detail between different administrative levels, *Figure 8* below shows the boundaries of three different administrative levels, 'districts', 'Traditional Authorities' (TAs) and 'Group Village Headmen' (GVHs) in Malawi as visualized in the Community Risk Assessment dashboard (510 Global, 2017A).

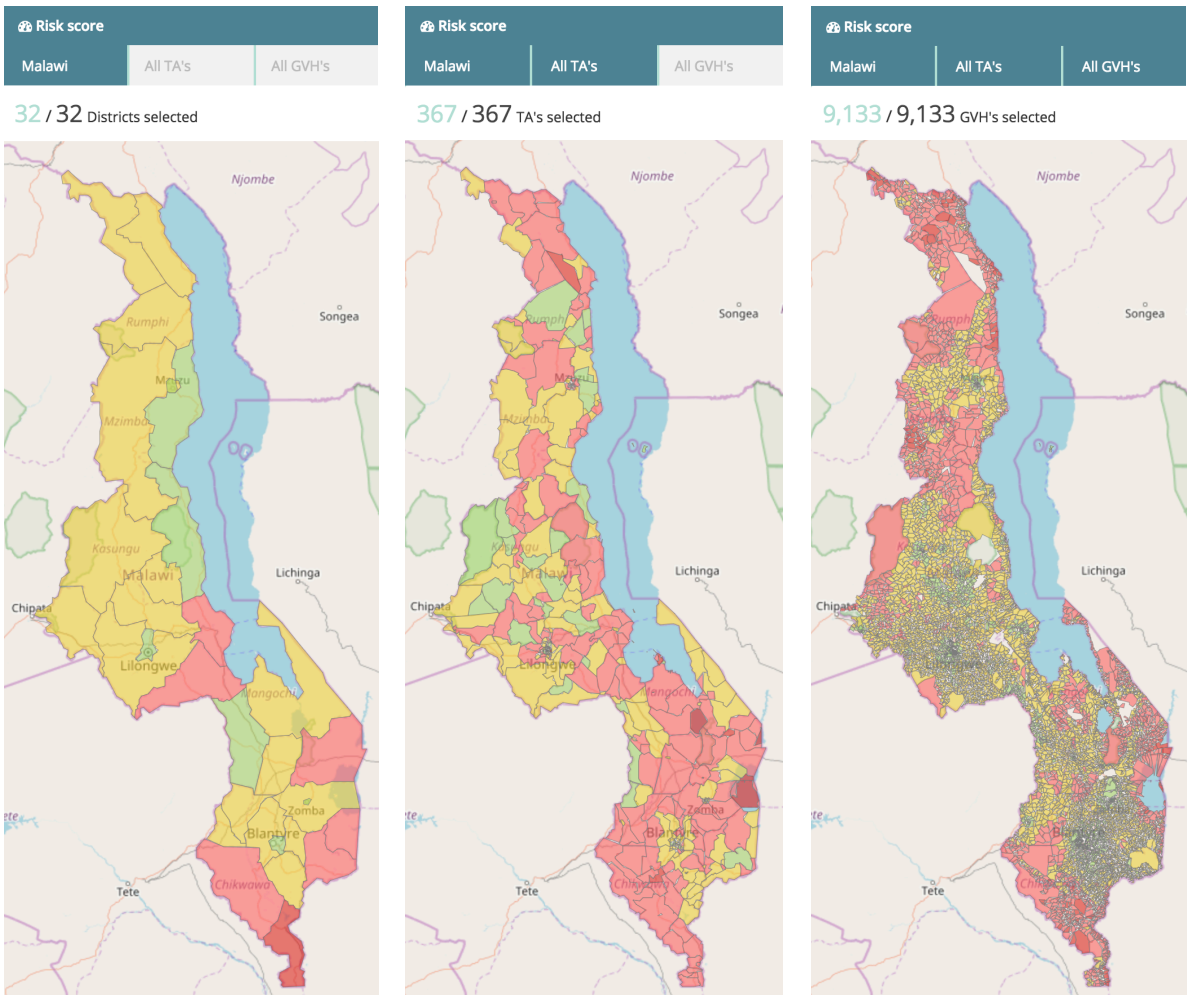


FIGURE 8 – Visualization of differences in administrative levels for Malawi (510 Global, 2017A)

2.2.2. ACTOR COMPLEXITY

A difficulty faced in both the preparedness and the response phase, is that required data is scattered across organizations and across sectors (Van den Homberg, Monné & Spruit, 2017). This introduces another type of complexity in the context of the Community Risk Assessment and Prioritization toolbox: the complex and fragmented actor environment.

There are several aspects that contribute to the complexity and fragmentation of the actor environment around the toolbox. In general, the humanitarian sector is characterized by many different (types of) parties being involved, and a constant movement of those parties and people (Van Lint, 2016). It is not uncommon that a duplication of efforts takes place (Global Pulse, 2015), and the institutional structure is often fragmented. To get an idea of the fragmentation when it comes to the humanitarian sector: in Sri Lanka, over 420 agencies are involved in the monitoring of certain data collection goals.

The result of this complex and fragmented structure is that many parties only focus on their own objectives. They individually start initiatives they can benefit from, but without overseeing the complete picture and having an overview of what other organizations are already doing. There is a lack of central governance to coordinate and align all these individual activities (Berens et al., 2016; Raymond et al.,2016; Sabou, 2016). This trend is amplified by an insufficient level of trust between humanitarian organizations on the ground (Mans, Berens & Shimshon, 2015). Van Lint (2016), who did a field study in the Philippines during the aftermath of the Typhoon Haiyan, found that “although the various agencies strive for sharing their knowledge and cooperating to reach the affected people that need assistance the most, they do not ‘trust’ upon information or numbers that are provided by others, and they trust information provided by their ‘own’ organization or by people they know very well more than information from unknown actors.” (p.57).

Figure 9 below visualizes the main stakeholders around the application of the Community Risk Assessment and Prioritization toolbox, and their roles. As can be seen, there is a very diverse group of data sources, of which an idea is given in *Figure 9*, from which 510 Global collects data. 510 Global either collects this data directly from the source organization (closed data), or retrieves it from open data platforms to which they uploaded their data (open data). It should be stressed here that these data providers differ per country, which also adds to the actor complexity.

510 Global then processes this data, which they made sure is in line with the dimensions from the INFORM risk index *(as described in in Section 2.2.1 – Technical complexity)* in their Community Risk Assessment and Prioritization toolbox, generating manageable information for humanitarian aid workers (decision-makers) in the form of the dashboard shown in *Figure 7.* If humanitarian aid workers integrate this information in their decision-making, the affected people eventually benefit from it (assuming it results in the provision of more timely and adequate aid). In their turn, the affected people could also function as data providers. This is not the case now, but when collected adequately, crowd-sourcing could be used to collect data at lower granular levels, and this data could serve as an input for the Community Risk Assessment and Prioritization toolbox.

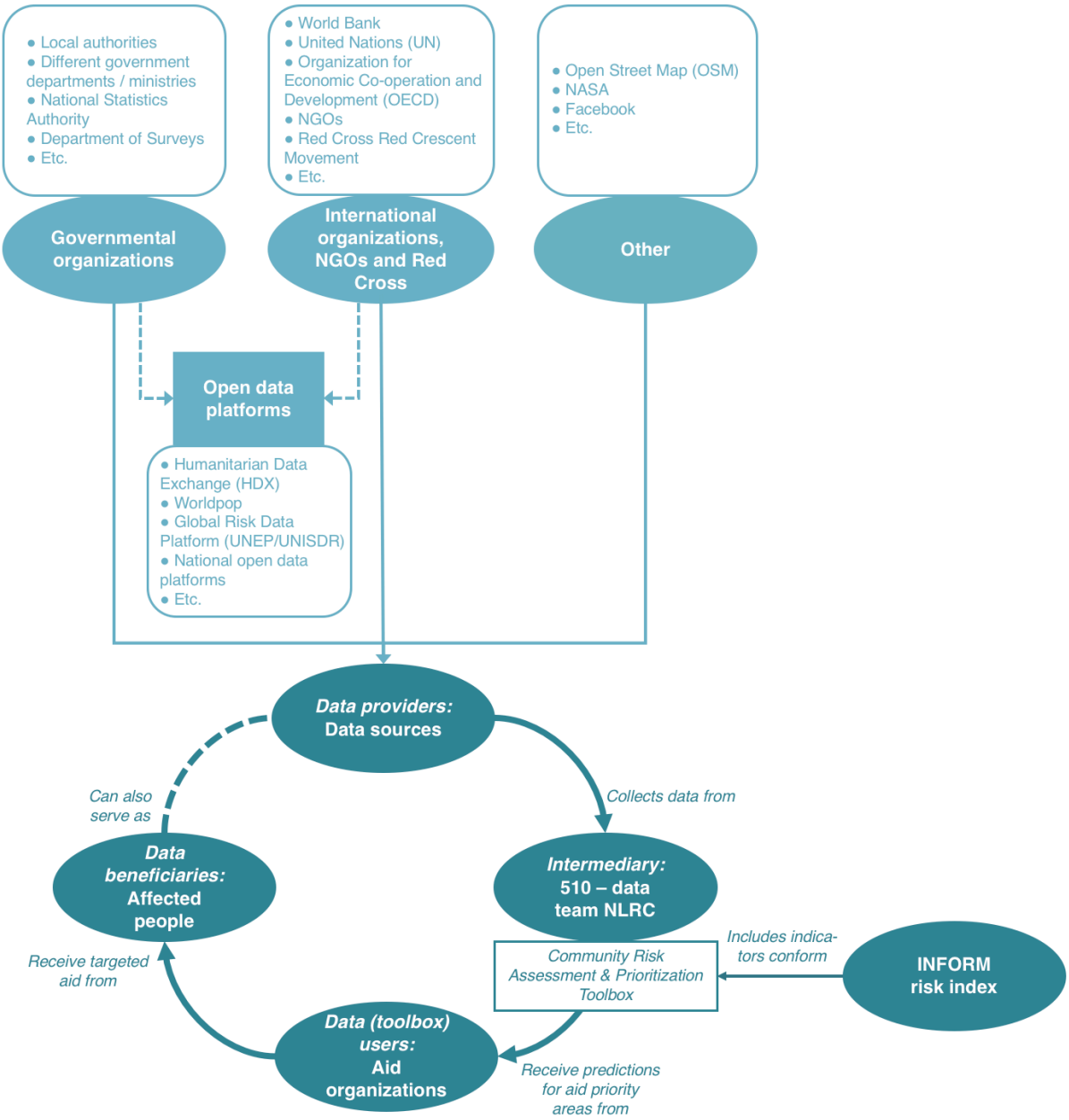


FIGURE 9 – Main stakeholders around Community Risk Assessment and Prioritization toolbox (own figure)

Besides the large variety of data sources, also the 'data user' group contributes to the high level of actor complexity. The data users in this context are the users of the Community Risk Assessment and Prioritization toolbox: the humanitarian decision-makers involved in emergency response. The many parties involved in disaster response can be categorized into three levels, as visualized in *Figure 10* below.

FIGURE 10 – Different operational levels of disaster response, as identified by Vitoriano, Montero & Ruan (2013). (Figure adapted from Van Lint, 2016)

The parties at the local level are the first to respond to a disaster. The parties at the National level usually get involved when a national government classifies an emergency as a disaster. The international response level is reached, when a national government authorizes an international humanitarian operation. This happens when national authorities do not have sufficient capacity to adequately respond to the disaster, due to for example the vulnerability of a country or the scale of the disaster (Vitoriano, Montero & Ruan, 2014).
These different parties collaborating on different levels in the highly dynamic and uncertain situation of disaster response contribute to these decision-making processes being very complex.
Many of the humanitarian decision-maker parties are potential users of the Community Risk Assessment and Prioritization toolbox. However, the initial focus groups of the toolbox are the operational departments of National Red Cross Societies, who are responsible for disaster management and send out field assessment teams to prioritize affected areas, as these are easiest to approach by 510 Global: 510 Global is a team of the Netherlands Red Cross and can therefore make use of the Red Cross network and connections. Additional target groups could be governmental organizations and UN-OCHA. They do their own field assessments, and could therefore also benefit from the toolbox predictions.

2.5. CHAPTER CONCLUSION

This chapter aimed to answer the following sub-question:

SQ1: WHAT CHARACTERIZES THE SYSTEM AROUND THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX?

The complex and dynamic process around scaling up the Community Risk Assessment and Prioritization toolbox should be approached from different perspectives: it is important to consider the technical system behind it, but also the social and institutional dynamics of the environment the toolbox is placed in. These characteristics are typical for so-called 'socio-technical systems'. To design appropriate solutions for problems in complex socio-technical systems, both the technical and the actor complexity must be understood properly.
The technical complexity around the Community Risk Assessment and Prioritization toolbox stems from a combination of different factors. In the first place, different types of data need to be combined. From the 'preparedness phase', pre-disaster secondary data on the affected area, historical hazard and geographic data serve as an input. The specific indicators to cover

the pre-disaster secondary data need to be determined per country based on the country context. In the 'response phase', disaster-specific hazard and exposure data and damage data must be collected. All this data must be collected timely. This means that countries must be 'data prepared', and have all the pre-disaster indicators readily available, accessible and up to date before a disaster strikes. Post-disaster data needs to be collected rapidly, under circumstances of chaos, uncertainty and severe time pressure. Lastly, all data should be collected at the lowers possible administrative level, to be able to generate detailed predictions. This is difficult, as data at these levels is often not stored properly, stored paper-based or not stored at all.

The actor complexity around the Community Risk Assessment and Prioritization toolbox is characterized by a high level of organizational fragmentation and a duplication of roles. Many different organizations are involved, and data is scattered across different organizations and sectors. There is a great variety of (potential) data providers, but also in potential users of the Community Risk Assessment and Prioritization toolbox. Moreover, the actors involved vary per country, and therefore need to be mapped for each country individually.

To be able to come to a sound governance strategy for scaling up the toolbox, this complexity needs to be dealt with and should be considered.

CHAPTER 3 – THEORETICAL PERSPECTIVE: DATA ECOSYSTEMS

3.1. INTRODUCTION

A perspective that can be used to deal with complexity and understand interrelationships between agents and entities in socio-technical systems, which is a crucial step to get an idea of a suitable generic governance structure for scaling up the Community Risk Assessment and Prioritization toolbox, is the 'ecosystems' approach.
All data producing and consuming actors in the humanitarian sector together form a complex and diverse 'humanitarian data ecosystem' (Raymond et al., 2016). According to Heimstädt, Saunderson & Heath (2014), to analyze the provision and use of (open) data by different parties, which needs to be done in this case, an appropriate framework for analysis is required. The biological understanding of an ecosystem, that opposes unidirectional models of causality and development, and focuses on understanding interrelationships and interdependencies between agents and entities, could function as this framework for investigation, as it has proved valuable in similar investigations before (Heimstädt, Saunderson & Heath, 2014). Academics in information intensive, socio-technical contexts have applied the ecosystems perspective to analyze and describe the diverse interrelationships between data users, data providers, data itself, institutions and material infrastructure (Harrison, Pardo & Cook, 2012), which all relate to the problematic issues in the context of scaling up the Community Risk Assessment and Prioritization toolbox, as described in *Section 1.2 – Problem definition.*
The metaphor can be used in any socio-technical domain (Harrison, Pardo & Cook, 2012).
This chapter will therefore provide a framework coming from a systematic literature review around 'data ecosystems', that can serve as the theoretical foundation for the start of the creation of a generic governance structure to scale up the Community Risk Assessment and Prioritization toolbox. This framework should reduce the complexity as presented in the previous chapter, and provide insight in the factors that should be considered to come to a generic governance structure, to ensure that no relevant aspects will be overlooked.
The following sub-question will be answered:

SQ2: WITHIN THE FIELD OF DATA ECOSYSTEMS, WHICH FRAMEWORK CAN BE USED FOR THE CREATION OF A GENERIC GOVERNANCE STRUCTURE FOR THE TOOLBOX?

a. *What is a data ecosystem?*
b. *What are the criteria for a successful data ecosystem, according to literature?*

This chapter is structured as follows. First, the concepts that describe what is meant by a (data) ecosystem, the characteristics, and its functioning will be discussed in paragraph 3.2. Next, in paragraph 3.3, the methodology used to collect suitable literature for the systematic literature review will described, and an overview of the selected literature will be provided. A review of

this literature will then provide insight in the essential components of data ecosystems in paragraph 3.4. In paragraph 3.5, these findings are summarized.
The output of this sub-question will be a framework of important aspects to consider when establishing, enabling or improving a data ecosystem, which can serve as a point of departure for scaling up the Community Risk Assessment and Prioritization toolbox.

3.2. THE ECOSYSTEMS APPROACH

3.2.1. WHAT AND WHY?

An ecosystem is defined as “a system of people, practices, values, and technologies in a particular local environment” (Nardi & O’Day, 1999: p. 49), or “the people and technologies collecting, handling, and using the data and the interactions between them” (Parsons et al., 2011: p.557). According to Harrison, Parco & Cook, “ecosystems are comprised of interacting components with substantial interdependencies” (p.906). The specific components are different for each ecosystem (Harrison, Pardo & Cook, 2012). Zuiderwijk, Janssen & Davis (2014) distinguish different stakeholders that are involved in ecosystems: agencies and departments, suppliers and partners, internal elements, communities and individual stakeholders. Additionally, often intermediaries are involved, who operate between the final customer and the data provider (Zuiderwijk, Janssen & Davis, 2014).
The idea of the ecosystem metaphor is an image derived from the natural world that is now widely used in different domains that deal with complex organizational and technological contexts, and interdependencies among technology, data and innovation. The metaphor has for example been applied to analyze strategical business contexts, to describe the ‘Internet’ and to explain open source software communities, like the Linux operating system (Harrison, Pardo & Cook, 2012).
Harrison, Pardo & Cook (2012) describe ecosystems as typically containing “interdependent social systems of actors, organizations, material infrastructures, and symbolic resources that can be created in technology-enabled information-intensive social systems” (p. 900). They also state that ecosystems can be “seeded, modeled, developed, managed, that is, intentionally cultivated for achieving a managerial and policy vision” (p.907). The metaphor can be used to outline existing conditions, or to develop desired conditions, stating that “its users often aim to provoke new thinking about the conditions and requirements necessary to actively cultivate development of an ecosystem to achieve a set of specific and desirable goals” (p.905). This application of the ecosystems approach is expected to be valuable in addressing the defined problem definition around scaling up the Community Risk Assessment and Prioritization toolbox, as it implies that insight in the requirements for the development of a well-functioning humanitarian data ecosystem will help to achieve the goal of coming to a generic governance structure.
A similar clear idea of how the ecosystems approach can be applied, comes from Dawes, Vidiasova & Parkhimovich (2016). They argue that the ecosystems perspective can be a means to assess existing conditions in a system, like for example strengths and weaknesses of different relationships and components. By doing so, the perspective can help designers to get an idea of the necessary changes in strategies, policies, practices, participants or relationships to improve those existing conditions; to see “where initial investments are most needed in policies, capabilities or connections to achieve their specific goals” (p.26). This application can directly be translated to the context of scaling up the Community Risk

Assessment and Prioritization toolbox. Additionally, Dawes, Vidiasova & Parkhimovich (2016) believe that an important characteristic of an ecosystem is that it is dynamic, that it changes over time, and that this is an important factor to consider, which also holds for the toolbox.
Following this introduction to data ecosystem perspectives and its potential applications, the context of the Community Risk Assessment and Prioritization toolbox can be perceived as a data ecosystem, as it includes interdependent technical and social systems. The ecosystems approach can help to structure, deal with and design solutions for the issues explained in *Section 1.2 - Problem definition.*
A systematic literature review around data ecosystems theories and the resulting criteria for a well-functioning data ecosystem therefore seems an appropriate means to form the bridge to the creation of a generic governance framework to facilitate the process of scaling up the Community Risk Assessment and Prioritization toolbox in a new country. It will allow for the assessment of the data-related interdependencies between the stakeholders involved in the application of the toolbox, and can inform policy development and implementation.

3.2.2. Different ecosystem applications

Harrison, Pardo & Cook (2012) state that the "metaphor may be applied to any existing socio-technical domain" (p.907); a claim supported by a scan of existing literature. The concept is used to describe 'business ecosystems', which are defined as "an economic community supported by a foundation of interacting organizations and individuals – the organisms of the business world" (Moore, 1996: p.9), 'innovation ecosystems', being "the collaborative arrangements through which firms combine their individual offerings into a coherent, customer-facing solution" (Adner, 2006: p. 41), and 'digital ecosystems', which consist of "interconnected, interrelated and interdependent digital species situated in a digital environment, that interact as a functional unit and are linked together through actions, information and transaction flows" (Hadzic & Sidhu, 2008: p.7). Moreover, regarding Harrison, Pardo & Cook (2012), there are multiple scholars who "propose the idea of an ecosystem as an organizing framework for discussions about open government" (p.907). This view takes up the idea that the government should adopt a central role, and should take the lead in achieving innovative goals within networked systems (Harrison, Pardo & Cook, 2012).
Yet another approach is to view open data as an ecosystem. This view partly overlaps with the open government ecosystems lens, because much open data is published by the government. In open data ecosystems however, it is possible for elements to be provided by the private sector as well (Zuiderwijk, Janssen & Davis, 2014). Open data ecosystems can be used "as a means for decision-making and planning" (Zuiderwijk, Janssen & Davis, 2014: p.18), to locate the "relative positions of the actors in the ecosystem (data providers, sources, resources and users)" (Van Schalkwyk, Willmers & McNaughton, 2016: p.72), and to "facilitate access to sharing and using data" (Welle Donker & Van Loenen, 2017: p.286). Immonen, Palviainen & Ovaska (2014) stress the importance of a 'collaborative environment' to create open data ecosystems, which is also acknowledged by Zuiderwijk, Janssen & Davis (2014).

3.2.3. The context of the Community Risk Assessment and Prioritization toolbox as a data ecosystem

This last ecosystem approach, the open data one, shows most similarities with the context of the Community Risk Assessment and Prioritization toolbox, in which humanitarian data needs to be shared and collected. However, this data does not necessarily need to be 'open' as

defined by Janssen, Charalabidis & Zuiderwijk (2012), who define open data as “non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distribution” (p.258). Apart from the fact that not all these criteria matter when collecting the data required for the toolbox (its funding is irrelevant; what matters is if it can be retrieved), it is not clear yet if all required data is, or needs to be open in this sense.
Additionally, Van Schalkwyk, Willmers & McNaughton (2016) describe that “it is acknowledged that data can be interpreted and classified along a continuum ranging from completely open to completely closed depending on the criteria used to assess openness and, ultimately, on how the authors of the assessment method expect open data to make an impact in a particular context” (p.71). It is therefore suggested to, in this specific humanitarian context, reflect on the degree of openness as a possible feature of a data ecosystem rather than an inherent characteristic.
Thus, for the humanitarian context, the establishment of a ‘data ecosystem’ in general rather than an ‘open data ecosystem’ should be considered: the data ecosystem around the Community Risk Assessment and Prioritization toolbox should be assessed as a whole, instead of ‘just’ the open data. The aim was therefore to take literature on data ecosystems in general as the theoretical basis for this study, but it was found that most of this literature is centered around open data. This is not an issue, as it will largely be applicable, though it will be important to remain a critical attitude towards aspects that are typical for open data.

3.3. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

A systematic literature review will help provide an overview of the criteria for a successful data ecosystem, combining the outcomes from relevant publications.
According to the ‘typology of reviews’ from Grant & Booth (2009), “literature reviews lack an explicit intent to maximize scope or analyze data collected” (p. 97). *Systematic* literature reviews are generally more scoped and detailed; they are characterized by taking a more rigorous and well-defined approach. A systematic literature review seeks to put together and synthesize all known knowledge on a topic area (Grant & Booth, 2009) and that is exactly what is needed in this case. This is the reason why a systematic literature review has been selected as the preferred methodology to create an overview of the criteria for a successful data ecosystem.
Khan et al. (2003) have defined the following steps to be completed when carrying out a systematic literature review:

1. *Framing questions for a review*
2. *Identifying relevant work*
3. *Assessing the quality of this work*
4. *Summarizing the evidence*
5. *Interpreting the findings*

These steps have been followed. The question to be answered with this systematic literature review is sub-question *2b* of this research: *What are the criteria for a successful data ecosystem, according to literature?*
The literature study was carried out between 17-04-2017 and 25-05-2017. Articles published after this data have therefore not been taken into consideration.

Suitable literature was found by means of a thorough search among journal and conference articles, books, reports and other informative documents. Scopus, Web of Science and Google Scholar and Google Search were consulted as online databases. The following (combinations of) keywords were used to search these databases:

- "Data ecosystem"
- "Humanitarian data ecosystem"
- "Open data ecosystem"
- "Government ecosystem"
- "Criteria" + "data ecosystem"
- "Element" + "data ecosystem"
- "Successful" + "data ecosystem"
- "Establish" + "data ecosystem"
- "Development" + "data ecosystem"
- "Data infrastructure"

The search of relevant articles was an iterative process. When scanning through found publications, it was for example found that the definition of 'data infrastructures' resembles the described definition of data ecosystems: "A SDI (Spatial Data Infrastructure) can be defined as the integration of a number of components to create a platform which enables a wide variety of stakeholders to access, share and use spatial data in an efficient and effective way" (Macharis & Crompvoets, 2004, p. 45). The term 'data infrastructures' was therefore added to the search terms.

By inserting (combinations of) these keywords, a multitude of articles was found. To give an indication of the amount of documents the searches resulted in; by searching on "data ecosystem" in Scopus, 147 documents were found (of which around ten were focusing on 'natural ecosystems', which is not applicable at all in this context). When changing the search term into "open data ecosystems", only 38 documents remained and searching for "Government ecosystem" resulted in five documents. Because inserting the term "data infrastructures" resulted in over 2000 documents – a remarkably high amount compared to the documents found on data ecosystems –, this search term was further specified by (individually – not at the same time) adding the terms "criteria", "elements" and "successful". Also, Google Search results (combining all the different search terms) were scanned extensively, to look for additional 'grey literature' not present in the scientific databases.

The found articles were judged on relevance by reading the titles and summaries, and by scanning the text. The most important aspect in this selection phase was the knowledge that the focus of the systematic literature review would not be on data ecosystems in general or on one specific part of or role in a data ecosystem, but on the whole range of *criteria* that need to be considered when *establishing* a data ecosystem: all selected articles must be centered around the successful establishment of a data ecosystem. To check for this requirement, the articles were sought for terms like '(essential) elements/components/criteria of data ecosystems' and 'enabling/building/creating/establishing/realizing/developing a data ecosystem'. Subsequently, a selection was made based on the presence of these keywords. Articles that did not include any of these keywords and did not explicitly address the success criteria or the development of a data ecosystem, were omitted.

When an initial selection was made, an additional search for literature took place based on the references used in these selected articles, which resulted in an increase of articles with potential. The new articles were scanned in the same manner, and selected based on the mentioned requirement.

In total, 17 publications were found to be relevant in this systematic literature review. *Table 1* below lists the selected articles, of which the review can be found in the next section. The objectives and methodologies described have been retrieved from the articles.

	AUTHOR	FOCUS	OBJECTIVE	DOMAIN OF AUTHORS	METHODOLOGY OF STUDY
1.	Attard, Orlandi & Auer (2016)	Economic data ecosystem	To project their vision of generating a new Economic Data Ecosystem that has the Web of Data as its core.	Enterprise Information Systems	- Model design by means of a literature and practice review - Model evaluation by means of a survey
2.	Barthélemy (2016)	Open data ecosystem	To provide an overview of the Belgian open data ecosystem	School of Management: MSc in Business Engineering	Master thesis: - Literature review - Surveys/interviews - Case study
3.	Davies (2012)	Open data infrastructures and ecosystem	To highlight some of the interventions that may be necessary to support realization of impact from open data initiatives.	Open Data & Open Government	Argumentative text based on domain description
4.	Dawes, Vidiasova & Parkhimovich (2016)	Open government data ecosystem	To develop a preliminary ecosystem model for planning and designing Open Government Data programs	E-governance / Open Government	Literature review combined with findings from two empirical case studies
5.	Harrison, Pardo & Cook (2012)	Open government ecosystem	To create a research and development agenda with questions essential to the development of Open Government Ecosystems.	IT in Government / Open Government	- Literature review serves as framework - To consider the outcomes of a workshop with a multi-sector expert group
6.	Heimstädt, Saunderson & Heath (2014)	Open data ecosystem	To identify a set of structural business ecosystem properties.	Open (Government) Data	- Review of popular Open Data definitions and business ecosystem theories - Applied to empirical data using a timeline analysis - Work informed by combination of discourse analysis and in-depth interviews
7.	Immonen, Palviainen & Ovaska (2014)	Open data based business ecosystem	To define the requirements of an open data based business ecosystem (an open data ecosystem from the business viewpoint)	Service / Software Engineering	- State-of-the-art knowledge explored from the literature - State of the practice on data-based business in the industry collected through interviews
8.	Jetzek (2017)	Open data ecosystem	To explore the possibilities for sustainable value generation in the Open Data Ecosystem	IT Management: Big Data & Open Data	Theoretical review combined with case study (interviews and participation in an open data initiative in Denmark)
9.	Lee (2014)	Open data ecosystem	To specify a series of specific elements critical for building an Open Data Ecosystem	Linked & Open Data	Literature review combined with liaising with stakeholders

	AUTHOR	FOCUS	OBJECTIVE	DOMAIN OF AUTHORS	METHODOLOGY OF STUDY
10.	Macharis & Crompvoets (2014)	Spatial data infrastructure	To evaluate development scenarios for the spatial data infrastructure for Flanders	Supply Chain Management and Policy Measures / Public Governance / Spatial Data Infrastructures	Multi Actor Multi Criteria Analysis (MAMCA) applied to a case study
11.	Parsons et al. (2011)	Science data ecosystem	To success several short- and long-term strategies to facilitate a socio-technical evolution in the overall science data ecosystem	Science Data	Case study
12.	Pollock (2011)	Open data ecosystem	To stress the importance of data cycles with feedback loops when building open data ecosystems	Open Knowledge	Blogpost
13.	Ponte (2015)	Open data ecosystem	To provide an overview of the issues to be addressed when enabling an open data ecosystem	Organization Science & ICT	Survey of the open data market
14.	Van Schalkwyk, Willmers & McNaughton (2016)	Open data ecosystem	To consider the supply, demand and use of open data, as well as the roles of intermediaries, using an ecosystems approach	Open data / Open ICT Ecosystems & Scholarly Communication	Case study
15.	Welle Donker & Van Loenen (2017)	Open data ecosystem	To develop an open data assessment framework based on three output indicators as conditions for a successful open data ecosystem	Open Data / Information Infrastructures	- Framework analysis (domain description) - Framework development - Case study
16.	Wiener et al. (2016)	Open data ecosystem	To enable an open data ecosystem for the neurosciences	Neuroscience & Data-sharing	Workshop with actors involved
17.	Zuiderwijk, Janssen & Davis (2014)	Open data ecosystem	To provide an overview of essential elements of open data ecosystems for enabling easy publication and use of open data	Open Data / ICT & Governance	-Literature review -Scenario analysis

TABLE 1 – List of articles reviewed (objectives and methodologies have been retrieved from the articles)

The completion and relevance of this set of publications was validated with a group of open data experts from the Delft University of Technology (which is further explained in *Section 3.5 – Expert validation*), and no additional articles were identified.
As can be seen, the majority of the articles is centered around open data ecosystems and many of the authors have a background in open data/open government. As already described in *Section 1.3.4 – Societal and scientific relevance,* searching for "humanitarian data ecosystem" led to zero scientific hits and a limited set of six documents classified as 'grey literature', which did not specifically address the success criteria for a well-functioning humanitarian data ecosystem, and were hence not included in the set of selected articles.

3.4. CRITERIA FOR A SUCCESSFUL DATA ECOSYSTEM

There is a diverse set of aspects that needs to be considered when establishing a successful data ecosystem. This section will provide an overview of these aspects, based on a systematic review of the literature that is available on this topic. The 'three output indicators as conditions for a successful open data ecosystem' of Welle Donker & Van Loenen (2017) have been used to structure this review. Welle Donker & Van Loenen (2017) have developed a framework that holistically assesses these three output indicators, being open data supply, open data governance and open data user characteristics. Their framework "evaluates the maturity of the open data ecosystem" (p.284) and can serve as "a tool to indicate which aspects of the open data ecosystem are successful and which require attention" (p.284). The success criteria for a data ecosystem that resulted from the systematic literature review have been organized according to these output indicators (which thus served as categories). The findings are summarized in *Table 2* below, and further elaborated on from *Section 3.4.1* onwards.

CATEGORY	CRITERION	DESCRIPTION	SOURCE
1. DATA SUPPLY	Data availability and accessibility	It should be clear for actors in a data ecosystem how to find data, where to find it and how to access it	• Attard, Orlandi & Auer (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016) • Zuiderwijk, Janssen & Davis (2014)
	Metadata	The existence of appropriate metadata can help to improve the availability and accessibility of the data	• Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016) • Zuiderwijk, Janssen & Davis (2014)
	Licensing and legal conditions	Another important aspect to enhance the accessibility of data in a data ecosystem is the type	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Immonen, Palviainen & Ovaska (2014) • Lee (2014)

CATEGORY	CRITERION	DESCRIPTION	SOURCE
		of licensing associated with the data, which is necessary to ensure the legal grounding for the potential (re)use of data	• Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Zuiderwijk, Janssen & Davis (2014)
	Privacy	It should be ensured that data protection laws are followed	• Dawes, Vidiasova & Parkhimovich (2016) • Immonen, Palviainen & Ovaska (2014), • Lee (2014) • Wiener et al. (2016)
	Data usability	High quality of the data shared in a data ecosystem should be ensured to enhance its usability	• Attard, Orlandi & Auer (2016) • Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Zuiderwijk, Janssen & Davis (2014)
	Data standards	An important aspect that determines the quality of data is the presence of standards to facilitate data interoperability	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
2. DATA GOVER-NANCE	Vision, communication and stimulation	A collaborative, interactive environment should be established and cooperation between stakeholders should be stimulated	• Attard, Orlandi & Auer (2016) • Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Immonen, Palviainen & Ovaska (2014) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016) *(And other scholars implicitly: a 'collaborative environment' is a characteristic inherent to ecosystems)*
	Division of roles and responsibili-ties	There should be a clear division of the roles and responsibilities of the actors in an ecosystem	• Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Van Schalkwyk, Willmers & McNaughton (2016) • Zuiderwijk, Janssen & Davis (2014)
	Feedback	The data ecosystem should include feedback mechanisms to enable data users to provide feedback to data providers	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Pollock (2011)

CATEGORY	CRITERION	DESCRIPTION	SOURCE
			• Zuiderwijk, Janssen & Davis (2014)
	Leadership and incentivization	Activities in an ecosystem should be stimulated, incentivized and coordinated (either top-down or bottom-up, depending on the context)	• Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
	Self-organizing capacity	Data supply should match data demand in an ecosystem; public bodies require certain capacities to be able to participate in an ecosystem	• Attard, Orlandi & Auer (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Lee (2014) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
	Sustainability: financing and value creation	For a data ecosystem to become sustainable, sustainable financing should be arranged and the value should be generated for the ecosystem stakeholders	• Davies (2012) • Harrison, Pardo & Cook (2012) • Heimstädt, Saunderson & Heath (2014) • Jetzek (2017) • Lee (2014) • Parsons et al. (2011) • Ponte (2015) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
3. USER CHARAC-TERISTICS	User capabilities	The capabilities of the data users in an ecosystem should be considered	• Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Lee (2014) • Welle Donker & Van Loenen (2017) • Zuiderwijk, Janssen & Davis (2014)
	User needs	The needs of the data users in an ecosystem should be considered	• Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Welle Donker & Van Loenen (2017)
A. EXTERNAL CONTEXT	-	The external context of the data ecosystem should be considered	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Zuiderwijk, Janssen & Davis (2014)

TABLE 2 – Criteria for a successful data ecosystem found in the systematic literature review

3.4.1. CATEGORY 1 - DATA SUPPLY

The first factor that indicates the success of an open data ecosystem identified by Welle Donker & Van Loenen (2017) is 'data supply': "the way in which data are provided as open data" (p.288). Welle Donker & Van Loenen (2017) use the concentric shell model of Backx

(2003) to get insight in the supply of open data from a user perspective: what are the steps users should follow to assess if data meets their requirements?

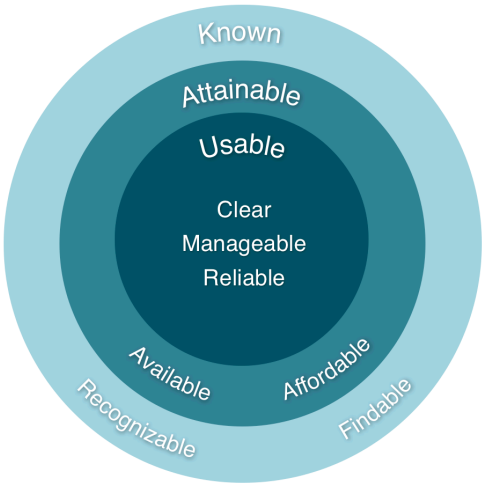


FIGURE 11 – Concentric shell model of Backx (2003)

Element 1.1 - Data availability and accessibility

As can be seen in *Figure 11*, this model consists of three different shells. In the first place, the user needs to know how to find the data ('recognizable'), and where to find it ('findable'). If a dataset is available but it cannot be found through a simple search, then it is conceivable that the data will be overlooked. The second shell relates to the 'attainability' of the data, which indicates that a user must be able to physically access, collect and afford the data, and to be allowed to use the data

The importance of these two first shells for the success of a data ecosystem is acknowledged in other literature as well, though more often designated by the terms 'availability' (known) and 'accessibility' of the data.

According to Immonen, Palviainen & Ovaska (2014), one of the requirements of an ecosystem is that it should provide information about the available data. Harrison, Pardo & Cook (2012) state about open government ecosystems that a critical assumption is that "the data itself is readily available and once made accessible, is fit for any use a citizen, entrepreneur, activist or government might make of it" (p. 919).

Macharis & Crompvoets (2012) defined access to and availability of data as one of the main criteria for a so called 'Spatial Data Infrastructure' (SDI), a concept that shows similarities with the characteristics of a data ecosystem: they describe it as "the integration of a number of components to create a platform which enables a wide variety of stakeholders to access, share and use spatial data in an efficient and effective way" (p. 45). Moreover, Davies (2012) highlights that several open data initiatives involve the creation of new data infrastructures and that the deployment of twin concepts of infrastructure and ecosystem to describe technical and socio-technical systems is nothing new, which confirms the overlap and hence applicability of the SDI criteria in this data ecosystems context.

Wiener et al. (2016) identify 'discoverability' as one of the factors to be addressed when initiating a data ecosystem and enabling its survival, stating that sharing alone will not result in the desired open data benefits if datasets are not discoverable. In other words: both data availability and accessibility are required. In Attard, Orlandi & Auer (2016), 'data distribution' and 'data discovery' are the two most important components of their 'Data Value Network',

which models the co-production of value through the interaction of involved stakeholders to project their vision of the generation of a new 'Economic Data Ecosystem'. They claim that without the existence (availability) of data and its consumption (accessibility through distribution), "there can hardly be any value creation upon a data product" (p. 455). Also, Van Schalkwyk, Willmers & McNaughton (2016) stress the importance of availability and accessibility of data for a successful data ecosystem, focusing on ways to improve data access. Additionally, they state that "datasets indicated as closed are not necessarily inaccessible" (p.74), which is important to keep in mind when analyzing the data context of the Community Risk Assessment and Prioritization toolbox. However, it is conceivable that even though closed data is accessible, this does have implications for how the data should be collected. As most literature concerns 'open data' specifically, it is mostly assumed that when data is made accessible, it can be collected automatically. This does not hold for 'closed' data. The result is that most open data ecosystems literature does not focus on data collection specifically. Remarkably, the only article on data ecosystems found for this literature review that does not specifically focuses on 'open' data ecosystems in one or another form, but on general data ecosystems (although specified for science data - Parsons et al., 2011), also structurally uses the term 'data collection'.

To continue, Zuiderwijk, Janssen & Davis (2014) implicitly mention data availability and accessibility in their first two of the four key elements that open data ecosystems should capture (1: "releasing and publishing open data on the internet" (p.30) – which enhances availability, and 2: "searching, finding, evaluating and viewing data and their related licenses" (p.30) – which relates to accessibility). Furthermore, 'data access' and 'data discovery' (described as related to a user knowing certain data exists and knowing where to find it) are two of the 'Open Data Ecosystem Elements' from Lee (2014), who says that "each element plays a crucial role in an Open Data Ecosystem and needs to be addressed in order to support a holistic and sustainable initiative" (p. 354). In Dawes, Vidiasova & Parkhimovich (2016), 'data publication' and 'data use' (search and identification), which both encompass data availability and accessibility, are 'essential components' in their 'preliminary ecosystem model of open government data programs'. And lastly, also Parsons et al. (2011) believe that data in an ecosystem should be available and accessible, saying that "data must be capable of being located, identified and generally assessed" (p.557).

Element 1.1.1 - Metadata

An aspect that can help to improve the availability and accessibility of data, is the existence of metadata (Welle Donker & Van Loenen, 2017). Metadata provides descriptions about other data by summarizing basic information, for example how the data is formatted, and when and by whom a particular set of data was collected. Parsons et al. (2011) state that "to date, most approaches to facilitate data discovery require the creation of formally structured descriptions, i.e. metadata records, that are then submitted to centralized registries" (p. 558). Furthermore, the study of Macharis & Crompvoets (2012) showed that correct and adequate metadata can improve the access to different types of data, and also Davies (2012), Harrison, Pardo & Cook (2012), Lee (2014) and Wiener et al. (2016) state that appropriate metadata about datasets allows users to identify and find certain data. According to Zuiderwijk, Janssen & Davis (2014) "metadata may improve storing, preservation, accessibility, visualization and multilinguality of

open data" (p.29), and Dawes, Vidiasova & Parkhimovich (2016) mention that there should be metadata requirements for the publication of data for public use.

Element 1.1.2 - Licensing and legal conditions

Another important aspect in the accessibility of data is the type of licensing associated with the data, which is necessary to ensure the legal foundation for the potential (re)use of data (Welle Donker & Van Loenen, 2017). An individual, organization or company that wants to use data, requires assurance of what they are legally allowed to do and not to do with the data. If no license is indicated, the data owner or publisher must be contacted by the data user on a case-by-case basis (Lee, 2014). Unclear license conditions may form a barrier for potential users. Therefore, according to Welle Donker & Van Loenen (2017), "it is important that there are policies in place that define the legal context" (p.286). Apart from licensing being one of the Open Data Ecosystem Elements according to Lee (2014) and Dawes, Vidiasova & Parkhimovich (2016) stating that "a legal framework that sets clear responsibilities and limitations lays the essential foundation for an Open Government Data program" (p.24), also Barthélemy (2016), Immonen, Palviainen & Ovaska (2014), Van Schalkwyk, Willmers & McNaughton (2016) and Zuiderwijk, Janssen & Davis (2014) mention licensing and related legal conditions as an important element to be addressed in a data ecosystem.

Element 1.1.3 - Privacy

Remarkably, the importance of privacy when establishing a data ecosystem is not mentioned a lot in the literature reviewed. It comes forward in the article of Immonen, Palviainen & Ovaska (2014), who discuss open business data: the companies they interviewed highly valued privacy, as a large amount of their data concerned customer's data or data about individual persons. Also Wiener et al. (2016), who assess the establishment of open data for the neurosciences, mention the importance to address ethics and privacy issues related to "maintaining datasets containing sensitive, proprietary, or identify information" (p.619).
Lee (2014) argues that it should always be ensured that data protection laws are followed and that data cannot be linked back to an individual and also Dawes, Vidiasova & Parkhimovich (2016) acknowledge this, stating that when publishing data for public use, there should be requirements regarding privacy and confidentiality protections.

Element 1.2 - Data usability

To get back to the model of Backx (2003) in *Figure 11,* the last success criteria related to the supply of data is about usability of the data. This is the inner shell, because users will only be able to see whether they consider the data to be suitable after the data can be physically inspected; after it has been made accessible and has been collected. According to Welle Donker & Van Loenen (2017), aspects within this shell relate to data quality, which they think is determined by for example "available data formats, available documentation/metadata, level of coverage, timeliness and update frequency" (p.289). Additionally, they mention reliability and guarantees for continuous availability as being part of data quality (Welle Donker & Van Loenen, 2017). Lee (2014), who identify 'High-Quality Data' as one of the Open Data Ecosystem Elements, describe similar aspects that together make up the data quality. Also in Macharis & Crompvoets (2012) data quality and accuracy comes forward as a success criterion for their 'Spatial Data Infrastructure', and in Attard, Orlandi & Auer (2016) data 'quality

aspects' are mentioned as a factor in one of the Data Value Network components ('data curation'). Van Schalkwyk, Willmers & McNaughton (2016) stress the importance of data 'utility' and state that Iit requires the investment of resources to provide and collect data in "a format and context that ensures greater probability of use and impact" (p.77). According to Immonen, Palviainen & Ovaska (2014), one of the requirements of a data ecosystem that enhances its usability is that there should be a data quality verification service in the ecosystem, and that the ecosystem should define how data is created. In the study of Ponte (2014) on how to enable an open data ecosystem, it was found that one of the most important open data ecosystem issues to be addressed is the technical quality of the published data; data providers need to assure that data is correct. This is also acknowledged by Zuiderwijk, Janssen & Davis (2014), who identify the presence of a quality management system that enables users to determine whether they find the quality of the open data satisfactory as an essential element for open data ecosystems. Other scholars focusing on data usability and quality as a critical component in data ecosystems are Barthélemy (2016), Davies (2012), Dawes, Vidiasova & Parkhimovich (2016), Harrison, Pardo & Cook (2012) and Parsons et al., (2011).

Element 1.2.1 - Data standards

An aspect of data quality (as defined by Lee,2014, and Welle Donker & Van Loenen, 2017) that comes forward more explicitly in literature is the presence of standards to facilitate data interoperability. Data standards help to provide a universal meaning to data, which is especially important when data is shared between different organizations, or even countries (Lee, 2014). Immonen, Palviainen & Ovaska (2014) state that in an ecosystem there should be standard definitions of how to generate and store information, as the use of standard data formats ensures interoperability. In Macharis & Crompvoets (2012) the degree of data standardization was identified as an important criterion for the success of a Spatial Data Infrastructure, and also Barthélemy (2016), Dawes, Vidiasova & Parkhimovich (2016) and Wiener et al. (2016) stresses the importance of standards. In the study of Harrison, Pardo & Cook (2012) on the questions essential to the development of the development of an Open Government Ecosystem, they found that participants of their workshop called for new research on the development and adoption of standards, "noting a fundamental disconnect between the recognition that standards are key to scale and use of data" (p. 920).

3.4.2. CATEGORY 2 - DATA GOVERNANCE

So far, the success criteria for data ecosystems related to data supply. The second output indicator of Welle Donker & Van Loenen (2017) used to categorize success criteria resulting from literature is 'data governance'. They consider governance to be the framework of policies, processes and instruments to realize common goals in the interaction between public and/or private sector bodies. Data governance will help to facilitate the data supply model of Backx (2003), as visualized in *Figure 11*, and to assist the user when having issues with the data supply. Welle Donker& Van Loenen (2017) adapted the five elements for assessing the governance of geographical information infrastructures of Kok & Van Loenen (2005) to apply it to the data ecosystems governance context: vision, leadership & control, communication & stimulation, self-organizing capacity and financing (see *Figure 12* below).

FIGURE 12 – Aspects of the governance of a data infrastructure (Kok & Van Loenen, 2005)

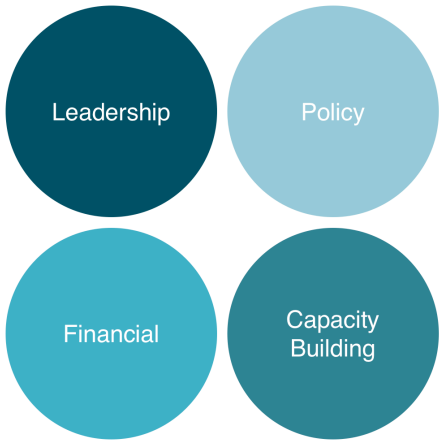


FIGURE 13 – Open Data Support Mechanisms for Public Bodies (Lee, 2014)

Also Harrison, Pardo & Cook (2012), Lee (2014) and Wiener et al. (2016) acknowledge the importance of a well-developed data governance model in an ecosystem, for example to be able to connect with potential users, to identify which data to publish and to maintain the data. Lee (2014) identifies four 'Open Data Support Mechanisms for Public Bodies' (*Figure 13*), which are similar to the aspects of the governance of a data infrastructure as shown in *Figure 12*. In *Figure 13*, the 'Communication & stimulation' and 'Vision' element are merged together in the 'Policy' element, which encompasses the outline of "a government's open data commitments, goals and principles" (Lee, 2014: p.257).
The governance aspects as visualized in these figures (*12* and *13*) are found to be backed up a lot in other literature. This systematic literature review resulted in a slight adjustment of the aspect names, to also have them cover aspects coming forward in other articles.

Element 2.1 - Vision, communication and stimulation

Welle Donker & Van Loenen (2017) describe the 'Vision' aspect as "to provide a common goal, to avoid a fragmented approach and to stimulate cooperation between stakeholders" (p. 296). Establishing cooperation and the need for a collaborative environment in which interaction takes place appears to be an important theme in the creation of a data ecosystem in general, that is also put forward by other scholars, for example by Barthélemy (2016), Attard, Orlandi & Auer (2016), Dawes, Vidiasova & Parkhimovich (2016), Immonen, Palviainen & Ovaska (2014), Parsons et al. (2011) and Wiener et al. (2016). Yet, this seems rather logical, as these are characteristics inherent to what comprises an ecosystem. The 'vision' aspect is taken together with 'communication and stimulation', as this 'vision' definition of Welle Donker & Van Loenen (2017) as presented closely relates to 'communication and stimulation', which also concerns the cooperation between stakeholders.

Element 2.1.1 - Division of roles and responsibilities

A more explicit ecosystems criterion that appears in literature related to this 'stimulation of cooperation between stakeholders' is the need to clearly assign different roles and responsibilities in the ecosystem. To do so, first all actors in the ecosystem should be identified (Barthélemy, 2016; Dawes, Vidiasova & Parkhimovich, 2016; Immonen, Palviainen & Ovaska, 2014), and then the relative position of the actors in the ecosystem can be located (Schalkwyk,

Willmers & McNaughton, 2016). Zuiderwijk, Janssen & Davis (2014) argue that in an open data ecosystem, persons and organizations can have multiple roles, and Dawes, Vidiasova & Parkhimovich (2016) also recognize this. Additionally, Dawes, Vidiasova & Parkhimovich (2016) mention the need for accountability mechanisms in a data ecosystem, as part of the ecosystem policies and strategies.
According to Immonen, Palviainen & Ovaska (2014), 'division of work' is one of the criteria of a well-functioning data ecosystem; "each actor must have a natural role in the ecosystem" (p.91). Also in Harrison, Pardo & Cook (2012) it was found that in an ecosystem there is a need to decide which parties have which responsibilities, like for example creating the technical architecture infrastructure and implementing the required policies. Furthermore, Davies (2012) and Lee (2014) mention the division of responsibilities and roles in an ecosystem as an important factor.

Element 2.1.2 - Feedback

Another important ecosystem element related to the stimulation of cooperation between stakeholders coming forward in literature is the importance of the presence of feedback mechanisms to enable the provision of feedback from data users to data providers. Pollock (2011) already found that data cycles with feedback loops are an absolute necessity in an open data ecosystem. In Zuiderwijk, Janssen & Davis (2014), "interpreting and discussing data and providing feedback to the data provider and other stakeholders" (p.30) is identified as one of the four key elements that should be captured to create an open data ecosystem. Also in Dawes, Vidiasova & Parkhimovich (2016) 'feedback and communication' is an essential element in their ecosystem model of open government data programs. About the ways to realize this, they state that "government leaders and administrative agencies can implement different ways to acquire input and feedback about the data and its use including not only discussion forums and feedback request forms but also consultations, meetings, workshops and other virtual and face-to-face discussions" (p.24). In Immonen, Palviainen & Ovaska (2014) 'co-creation', including the need for feedback, is found to be one of the elements essential for an open business data ecosystem. They also acknowledge that users can contribute to improving the correctness and quality of data by providing feedback to data providers. Furthermore, Barthélemy (2016), Harrison, Pardo & Cook (2012) and Lee (2014) recognize the importance of feedback on the data published in a data ecosystem.
The inclusion of feedback mechanisms in a data ecosystem also contributes to meeting the user's needs (Barthélemy (2016), as elaborated on in *element 3.2* in this section.

Element 2.2 - Leadership and incentivization

'Leadership and control' involves the need to have a problem owner in an ecosystem who coordinates and stimulates open data activities (Welle Donker & Van Loenen, 2017) - a governance element also identified by Lee (2014). This includes the creation of awareness and capacity-building (elaborated on later), which may lead to work floor support and political support for open data – an important success factor identified by Craglia et al. (2002). This is in line with Immonen, Palviainen & Ovaska (2014), who claim that "the ecosystem has to enable different open data providers to the ecosystem from heterogeneous data sources" (p.92). Dawes, Vidiasova & Parkhimovich (2016), found that governments can play an

important role in stimulating and supporting data producers and users in an ecosystem, for example by encouraging data publication and use.
Harrison, Pardo & Cook (2012) mention an aspect related to this, namely that key actors should be incentivized to take up an active role in the data ecosystem and to share data. Wiener et al. (2016) also recognize this; they identify the importance of 'incentives' as one of the factors that need to be addressed when initiation a data ecosystem and enabling its survival. According to them, data providers need to be motivated to share data. They suggest that this can be done by the community, who can for example "adopt a metric to assess the use of shared data" (p.618) and encourage collaborative efforts so data sharers feel they are intrinsically rewarded, or by "lowering the barrier to entry for data-sharing" (p.618) as this is vital to participation.
The importance of incentivization to share data is also acknowledged Barthélemy (2016), who assigns this task to the top management. Davies (2012) however, does not share this view. He argues that the management strategies that open data ecosystems adopt should be based on the actors operating as equal, networked partners rather than taking a top-down approach. The explanation Welle Donker & Van Loenen (2017) give for the existence of these different perspectives is that open data are promoted for different reasons. As a consequence, there is a lot variation in the perception of which organization provides leadership and how this is done. Leadership in this context should therefore not necessarily be interpreted as a 'position' reserved for one specific party in the ecosystem; the 'leadership' tasks (to stimulate, incentivize and coordinate) could also be carried out by a group of parties, or realized using a bottom-up approach, creating 'community responsibility' (Barthélemy, 2016).

Element 2.3. Self-organizing capacity

The governance element of 'self-organizing capacity' concerns the way in which data supply matches demand, an aspect that also Attard, Orlandi & Auer (2016) focus on by proposing a 'Demand and Supply Distribution Model', which provides a starting point for stakeholders to create value out of data. This element partly overlaps with the previous 'leadership' element, in the sense that to achieve this, open data activities need to be stimulated. Lee (2014) focuses mainly on capacity building for actors engaging in the open data ecosystem. Public bodies for example, require certain capacities to be able to participate in the ecosystem, such as technical knowledge on the systems and technologies involved, data management knowledge on how to ensure high data quality, and operational knowledge on how to incorporate the open data activities into their current practices (Lee, 2014). This is also acknowledged by Dawes, Vidiasova & Parkhimovich (2016) and Wiener et al. (2016); the latter additionally stresses the importance of data providers being able to meet the data demands.

Element 2.4 - Sustainability: financing and value creation

Heimstädt, Saunderson & Heath (2014) have identified 'sustainability' as one of the structural properties of an ecosystem.
In literature, it was found that many scholars, among which Lee (2014), Welle Donker & Van Loenen (2017) and Wiener et al. (2016), mention 'sustainable financing' an important factor in achieving this ecosystems sustainability. According to Ponte (2015), one of the main issues to address when enabling an open data ecosystem is the sustainability over time; he also states that this can be realized by achieving economic sustainability. Parsons et al. (2011) agree,

saying that a sustainable business model is 'critical' to "facilitate a socio-technical evolution in the overall science data ecosystem" (p.558) they analyze, which stresses the importance of including sponsors and funding agencies.
Besides financing, 'value creation' comes forward as an important factor in the achievement of a sustainable data ecosystem. Van Schalkwyk, Willmers & McNaughton (2016) found that data ecosystems appear to be more stable and sustainable when users perceive value in the provision of data. In the study of Harrison, Pardo & Cook (2012), 'valued outcomes' emerged as an essential dimension of an open government ecosystem, as it can serve as a basis for ecosystem sustainability; the sustainability of an open government ecosystem relies on the ability to explain the value generated by the data-sharing to all relevant stakeholders. They hence imply that value perception is something that actors in an ecosystem can influence themselves.
Heimstädt, Saunderson & Heath (2014) argue that long-term sustainability in an ecosystem can only be realized when data suppliers recognize a significant benefit, and also Davies (2012) and Jetzek (2017) stress the importance of value generation and perceived value for ecosystem sustainability.

3.4.3. CATEGORY 3 - USER CHARACTERISTICS

To derive actual value out of the publication of data, user characteristics must be considered when establishing an open data ecosystem (Dawes, Vidiasova & Parkhimovich, 2016; Lee, 2014 and Welle Donker & Van Loenen, 2017). The ecosystems assessment framework of Welle Donker & Van Loenen (2017) distinguishes itself by not only focusing on open data readiness (through the analysis of 'data supply' and 'data governance'), but also including the user's perspective. This is important to engage data users and to stimulate open data re-use (Lee, 2014). Harrison, Pardo & Cook (2012) state that "at the heart of the open government ecosystem is the assumption that government possesses information that users want and will use" (p.917). They identify the user aspect, which they describe as "cultivating new engagements between government and users of all types (citizens, businesses, civil society) to assess users' expectations and requirements for information and data on the one hand, and users' abilities to consume and create information and data on the other hand" (p.922), as an essential dimension of an open government ecosystem.

Element 3.1 - User capabilities

As described by Harrison, Pardo & Cook (2012), and also stressed by Lee (2014), to ensure that users *will* use data available and accessible in an ecosystem, it is important to take into account potential user capabilities. Once users have access to the data, they might need a certain level of technical knowhow to be able to retrieve it. Also, they might need specific skills to transform the data, or they simply are unfamiliar with the opportunities (Welle Donker & Van Loenen, 2017). According to Harrison, Pardo & Cook (2012), strategic ecosystem thinking comprises that governments engage with users to assess their "capabilities for information consumption and creation, and explore the circumstances in which data is best used" (p.917). Also Dawes, Vidiasova & Parkhimovich (2016) identified the capabilities of users in an open data ecosystem as an essential element.
Barthélemy (2016) additionally found that "it should be easier for re-users to contribute to datasets released, for instance by helping to update the data" (p.33), which is in line with the

call of Lee (2014) to “encourage data users to take ownership as active participants as opposed to passive data recipients” (p.357).
A proper and user-friendly data presentation and visualization could help to respond to limited user capabilities (Davies, 2012 and Harrison, Pardo & Cook, 2012).

Element 3.2 - User needs

To ensure that users *want to* use data available and accessible in an ecosystem, diverse user needs should be considered. Because what do different types of users want? What are their interests and expectations (Lee, 2014)? Interaction between data providers and data users can help answer these questions (Dawes, Vidiasova & Parkhimovich, 2016 and Harrison, Pardo & Cook, 2012). In this regard, Dawes, Vidiasova & Parkhimovich (2016) state that when users are consulted about their specific goals for data use, this “can stimulate more targeted applications (of the open data, red.) that may have lasting value because they address real community needs” (p.24).
When assessing the user needs, different user types should be identified and addressed, as there is a great diversion in user needs - there is not one single user goal (Harrison, Pardo & Cook, 2012; Immonen, Palviainen & Ovaska, 2014 and Welle Donker & Van Loenen, 2017).

3.4.4. CATEGORY A - EXTERNAL CONTEXT

One last aspect related to data ecosystems coming forward in literature is the importance of considering the external context of an ecosystem. This is not exactly a ‘criterion’ for a successful data ecosystem; it is merely an overarching factor that should be considered when analyzing a data ecosystem. This aspect is therefore not numbered, but forms a separate aspect category *A,* that does not relate to any specific criteria (*see Table 2*).
The importance of considering the ecosystem context is mentioned quite frequently in literature, either explicitly or implicitly. Authors in general state that the functioning of a data ecosystem also depends on the ecosystem context, like for example the local culture, the political system and historical influence (Dawes, Vidiasova & Parkhimovich, 2016; Heimstädt, Saunderson & Heath, 2014). These aspects all together form the institutional conditions in which the ecosystem is or needs to be embedded, that influence how actors in the ecosystem function and how different ecosystem elements are arranged (Van Schalkwyk, Willmers & McNaughton, 2016). Related to this, is the emphasis of Parsons et al. (2011) on the need to taking an organic approach: to optimize a data ecosystem, the system must be understood in totality rather than just looking at the individual components. A data ecosystem must be flexible and adaptable to be able to deal with contextual dynamics (Parsons et al., 2011).
Other authors mentioning the importance of considering the ecosystems context are Barthélemy (2016), Harrison, Pardo & Cook (2012), Macharis & Crompvoets (2012) and Zuiderwijk, Janssen & Davis (2014).

3.4.5. VISUALIZATION OF FINDINGS

The initial findings from the systematic literature review on the criteria for a successful data ecosystem as summarized in at the beginning of this section in *Table 2* and further elaborated on in the previous paragraphs are visualized in *Figure 14* below.

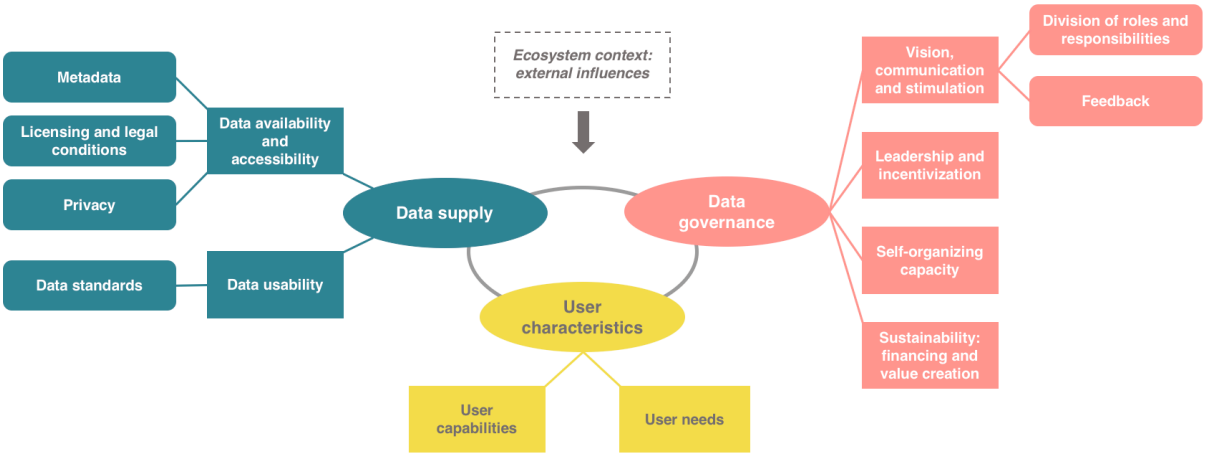


FIGURE 14 – Visualization of criteria for a successful data ecosystem found in the systematic literature review

3.5. SYSTEMATIC LITERATURE REVIEW EVALUATION: EXPERT VALIDATION

3.5.1. APPROACH

The outcomes of the systematic literature review have been validated in two different ways.

In the first place, by means of an internal feedback session with the 510 Global team on 22-05-2017. During this feedback session of one hour, the context and outcomes of the systematic literature review were presented to six of the team members, among which staff with practical experience in the collection of humanitarian data. After the presentation of 15 minutes, there was room for an open discussion. Attendants were first asked for feedback on the individual ecosystem elements: did they agree with the content? Did they think other elements should be included? During this part of the discussion it was questioned whether different data collection methodologies were accounted for in the framework. It was suggested to include a separate 'data collection' element at the data supply part of the framework.

In the second part of the discussion, the application of this framework in the humanitarian context was discussed. It was mainly pointed out that it is important to identify the actors taking the lead in the development of this framework, and that the humanitarian data ecosystem is still very much in its infancy. This will be further elaborated on in *Chapter 4.*

To be able to receive more content-driven, theoretical feedback on the data ecosystems framework derived from the systematic literature review, a second validation session took place at the Delft University of Technology. On 23-05-2017, the review outcomes were presented at an (open) data meeting of researchers and professors involved in (open) data research of two different faculties of the Delft University of Technology (the Faculty of Technology, Policy and Management and the Faculty of Architecture and the Built Environment). 17 researchers from various disciplines attended:

Faculty of Architecture

- Bastiaan van Loenen
- Frederika Welle Donker
- Hendrik Ploeger
- Lorenzo Dalla Corte
- Dirk Dubbeling
- Agung Indrajit
- Glenn Vancauwenberghe

Students

- Elise Haak (SEPAM)
- Stephan Kool (SEPAM)

Faculty of TPM

- Darus Salam
- Ahmad Luthfi
- Arie Purwanto
- Ricardo Matheus
- Dhata Praditya
- Jolien Ubacht
- Ben Zhu
- Boriana Rukanova
- Anneke Zuiderwijk - van Eijk
- Iryna Susha (after the meeting)

Among the attendants were Frederika Welle Donker and Bastiaan van Loenen, the authors of the article that was used to structure the systematic literature review (Welle Donker & Van Loenen, 2017).
The aim of the presentation was to evaluate the review outcomes and to receive expert feedback, both on the data ecosystems framework and on the completion of the literature list. Prior to the presentation, all attendants were provided with a hand-out of the overview of the literature used (*Table 1*) and a hand-out of the list of outcomes of the literature review (*Table 2*), so they could read along and could go through the detailed descriptions of the found ecosystem elements and the literature list. The presentation slides can be found in *Appendix B.*
After the 15-minute long presentation, there was room for an open discussion of 25 minutes. The following questions were asked:

1. Did the researcher miss other relevant literature on the criteria for successful data ecosystems?
2. Do the attendants think this overview of elements (*Figure 14*) is complete, or are certain aspects missing or overlapping?

The attendants did not have suggestions for further relevant literature. It was therefore expected that the literature used was adequate.
To structure the discussion on the completeness and content of the elements found in literature, the researcher first presented the aspects she found remarkable herself to the attendants, to discuss those. Subsequently, everyone was free to provide additional feedback in a group discussion. Everyone got the opportunity to contribute – there was sufficient time for everyone to pose the questions they had and raise points for discussion.

3.5.2. INITIAL REMARKS OF THE RESEARCHER

Remark 1 - Privacy is mentioned limitedly

Privacy was only mentioned in four of the 17 articles reviewed. However, it was expected to be considered more important in the context of data ecosystems, as in literature on open data issues related to privacy generally pop up. Frederika Welle Donker attributed this to the fact that most of the reviewed literature concerns open data, and that in their article (Welle Donker

& Van Loenen, 2017), they stated in the introduction that they would not consider privacy since they were dealing with open data and assumed that for open data there would be no privacy issues. A discussion unfolded of which the main conclusion was that the presence of privacy issues in a data ecosystem is a typical example of an ecosystem element that is dependent on the external context: it depends on the ecosystems context whether privacy of the data isconsidered to be an issue.

Remark 2 - Importance of trust or transparency not mentioned at all
Based on previous readings on data-sharing in the humanitarian sector, it was found remarkable that in none of the articles the importance of trust between data providers and data users was mentioned. In the humanitarian sector, it is essential for data providers to trust data users, as this affects their willingness to share their data, and vice versa: data users only want to use the data if it comes from a trusted party that is transparent about where they got their data from. This will be further discussed in Chapter 4, when looking at the humanitarian data ecosystem specifically, but it is conceivable that this also holds for sectors other than the humanitarian sector. This was indeed confirmed by open data researcher Anneke Zuiderwijk – van Eijk, who mentioned that she thought a 'trust' element was lacking in the systematic literature review outcomes even before these doubts were presented to her. Other attendants agreed. It is therefore suggested to include 'trust and transparency' element in the 'data governance' category. This remark did not result in any adjustments of the framework.

Remark 3 - Importance of data collection neglected
The goal of the systematic literature review is to create an overview of critical elements of data ecosystems. As mentioned before, most articles that were reviewed concern *open* data ecosystems, and this could be the reason why the importance of active, manual data collection - required to retrieve *closed* data - is neglected in this literature. When open data is made accessible this automatically means it can be retrieved, but for closed data, it must be possible to collect it to become accessible – an aspect also pointed out during the feedback session at the Netherlands Red Cross. This was acknowledged by the attendants. Hence, to account for various degrees of openness of data in data ecosystems, attention should be paid to possibilities for data collection, which is why this element should be included as part of 'data availability and accessibility'.

Remark 4 - Limited focus on user characteristics
When comparing it to the presence of data supply and data governance elements in literature, the importance of considering user characteristics in a data ecosystem comes forward relatively limitedly. On top of that, in most of the articles in which it is mentioned, it does not have a prominent role. This is in line with literature; Dawes, Vidiasova & Parkhimovich (2016) for example state that it is problematic that open data initiatives are generally biased toward the supply part – that they lack sufficient attention to the user perspective. This is acknowledged by the attendants. Especially Frederika Welle Donker and Bastiaan van Loenen stressed that there is very little knowledge yet on user characteristics. It is very important to get a better idea of the user capabilities and needs in a data ecosystem, as there is often a mismatch between data supply and governance, and the user characteristics. A clear overview of how this problem can be solved is currently lacking. It would therefore be interesting to map this for the humanitarian sector.

3.5.3. ADDITIONAL COMMENTS OF PARTICIPANTS

Additional comments brought forward by the attendants are listed below. These comments mainly concern the visualization of (and relations between) the elements as shown in *Figure 14* and more fully discussed in *Section 3.6.4.*

- Metadata and data standards are interrelated, as there are many metadata standards as well, which are important for both data availability and accessibility, and data usability. It is suggested to visualize this by creating a dotted line between the two elements.
- Data quality seems to be lacking. This is not the case - it is part of the definition of 'data usability' ("High quality of the data shared in a data ecosystem should be ensured to enhance its usability") but it is understandable this does not directly become clear when looking at the visualization. It was therefore decided to explicitly mention data quality as an aspect of data usability in the framework.
- The price that needs to be paid to access the data at the supply part is lacking: data should be affordable. According to the attendants, this is not included in the 'sustainable financing' aspect of the data governance. It is specifically related to the data availability and accessibility, and should therefore be included as a separate element. A reassessment of the reviewed literature showed that this idea is backed up by Welle Donker & Van Loenen (2017), and also mentioned by Heimstädt, Saunderson & Heath (2014), who refer to the 'usage cost' of data, Immonen, Palviainen & Ovaska (2014), who discuss different pricing models ("there could be a small fee for the used data" (p.95)), Jetzek (2017, mentioning 'prices to usability' of data, and Macharis & Crompvoets (2012), who refer to 'the price to purchase data'.
- The term 'data governance' is inadequate, as it is not just the data that needs to be governed, but the entire ecosystem (e.g. also the actors operating in it). It would be better to change it to the more general 'governance'.
- For the overview, financing and value creation could be visualized as the two (light orange) sub-elements of sustainability. Then the framework allows for the two aspects to be assessed individually, which makes sense as they represent different things.
- The term 'self-organizing capacity' does not seem to clearly reflect the content of this element on the governance part of the framework. The element entails that data supply should match data demand in an ecosystem and that data providers require certain capacities to be able to make this happen, but the term 'self-organizing' suggests something different. It is therefore suggested to change this name into something that better reflects the content, like 'participatory capacity'.

3.5.4. CONCLUSIONS OF EXPERT VALIDATION

Overall it can be concluded that the evaluation session was very useful, and that it resulted in a successful expert validation of the framework of elements for a successful data ecosystem. In general, the outcomes of the literature review were considered to be rather complete, but some additions and adjustments were suggested. This led to the following changes to the initial set of elements visualized in *Figure 14:*

- The addition of the 'Data collection' element
- The addition of the 'Affordability' element
- The addition of the 'Trust and transparency' element

- The separation of ‘Data quality’ as an individual element. It was already part of data usability, but is now shown as a separate element to enhance clarity.
- The creation of a dotted line between ‘metadata’ and ‘data standards’ to show they are interrelated
- The change of the ‘data governance’ category into the more general ‘governance’ category, since this better covers its load
- The separation of ‘sustainable financing’ and ‘value creation’ as two sub-elements of sustainability.
- The name-change of the element ‘Self-organizing capacity’ into ‘Participatory capacity’, aiming to reflect that ecosystem actors other than the data users also require certain capacities to be able to effectively participate in the ecosystem. The term ‘capacity’ is still considered to be accurate. A definition of capacity of for example Chaumba & Van Geene (2003) is: “Capacity is the power of something (a system, an organisation, a person) to perform or to produce. It may be defined as the ability of individuals and organisations to perform functions effectively, efficiently, and sustainably” (p. 69).

In the next section, the final framework of elements for a successful data ecosystem, in which these expert comments are incorporated, will be presented.

3.6. Chapter conclusion: Final framework of criteria for a successful data ecosystem

This chapter aimed to answer the following sub-question:

SQ2: Within the field of data ecosystems, which framework can be used for the creation of a generic governance structure for the toolbox?

To do so, 17 relevant articles on data ecosystems were systematically reviewed to create an overview of success criteria for a successful data ecosystem. Both the selection of relevant articles and the overview of success criteria were evaluated with a group of 17 data experts of two faculties of the Delft University of Technology, which resulted in the addition of a few elements and some minor adjustments.
In *Figure 15* the final framework of criteria for a successful data ecosystem is visualized. The elements that have been adjusted are numbered:

1 = Element added after the expert validation
2 = Element was already implicitly present in the initial framework, but now included as an individual element for clarity’s sake
3 = Renamed element

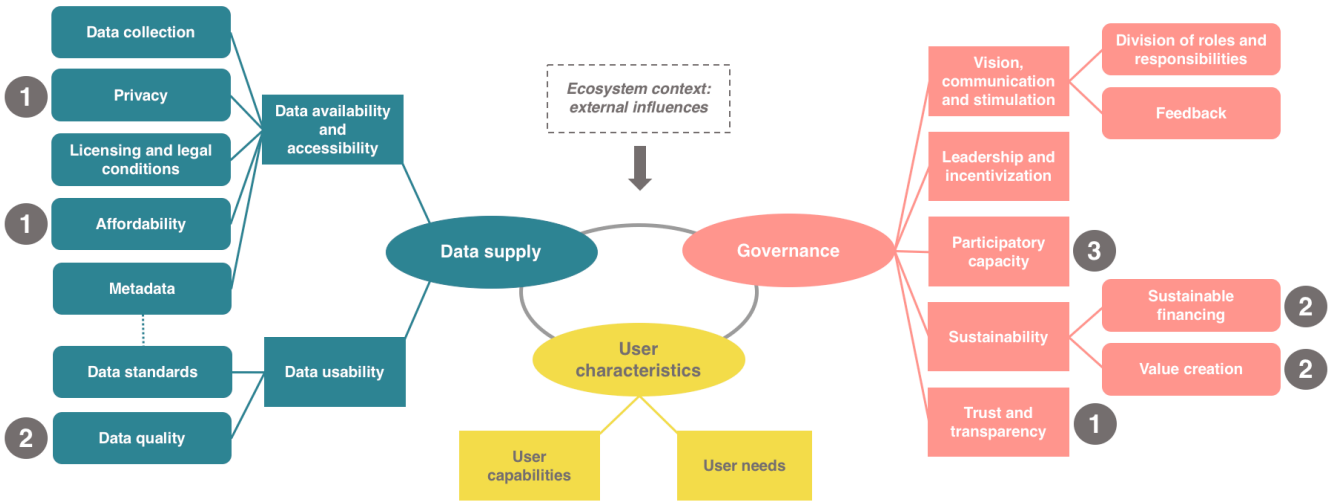


FIGURE 15 – Final framework of criteria for a successful data ecosystem – findings from systematic literature review and expert validation combined

CHAPTER 4 – THE HUMANITARIAN DATA ECOSYSTEM

4.1. INTRODUCTION

As described in *Chapter* 3, the ecosystems perspective can be a means to assess existing conditions in a system, like for example strengths and weaknesses of different components and relationships, and by doing so, the perspective can help designers to get an idea of which changes and investments in strategies, policies, practices, participants or relationships are necessary to improve those existing conditions. These are the insights that the framework of criteria for a successful data ecosystem as developed in the previous chapter should provide for the data ecosystem around the Community Risk Assessment and Prioritization toolbox.
In this chapter, it should become clear how this can be done: how can the framework of criteria for a successful data ecosystem be applied in the context of the problem definition – scaling up the toolbox? How can the framework be made operational? That is the question that will be answered in this chapter:

SQ3: HOW CAN THIS DATA ECOSYSTEMS FRAMEWORK BE APPLIED TO THE CONTEXT OF SCALING UP THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX?

a. *Which criteria for a successful data ecosystem as derived from literature are relevant in the context of the Community Risk Assessment and Prioritization toolbox?*
b. *How can these criteria be made operational, so they can become of use in the process of scaling up the Community Risk Assessment and Prioritization toolbox?*
c. *How should these operationalized criteria be arranged in a country, for the country to be eligible for application of the Community Risk Assessment and Prioritization toolbox?*

It is unclear yet whether all the success criteria that resulted from the systematic literature review, in which data ecosystems in general were considered, are relevant for the humanitarian data ecosystem specifically, and whether they are all equally important. It could for example be possible that in the context around scaling up the Community Risk Assessment and Prioritization toolbox, some of the criteria require more immediate attention than others. This is why sub-question 3a is important to answer, before continuing to the operationalization of the criteria. This sub-question will be answered by explaining the development of the emerging data ecosystem around the Community Risk Assessment and Prioritization toolbox (*Section 4.2*) This is done by means of a desk research combined with expert input – the relevance of the different framework criteria in the context of the Community Risk Assessment

and Prioritization toolbox have been discussed with the 510 Global team in an internal feedback session.
Practical experience from the development of the toolbox for the Philippines, Malawi and Nepal will subsequently serve as an input to answering sub-question 3b and c, as it is expected that this will provide the required operational insights (*Section 4.3*). The outcome of this research step will be a list of measurable conditions per success criteria relevant in the context of scaling up the Community Risk Assessment and Prioritization toolbox. Assessing the state, or 'maturity' of these conditions in a country can provide a scan of the situation in a country in terms of its eligibility for the application of the Community Risk Assessment and Prioritization toolbox. This scan helps to identify the factors to be addressed when trying to improve this eligibility: the aspects the governance should focus on (*Chapter 5*).

4.2. THE DATA ECOSYSTEM AROUND THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX

In this section, the question which data ecosystem success criteria (or elements) as derived from literature are relevant in the humanitarian data ecosystem around the Community Risk Assessment and Prioritization toolbox specifically will be addressed. The outcome of this section is a set of selected framework elements that need to be made operational – which will be done in *Section 4.*3 – to create insight in how they can be positively influenced to enhance a country's humanitarian data ecosystem.

4.2.1. AN EMERGING DATA ECOSYSTEM

In the feedback session with a group of team members from 510 Global, as described in *Section 3.5.1,* not only was the theoretical framework evaluated and were no additional elements identified, also its relevance for application in the humanitarian sector was discussed. Something that all present team members agreed on, and of which the lack of existing literature on the topic could be an indication, is that the humanitarian data ecosystem is still very much in its infancy. It is emerging, but immature. Where the literature on data ecosystems focuses on efforts to open up existing data, in the humanitarian data ecosystem much of the existing data is yet to be identified, and stored in a way that allows for sharing it on an open data platform. This has implications for the immediate relevance and presence of some of the data ecosystem success criteria identified in literature.
In *Figure 16* on the next page this is visualized for the data ecosystem around the Community Risk Assessment and Prioritization toolbox. The development of the toolbox knows different phases, in which different success criteria require attention. The elements that are grey-colored are not focused on in that ecosystems phase. Eventually all criteria will need to be considered to establish a successful and sustainable humanitarian data ecosystem, but as can be seen, during the initial development of the toolbox, there is a strong focus on a selective set of elements. This will be described in the next paragraphs, per category of the framework.

4.2.2. DATA SUPPLY IN THE EMERGING DATA ECOSYSTEM

In the first phase of the toolbox development (see *Figure 16*), there was a strong focus on figuring out which data to collect, of (sufficiently) high quality and at low cost. This trend continues in the current situation. To get the model running, the first concern is to collect all the required data. As described before, this is a very challenging process in the humanitarian

sector, due to the data being scattered across many organizations and sectors, and, especially at lower administrative levels, often being stored only on a local hard drive, or even paper-based. Concerns about data privacy, licensing and legal conditions, meta data and data standards are also highly relevant, but they are of a different order: their importance increases as the ecosystem evolves and more data-sharing starts to take place. Their importance stems from the elements that determine the level of data collection that can take place ('data collection', 'affordability' and 'data quality'): in a country for which the toolbox has not been developed yet, but you want to start developing the toolbox, 'collection', 'affordability' and 'data quality' *need* to be in place. They are critical for a country to be eligible for the toolbox application in the first place. The remaining *Data supply* elements only become relevant in later stages of the toolbox development; they are hence to be considered of secondary concern.

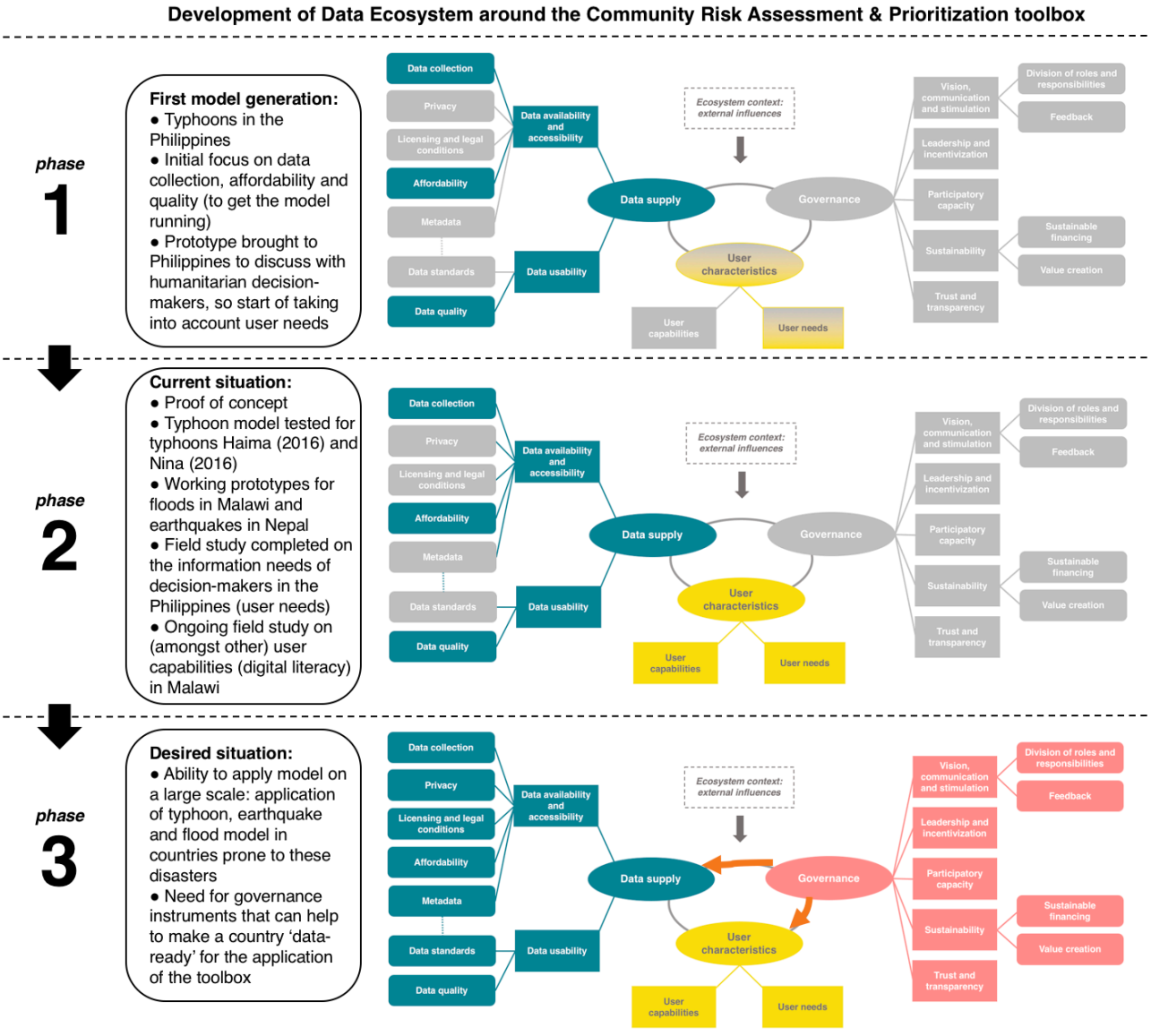


FIGURE 16 – Development of data ecosystem around the Community Risk Assessment & Prioritization toolbox

4.2.3. USER CHARACTERISTICS IN THE EMERGING DATA ECOSYSTEM

In the first phase of the data ecosystem around the Community Risk Assessment and Prioritization toolbox, the user characteristics are considered partially. The indicators that serve

as an input for the toolbox are based on the three dimensions of INFORM. According to INFORM, “each dimension encompasses different categories, which are user-driven concepts related to the needs of humanitarian and resilience actors” (INFORM, 2016). This was confirmed by an INFORM Disaster Risk Consultant (Interviewee 1, personal communication, June 28, 2017). This means that the basis of the Community Risk Assessment relies on user needs.

In the current situation (*phase 2*), the role of the user characteristics has increased. This is important, as both user characteristics have a large influence on the impact of the Community Risk Assessment and Prioritization toolbox. When scaling up the toolbox without matching the user capabilities or information needs of humanitarian decision-makers, it will not achieve the desired effect. It is important that aid workers have the capabilities to integrate the Community Risk Assessment & Prioritization toolbox predictions in their humanitarian decision-making, and that the toolbox generates information that matches the information needs of the aid worker. A challenge here is to make it clear to aid workers how they can benefit from the toolbox output. Aid workers with years of experience in one specific region tend to want to rely on their own expert judgments rather than on (to them) ‘external’ data. Their judgements might be right, but it is difficult to objectivize, scale and combine this information.

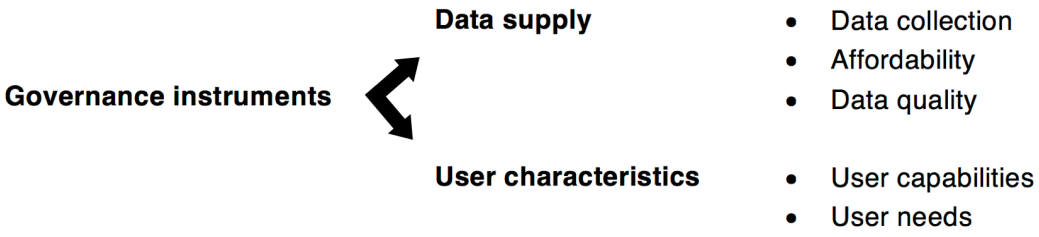
4.2.4. GOVERNANCE IN THE EMERGING DATA ECOSYSTEM

As already identified in *Section 1.2 – Problem definition,* an issue in the process of trying to scale up the Community Risk Assessment and Prioritization toolbox, is that a governance-structure is lacking. In the context of the toolbox, there are no – or very few – governance structures in place yet that can stimulate the data ecosystem development; the governance part is underdeveloped and many organizations only pursue their own objectives. As Sabou (2016) describes, “the existing humanitarian – ‘digital data’ – ecosystem is essentially a collection of ungoverned pilot programs”. Within many organizations involved in the application of the toolbox, the internal governance is not even in place. For example, Red Cross societies often do not have their internal data-sharing (for example between different projects and departments) sorted out.

This means that to realize further developments of the humanitarian data ecosystem, governance processes in line with the elements on the governance part of the data ecosystems framework should be introduced. This governance should affect the Supply part and User characteristics elements in such a way that they will enable a scale up of the Community Risk Assessment and Prioritization toolbox: the transition from *phase 2* to *phase 3* in *Figure 16.*

4.2.5. SET OF FRAMEWORK ELEMENTS TO BE OPERATIONALIZED

To realize this transition and be able to apply the toolbox on a larger scale, these governance processes to be introduced should contribute to the generation of an optimal context for the application of the Community Risk Assessment & Prioritization toolbox in a country. To do so, they should initially address the elements that are of primary concern; the elements that *need* to be in place to be able to develop a functioning Community Risk Assessment and Prioritization toolbox for a country:



These are therefore the elements that will be operationalized in the next section, to it becomes clear how to influence them. As explained, the remaining 'Data supply' elements are not irrelevant, however they do not have immediate priority: their relevance increases as the ecosystem evolves and more data-sharing starts to take place. Additionally, another team of 510 Global is currently working on a long-term strategy to deal with 'data responsibility' – issues around privacy, licensing and legal conditions, which another reason to leave these elements out of the scope of this research.

4.3. Operationalization of relevant ecosystem criteria

In the previous section, it was explained that to be able to develop the Community Risk Assessment and Prioritization toolbox for a country, five basic data ecosystem elements need to be in place. If they are not (or insufficiently) in place, it should be possible to deploy a certain governance approach to positively influence these elements, or form a favorable context for these elements to be developed. But to get an idea of how exactly to influence these elements and of the type of governance structure that needs to be developed, it first must become clear how these elements can be 'measured': they must be operationalized. Because how can it be assessed whether these elements are present in a country – whether a country is 'data ready' for the toolbox application? This is the question that will be answered in this section, by developing an assessment tool that can help to determine whether a country is eligible for the toolbox application, and if not, which elements need further attention.
To develop this tool, several steps need to be completed:

1. In the first place, the elements need to be transformed into measurable conditions that a country should meet to be eligible for the toolbox application (*Section 4.3.1*).
2. In the second place, the metrics for these conditions that can help to determine to what extent a condition is fulfilled need to be defined: which metrics can be used to determine the maturity level of a country per condition, and how to specify those different maturity levels? (*Section 4.3.1*).
3. In the third place, the threshold values per condition need to be determined: to what extent should a condition be fulfilled for a country to be eligible for the toolbox application? Which maturity level needs to be present? (*Section 4.3.2*)

Guided by the theoretical description of the elements, the defined conditions in the first step followed from the functioning of the toolbox (*Chapter 2*), and have been listed in consultation with the toolbox' head developer, who had the lead in developing the toolbox for the Philippines, Malawi and Nepal. A desk research combined with expert input from 510 team members involved in the development of the Community Risk Assessment and Prioritization toolbox led to the completion of step 2 and 3.

4.3.1. CONDITIONS TO BE FULFILLED AND RELATED METRICS

From the key theoretical framework elements that need to be in place to develop a functioning Community Risk Assessment and Prioritization toolbox for a country, a set of conditions has been defined. These are the conditions need to be fulfilled, at least to a certain extent, to be able to apply the Community Risk Assessment and Prioritization toolbox in a given country. The conditions and the corresponding metrics, which can be found in *Table 3* and *Table 4,* are further explained and elaborated on later in this section.

As can be seen, the data collection conditions are grouped per data type as explained in *Section 2.3.1 - Rapid execution and timely collection of* data (pre-disaster and post-disaster data), as each data type requires different conditions to be fulfilled. The 'affordability' condition holds for any type of data, and the 'data quality' conditions are specified again for the pre-disaster data, as there is a clear quality demarcation for this data and the collection of this data is a long-term process with room for improvement. As for the post-disaster data, there is more of a culture of 'work with what you have' due to severe time pressure.

Category		Element	Condition	Metric	Clarification
Data supply	Data availability & accessibility	***Data collection***	• *Preparedness phase - Community Risk Assessment (pre-disaster)*		
			1. It should be possible to get an overview of data providers and available data sets	Scale: Does this overview exist yes/partly/no	This overview should include the data providers in a country, the data they hold and if it can be shared
			2. It should be possible to access / collect the data	Scale: Rating of the country on the key open data indices (ODB, GODI, ODIN, OECD Index), combined with expert assessment	Gives an indication of the effort needed to collect the data: how can the available data be accessed?
			3. The gap between data needed and data available should be sufficiently small (*Completeness - DSPI*)	DSPI: Completeness component	DSPI = Completeness * Recency * Accuracy & Reliability
			• *Preparedness phase - Priority Index (pre-disaster)*		
			4. Historical damage data must be available	Scale: Expert score based on scan of EM-DAT International Disaster Database (http://www.emdat.be)	Historical damage data is needed to be able to run the model
			5. The damage data collection methodology should be transparent and reliable	Scale: Expert score based on perceived insight in and reliability of these processes	Transparency in the damage data collection methodology provides insight in their reliability. The more reliable the damage counts, the more reliable the toolbox predictions
			• *Response phase (post-disaster)*		
			6. It should be possible for exposure data (weather and hazard) to be collected fast	Time scale	To be rated based on speed of availability of historical exposure data in this country
			• *Overall*		
			7. Data should be collected at the lowest possible administrative levels	Expected level of granularity to be reached	If many of the indicators are only available at district level, than it is not possible to generate predictions at municipal level
		Affordability	8. The data collection cost should be proportional to the added value of collecting the data	Scale: high cost / medium cost / no cost	It is generally cheaper to retrieve open data than closed data. However, there might als be fees to open data
		Data quality	9. The available data should be sufficiently recent (*Recency - DSPI*)	DSPI: Recency component	DSPI = Completeness * Recency * Accuracy & Reliability
			10. The available data should be sufficiently accurate and reliable (*Accuracy & Reliability - DSPI*)	DSPI: Accuracy & Reliability component	DSPI = Completeness * Recency * Accuracy & Reliability
User characteristics	**User capabilities**		11. The toolbox users should be capable of using the toolbox predictions	• Expert score on user capability	From for example an information manager within the organization
				• Data Poverty Index (proxy: level of data literacy)	Determined by internet speeds, computer owners, internet users, mobile phone ownership, network coverage and higher education
	User needs		12. The toolbox output should match the user's information needs	Scale: Matching score of toolbox output with expected country-specific information requirements	If the toolbox predictions do not match the information needs of the intended user, those intended users are unlikely to start using them

TABLE 3 – Overview of conditions to be used to assess the eligibility of a given country for the application of the Community Risk Assessment and Prioritization toolbox

Category	Element	Condition	Maturity level			
			Lagging	Basic	Advanced	Unknown
Data supply	*Data collection*	• *Preparedness phase - Community Risk Assessment (pre-disaster)*				
	Data availability & accessibility	1. It should be possible to get an overview of data providers and available data sets	☐ There is no overview	☐ There is a basic overview	☐ There is a clear overview	☐ *Unknown*
		2. It should be possible to access / collect the data	☐ Majority of the data is closed	☐ Data is partly open / partly closed	☐ Majority of the data is open	☐ *Unknown*
		3. The gap between data needed and data available should be sufficiently small (DSPI: *completeness*)	☐ Large gap: < 50 %	☐ Medium gap: 50 - 75 %	☐ Small gap: 75 - 100%	☐ *Unknown*
		• *Preparedness phase - Priority Index (pre-disaster)*				
		4. Historical damage data must be available	☐ No data available	☐ Some data available	☐ Much data available	☐ *Unknown*
		5. The damage data collection methodology should be transparent and reliable	☐ Not transparent / reliable	☐ Some transparency / reliability	☐ Transparent & reliable	☐ *Unknown*
		• *Response phase (post-disaster)*				
		6. It should be possible for exposure data (weather and hazard) to be collected fast	☐ Longer than 12 hrs	☐ 8-12 hrs	☐ < 6-8 hrs	☐ *Unknown*
		• *Overall*				
	Data usability	7. Data should be collected at the lowest possible administrative levels	☐ National/regional/district level (admin 1, 2)	☐ Municipal level (admin 3)	☐ Community level (admin 4)	☐ *Unknown*
		8. The data collection cost should be proportional to the added value of collecting the data	☐ High or unacceptable collection cost	☐ Medium or acceptable collection cost	☐ No collection cost	☐ *Unknown*
		9. The available data should be sufficiently recent (DSPI: *recency*)	☐ Low recency (< 0,5)	☐ Medium recency (0,5 - 0,75)	☐ High recency (0,75 - 1)	☐ *Unknown*
		10. The available data should be sufficiently accurate and reliable (DSPI: *accuracy & reliability*)	☐ Not sufficient (< 0,5)	☐ Medium (0,5 - 0,75)	☐ Sufficient (0,75 - 1)	☐ *Unknown*
User characteristics	User capabilities	11. The toolbox users should be capable of using the toolbox predictions	☐ No capability	☐ Medium capability	☐ High capability	☐ *Unknown*
	User needs	Possible proxy if 11 *Unknown*: Score on Data Poverty Index	☐ High data poverty (> 3.62)	☐ Above average data poverty (2.42 - 3.62)	☐ Low or below average data poverty (< 2.42)	☐ *Unknown*
		12. The toolbox output should match the user's information needs	☐ No match	☐ Medium match	☐ Match	☐ *Unknown*

TABLE 4 – Data maturity assessment tool to assess the eligibility of a given country for the application of the Community Risk Assessment and Prioritization toolbox

In *Table 4,* the metrics per condition as described in *Table 3* are further specified by defining different possible maturity levels on which a country can be rated on. A red score is considered problematic, a yellow score basic but acceptable, and a green score represents the ideal score. Further interpretation of the scores is discussed in *Section 4.3.2.*
The aimed users of the data maturity assessment tool presented in *Table 4* are the Community Risk Assessment and Prioritization toolbox developers of the 510 Global. They are also considered to be experts being able to provide an expert score on certain metrics when needed. Below, the (origin of the) conditions, metrics and maturity levels are clarified.

Data collection: *Preparedness phase – Community Risk Assessment*

1. It should be possible to get an overview of data providers and available data sets

Clarification: An essential element when deciding on the eligibility of a country for the application of the Community Risk Assessment and Prioritization toolbox, is to assess the data that is available in a country. The data providers both inside and outside the national society should be mapped, including the data they hold and if it can be shared.
Metric: Extensively assessing this is an intensive and time-consuming task, but initial insights can be gained by means of an internet search of key organizations and portals that share their data online and allow for the data to be collated. For example, to what extent is there one specific organization that keeps track of all the data sets available in a country? If a country has a high quality open data portal that combines many different sources, the country scores high on this condition. But a country can also score high on this condition if there are already good connections in this country (with either the Red Cross National Society or other organizations) that have created a good overview of data providers and data sets by means of their in-country experience and have shared this overview with 510 Global.
Maturity level: This internet search in addition with other relevant background knowledge or experience of the expert filling in the tool should result in one of the following judgments (see *Table 4*):

- There is no overview of data providers and available data sets
- There is a basic overview
- There is a clear overview of data providers and available data sets

If there is no overview of data providers and available data sets at all, it will be difficult to start developing the toolbox for this country. Depending on the ratings on the other conditions and the strength of the desire or need to pursue with it, it should be determined how to continue. A field trip could then for example be a means to make a start, and also the possibilities for governance as addressed in *Chapter 5* and *6* could help to increase this maturity level.

2. It should be possible to access the data

Clarification: Data can be available, but it must also be accessible. In an ideal situation, all data is published on an online platform, openly available

to anyone. However, during the toolbox development it became clear that this is often not the case.

Metric: It is possible for a country to score high on the first condition, and still score low on this condition; if the data providers and available data sets are known but this data is not shared on an open data platform it can still be difficult to actually obtain this data.

To be able to give a rating to this condition, the level of openness of the available data serves as a proxy. It is hereby assumed that closed data is more difficult to access / collect than open data.

Maturity level: According to Open Data Charter (2017), which is a collaboration between governments and experts working to open up data, there are four key open data indices:

- Open Data Barometer (OBD), which is "a global measure of how governments are publishing and using open data for accountability, innovation and social impact" (OpenData Barometer, 2017). This index also includes a feature that makes it possible to visualize the comparison between two different countries.
- Global Open Data Index (GODI), which "provides comprehensive snapshots available of the state of open government data publication" (Open Knowledge Network, 2017).
- Open Data Inventory (ODIN), which "assessed the coverage and openness of official statistics to help identify gaps, promote open data policies, improve access, and encourage dialogue between national statistical offices (NSOs) and data users" (Open Data Watch, 2017).
- OECD OURdata Index, which "assesses government's efforts to implement open data in the three critical areas – Openness, Usefulness and Re-usability of government data" (OECD, 2017).

A combined consultation and interpretation of the applicable indices (they all do not have a global coverage), in combination with a validating internet search for open data relevant for the Community Risk Assessment and Prioritization toolbox, should result in one of the following judgments (see *Table 4*):

- The majority of the data is closed
- The data is partly open / partly closed
- The majority of the data is open

If this condition is rated green, it should be possible to develop an initial version of the toolbox for this country without further action. If this condition is rated yellow, this could be given a try, but further action might be needed. If it is rated red, a best strategy should be determined on how to create access to the required data: how can it be collected? There are many possible ways to do so, varying for example from the remote organization of 'mapathons' (in which satellite imagery is used to map roads and houses in areas that have not been mapped yet on the Open

Street Map (OSM) portal) to in-country data collection through enumerators going into the field. A way to categorize data collection methodologies is by distinguishing between data collected 'on' and data collected 'with' communities. *Table 4* provides an overview of different ways of acquiring disaster risk management data 'on' and 'with' communities along a spectrum of non-digital and digital. These classifications are not clear-cut but rather present a spectrum. A survey on a community can be 'with' them if the survey is truly interactive and has a lot of open questions; it can be labelled 'on' them if it is a rigid one-way list of questions. Data acquired in a non-digital way can be digitized later, and also surveys can be executed paper-based and be digitized afterwards. Participatory maps can be created with a community by writing in the sand and using sticks and stones, to digitize this information afterwards.

Table 5 on the next page brings together a - however limited and non-exhaustive - diversity of qualitative and quantitative research methods and instruments from different disciplines. Political, military, economic, humanitarian and development or scientific stakeholders can all collect data on communities with different purposes in mind.

The first three categories are used for collaborating between organizations, whereas the latter four are mostly used within an organization. In cases where the data literacy is low, a hybrid form of services can be used, combing low-tech/paper-based with high-tech (Van den Homberg & Van der Veen, 2017)

Eventually, a shift from non-digital to digital should be generated. How to realize this, will be discussed in *Chapter 5* and *6.*

When it comes to data collection and accessibility of data, it should also be monitored that only data that contributes and generates value is being collected/made accessible. About responsibly handling data in humanitarian action, Raymond et al. (2016) state that "data should never be used simply because they can be; the humanitarian need and potential benefits should be clear and defined" (p.11).

3. **The gap between data needed and data available should be sufficiently small**

Clarification: By mapping the available data on the set of data that is needed, the data and information gap can be determined. For the Community Risk Assessment and Prioritization toolbox to be able to generate accurate predictions, this gap should be sufficiently small.

To determine this gap, apart from having an overview of the available data, insight is needed in the pre-disaster data necessary for the toolbox to function well. As the toolbox is already under development for three different countries, a general idea of which data is needed already exists. As explained, the specific indicators to determine the vulnerability of the

SPECTRUM RANGING FROM	RESEARCH METHOD/INSTRUMENTS	NON-DIGITAL	DIGITAL
COMMUNITY INVOLVED, 'WITH' THEM	**Focus groups**	Face to face session	Closed groups. Messaging apps like Whatsapp
	Semi-structured interviews	Face to face interviews with key informants	Exchanges with key informants through email, skype
	Diaries of individuals in a community		
	Story telling	From visual (e.g. graffiti) up to oral	Tools like Sensemaker
	Participatory mapping	Participatory maps on paper	
	"Kiosk", place where people can go to share information	Volunteer disaster management committees; lowest governmental administrative	Associated digital "hubs", like in Bangladesh Digital Information Center (of the Access to Information Center).
	Vulnerability Community Assessment (VCA)	Often a combination of the methods above: participatory maps, focus groups, diaries.	Digitize the outcomes and migrate them to OSM.
	(Bounded) crowdsourcing, citizen science, citizen reporting (Open Street Map - OSM).		Open groups like civil society organizations for e.g. lobby and advocacy or academic organizations for science. Can be Facebook, Wikipedia, Social SMS, Ushahidi, FrontlineSMS, and Twitter, Fulcrum/PushPin, FormHub, Kobo Toolbox, and the Open Data Kit.
	Self-reporting	Via local government officials/NGO representatives etc.	Via a webportal
	Surveys by National Statistics Offices or NGOs etc. (resulting in micro-data)	Paper based survey	Survey on smartphone app, SMS surveys
	Open source information/intelligence (OSINF/OSINT)	Newspapers, journals, announcements.	Social media, radio and television, webscraping/crawling
	HumINT, human-source intelligence		
	Geo-spatial INT/Imagery INT		Satellite and UAV imagery, pictures
	Signals intelligence		Mobile data collection
COMMUNITY NOT INVOLVED, 'ON' THEM	**Public records**		

TABLE 5 – Different ways of acquiring disaster management data "on" and "with" communities (reprinted from Van den Homberg, Van der Veen & Visser, 2016)

affected areas (pre-disaster secondary data) might differ per country. The decision on which indicators to include in the model is therefore also steered by the availability of the data. There is no clear 'right' or 'wrong' here - the inclusion of different indicators can result in similar predictive accuracies, so it is also a process of trial and error. When determining on the data needed, it is important to consider the user needs (*Condition 12*).
Metric: In the context of the Community Risk Assessment (data preparedness) part of the toolbox, full 'completeness' (so zero gap) is defined as 'having data sources to cover all indicators of the Risk Index (*see Figure 5 in Section 2.3.1*). If for each subcomponent "at least one data source is available, then Completeness is defined as 100%" (Van den Homberg, Visser & Van der Veen, 2017: p.7). Completeness is also one of the components of 'quality' (see *condition 10*).
Van den Homberg, Visser & Van der Veen (2017) of 510 Global have developed a so called 'Data Sets Preparedness Index' (DSPI), which includes the calculation of the completeness. This index can therefore be used to determine the score on this condition.
Other components in the index are 'recency' (condition 9) and 'accuracy & reliability (condition 10).
Maturity level: Filling in the *Data Sets Preparedness Index* for the completeness component should result in one of the following judgments (see *Table 4*):

- There is a large gap between data needed and data available (< 50% completeness)
- There is a medium gap between data needed and data available (50 – 75% completeness)
- There is a small gap between data needed and data available (75 – 100% completeness)

In their article on the DSPI, Van den Homberg, Visser & Van der Veen (2017) explain that a score of 0.5 (50%) at all three DSPI components can be interpreted as reasonable.
Also in the context of the Community Risk Assessment and Prioritization toolbox it is aimed to reach this score for all three components, to have the predictions be based on a sufficiently solid data foundation.
In an ideal situation, the score on all three components should be equal to 1.
Based on the identified data gap (the *completeness*), a plan on how to collect that data can be established - e.g. either through field based mapping activities or through trying to establish data-sharing agreements with the data providers that have been identified that do have data in their possession, but did not yet open it up (Van den Homberg & Van der Veen, 2017).

Data collection: *Preparedness phase – Priority Index*

4. Historical damage data must be available

Clarification: A pre-condition for a country to be eligible for the toolbox application for a certain disaster, is that historical impact data (people affected, houses damaged and destroyed, casualties) of similar disasters that occurred in the past is available. This data is necessary to validate the model; without it, the model cannot be developed.

Metric: If available, this data can be retrieved from the EM-DAT International Disaster Database.

Maturity level: A search through this database of the expert filling in the tool should result in one of the following judgments (see *Table 4*):

- There is no historical damage data available
- There is some historical damage data available
- There is much historical damage data available

This condition can be a blocker, as historical damage data is needed to train the model. Hence, if no historical impact data is available, further country assessment can stop as this condition needs to be fulfilled – it is a critical condition, and one that is impossible to directly influence (efforts could be made to start collecting damage data from this point onwards, but historical damage data cannot be 'created' in retrospect).

If some data is available (*yellow*), the toolbox developer should judge whether this is sufficient data to be able to appropriately validate the prioritization model. A green score on this condition would mean that the availability of historical damage data is not an issue.

5. The damage data collection methodology should be transparent and reliable

Clarification: This condition also strongly determines the possibility for applying the Community Risk Assessment and Prioritization in a country. The model predicts damage, and hence a proper methodology to count this damage is essential.

Metric: As described earlier (in *section 2.3.1 - Rapid execution and timely collection of data*), it is expected that these damage counts are generally not very transparent nor reliable, due to i.a. a lack of centralized methodologies, wrong estimates being made and local governments reporting biased information. Thus, there are varying degrees of the transparency and reliability of damage data collection methodologies in different countries, which affects the accuracy of the toolbox predictions.

Maturity level: The toolbox developer filling in this tool could discuss the rating of this condition with a Red Cross employee – perhaps someone within the 510 Global team – who has experience in the country being assessed. This should result in one of the following judgments (see *Table 4*):

- The damage data collection methodology is not transparent / reliable
- There is some transparency / reliability
- The damage data collection methodology is transparent & reliable

Data collection: *Response phase*

6. It should be possible for exposure data (weather and hazard) to be collected fast

Clarification: As explained before, 510 Global aims to publish the Community Risk Assessment and Prioritization toolbox predictions within 12-24 hours after a disaster has struck. If the predictions are generated within this timeframe, the outcomes of the toolbox are considered useful for the toolbox users – the humanitarian aid workers.

Metric: Rainfall and wind speed data on affected areas usually becomes openly available around six to eight hours after a disaster has struck. If this data comes in by the times all other model elements have been made ready (which is currently mostly not the case), it still takes a couple of hours before the predictions have been generated, as this is the time the model needs to run. This means that it should not take too long before the exposure data comes in.

Maturity level: This condition could be rated based on the speed with which disaster data became available for past (similar) disasters.

If this is possible, to trace, it should result in one of the following judgments (see *Table 4*):

- It takes longer than 12 hours before exposure data can be collected
- It takes 8-12 hours before exposure data can be collected
- It takes shorter than 6-8 hours before exposure data can be collected

This is not a completely reliable proxy, since it does not necessarily say something about the speed with which new exposure data will become available, yet it could give an indication. Therefore, a red score on this condition (if filled in for previous disasters) should not have a very strong weight: it could be different for new disasters. Nevertheless, it is important to take this condition into account, or to be aware of it: when it takes a very long time before exposure data becomes available, it makes no sense to run the model anymore.

Data collection: *Overall*

7. Data should be collected at the lowest possible administrative levels

The reason this condition should be considered has also already been explained before (*Section 2.3.2 – Data from low administrative levels*): It is important to have data available from the lowest administrative levels (municipality or even community) so that detailed prioritization can take place based on the toolbox predictions. This will enable decision-makers to better target relief sources to the right locations. An issue is that there generally is a lack of data from these lower administrative levels, as it is more challenging to collect this data.

Metric: The Community Risk Assessment and Prioritization toolbox generates predictions as granular as its 'weakest link'. In the first place, all data should be organized by the same level of aggregation as the predicted variable (the damage). But if the predicted variable is of a much

lower administrative level than many of the indicators, these indicators determine the administrative level for which the model can generate predictions: if several indicators are only available at province level, then it is not possible to generate predictions at municipal level.

Maturity level: This condition could be rated based on the expected level of granularity to be reached (see *Table 4*):

- National or administrative level 1 or 2 (regional level / province or district level)
- Administrative level 3 (municipal level)
- Administrative level 4 (community level)

If there is only data available at a national, regional or province level, it will not make sense to develop the toolbox because in that case the toolbox application loses its effectiveness. The aim is to be able the generate predictions at a community level in the future, but up until then predictions at a municipal level are also perceived as useful.

Affordability

8. The data collection cost should be proportional to the added value of collecting the data

Clarification: Part of the data 510 Global collects as an input for the Community Risk Assessment and Prioritization toolbox is open data. When it comes to this open data, 510 Global is always looking for data sets that are freely available – they have so far never paid to retrieve online available data in the current toolbox development. However, once the toolbox will be scaled up and users request a certain improvement in for example granularity of the toolbox predictions, then the question will be: how much is this worth paying for? This is a relevant question, since – when not only using freely available open data – there are cost related to data collection. For example, field assessment methods are costly, so it would be relevant to then start comparing these with similar remote open data collection in terms of costs and benefits (e.g. quality).

The financial resources of 510 Global are limited, so it is important that this condition is considered once it becomes relevant: the condition must be considered as a boundary, especially in the future of the toolbox development.

Metric: Currently there are no direct data collection costs. Once resources will be deployed purely to 'fill' the Community Risk Assessment & Prioritization toolbox, this condition should be rated red until it is clear whether these costs are acceptable or not: whether the cost weigh up against the benefit this data collection will realize. In that case, a cost-benefit analysis could be a suitable basis for a decision. The specification of this cost-benefit analysis falls outside of the scope of this research, as it would not offer relevant added value now.

If there are collection cost but they are either low or acceptable, this condition can be rated yellow. If this condition is marked 'unknown', this

could be a blocker until it has been made clear that there are no high or unacceptable cost involved.
Maturity level: See explanation above.

- High or unacceptable collection cost
- Medium or acceptable collection cost
- No collection cost

Data quality

9. The available data should be sufficiently recent

Clarification: It is assumed that the higher the data quality, the more accurate the predictions of the Community Risk Assessment and Prioritization toolbox will be. But how can you determine 'data quality'? Several international organizations, like the European Union, The United Nations and the International Monetary Fund, have already developed different data quality assessment frameworks. The 'Post2015 Data Set Initiative' has merged these into their own framework, and from this framework Van den Homberg, Visser & Van der Veen (2017) identified three data quality components that are most relevant in the operational environment of 510 Global:

- Completeness
- Recency
- Accuracy & Reliability

The 'completeness' component has already been covered with *condition 2*.
Metric: Van den Homberg, Visser & Van der Veen (2017) explain 'Recency' as follows:

> *"Recency is a combination of when the data set was last updated and how long a data set remains representative of the reality. The more recent the source is, the higher the Recency score should be. How long a data set remains representative can also be termed "retention period". Some data sets stay valuable longer than others. For example, the retention period of a source about the geospatial distribution of earthquake risk is generally deemed longer than the retention period of a research on Good Governance Index of local municipal government, which might very well have changed considerably after a new local government was elected"* (p.7).

They additionally state that rating the retention period of a data source is a subjective matter, which can be done for example through averaging expert opinions (Van den Homberg, Visser & Van der Veen, 2017). It is, just like *completeness*, part of the *Data Sets Preparedness Index* (DSPI). This index can therefore be used to determine the score on this condition.
Maturity level: Filling in the *Data Sets Preparedness Index* for the recency component should result in one of the following judgments (see *Table 4*):

- Low recency (0 – 0,5)
- Medium recency (0,5 – 0,75)

- High recency (0,75 – 1)

For the assignment of these values, the same explanation holds as for *condition 3,* which relates to the completeness component of the DSPI. When most of the data is not very recent (and hence the recency score is low) this could cause issues for the accuracy of the toolbox predictions.

10. The available data should be sufficiently accurate and reliable

Clarification: The third component of data quality is 'Accuracy & Reliability'. Data sources can vary widely in their accuracy and reliability. An example of the determination of 'accuracy & reliability' provided by Van den Homberg, Visser & Van der Veen (2017) is that a "census is generally considered more accurate than a survey, because it measures the entire population instead of a sample" (p.8).

Metric: Accuracy & Reliability can also be evaluated by looking at data sources and its publishing organizations, which is quite a qualitative assessment (Van den Homberg, Visser & Van der Veen, 2017). It is, just like *Completeness* and *Recency* a factor in the Data Sets Preparedness Index, and this index can therefore be used to determine the score on this condition.

Maturity level: Filling in the *Data Sets Preparedness Index* for the accuracy & reliability component should result in one of the following judgments (see *Table 4*):

- Not sufficiently accurate and reliable (< 0,5)
- Medium accurate and reliable (0,5 – 0,75)
- Sufficiently accurate and reliable (0,75 – 1)

For the assignment of these values, the same explanation holds as for *condition 3* and *9,* which relate to the other components of the DSPI. If the score on this condition is red, this means it will not be possible to generate accurate and reliable predictions with the toolbox, and other data should be found and used.

User capabilities

11. The toolbox users should be capable of using the toolbox predictions

Clarification: Clearly, if this condition is not fulfilled, it will not make sense to apply the toolbox in a country.

Metric 1: A good proxy for this condition is the level of 'data literacy' in a country. Data-Pop Alliance (2016) define data literacy as "the desire and ability to constructively engage in society through and about data". It is expected that if aid workers are data literate, they will be better able to make sense of the outcomes of the Community Risk Assessment and Prioritization toolbox. Data literacy can be measured at different levels. In this context, it is most relevant to judge the capabilities of the actual toolbox users. This could be done based on for example the available (and

actually used) services and tooling within these organizations, as these could serve as an indicator for the level of data literacy.
The following tooling categories can be thought of:

- Communication/Instant messaging software
- Collaborative digital tooling/project management
- Cloud-based file sharing (inventory of country-specific geospatial data-sharing platforms)
- Mobile data collection (and other data-related tooling)
- Dashboard technologies for data analysis/visualization
- Regular office applications
- Graphics/Layout

Maturity level 1: An expert judgment or consultation of someone with experience in the organization and country subject to this judgment, should result in one of the following judgments:

- The toolbox users do not have the capability to use the toolbox predictions (no capability)
- The toolbox users have some level of capability to use the toolbox predictions (medium capability)
- The toolbox users have the capability to use the toolbox predictions (high capability)

If this condition is rated red (or yellow), then something needs to be done to make sure the toolbox will end up being (more) useful.

Metric 2: However, if the availability and use of these kinds of services and tooling is unknown, the 'Data Poverty Index' could be used as a proxy, to create a better understanding of the data literacy in a country at national level. The Data Poverty Index is based on internet speeds, computer owners, internet users, mobile phone ownership, network coverage and higher education (Leidig, Teeuw & Gibson, 2016). It provides results for 189 of the 214 countries listed by the World Bank (Leidig & Teeuw, 2015).
Maturity level 2: By consulting the Data Poverty Index, the following judgment can be made:

- The data poverty is high (> 3.62)
- The data poverty is above average (2.42 – 3.62)
- The data poverty is below average or low (< 2.42)

As this metric serves as a proxy for *Metric 1* for this condition, the same holds: if this condition is rated red (or yellow), then something needs to be done to make sure the toolbox will end up being (more) useful.

User needs

12. The toolbox output should match the user's information needs

Clarification: The toolbox output should match the information needs of its users, otherwise they will not perceive value out of the use of the toolbox and are not likely to start using it.

Metric: A qualitative assessment can provide insight in whether the toolbox outcomes match these information needs.
Maturity level:

- There is no match
- There is a medium match
- Thee is a match

It is generally expected that there is at least a medium match: a field study has been done on the information needs of aid responders in the Philippines (Van Lint, 2016) and it is expected for those information needs not to be drastically different in other countries. However, there could of course be differences and certain output that is very valuable in one country, does not necessarily need to be in a different country.

4.3.2. Threshold values of conditions to be fulfilled

Now all the conditions and related metrics and maturity levels have been defined, the 'threshold values' per condition should be determined: to what extent should a condition be fulfilled for a country to be eligible for the toolbox application? Which maturity level needs to be present?
It was originally aimed to make the metrics more quantitative and clear-cut, instead of the mainly qualitative metrics that have been defined now. It was even intended initially to make it possible to sum all scores on the conditions (all checked maturity levels) to one 'data readiness' score that could give an indication of the situation as to the data landscape in a country. However, this proved to be more difficult and complex in practice than expected up front, for various reasons.
In the first place, the aimed user of the assessment tool are internal developers of the Community Risk Assessment & Prioritization toolbox within the 510 Global team. For them, using the tool should be a rather quick way to make an initial assessment of a country, and get an idea of where it is positioned when it comes to data preparedness. For the toolbox developers to get the idea that filling in the tool does not require a lot of time, the maturity levels needed to be defined quite broadly.
Secondly, the tool does not provide a black or white assessment: you cannot indisputably state that the presence of five checks in a yellow (basic) box means that it is impossible to develop the toolbox for this country. Expert judgements should always be included; The tool outcomes need to be interpreted on a case by case basis. This leads to the third point: the tool is not country-specific; it should be possible to fill it in for any country for which the development of the Community Risk Assessment & Prioritization toolbox is expected to be of use. This means that the metrics and maturity levels cannot be country-specific either, and for those not to be country-specific, they must be expressed in a generic manner.
Lastly, it is difficult to make the tool mainly quantitatively-based if only three cases form the basis for its development. This is not a very large sample size and a biased impression of statistical significance would be given.
So, this has resulted in the tool not being a black-and-white assessment tool, but rather being a checklist for the toolbox developers to have an overview of aspects that could play a role in the success of the toolbox development for a country, that provides a direction on how they should proceed. It hence creates awareness. Filling in the assessment tool allows the

developers to make a scan of a country and to get an idea of the existing data landscape. The outcomes should be interpreted by the one using the tool; no black and white values can be attached to it. The outcomes could therefore also serve as an input for a discussion between different parties who need to come to a collective understanding of the data landscape in a country, after which an iteration of the filling in of the tool could take place to integrate and merge additional relevant knowledge of those parties.
In general, it can be stated that 'red' checks for conditions need to be given attention before the toolbox can be developed for a country. Within the red checks - and within the conditions in general - a distinction can be made between conditions 510 Global can directly influence, and conditions for which 510 Global is dependent (either partly or completely) on external parties or factors.
For those conditions for which 510 Global is dependent on others, a red maturity level ('lagging') could be a true process blocker, as 510 Global does not have the (direct) power to do something about it. Those conditions are:

Condition 4 – Historical damage data must be available.

Condition 6 – It should be possible for exposure data (weather and hazard) to be collected fast.

Condition 8 – The data collection cost should be proportional to the added value of collection the data.

If one of these conditions is checked red in the assessment tool, a warning message should pop-up stating that this condition is critical and that it should be investigated further before proceeding.
If *Condition 4* is checked red, the model cannot be trained – the availability of historical damage data is essential for a proper functioning of the model.
If *Condition 6* is checked red, there is a possibility that the outputs will be generated too late to be still of use. For this condition, it must be made sure though that this red check also holds for upcoming data after new disasters, something that is difficult to predict. Also, a red check on this condition has no effect on the possibilities for the development of the Community Risk Assessment dashboard, which can also be generated without the priority index component to become data prepared.
If the data collection cost (*Condition 8*) are too high, alternative data collection methodologies need to be explored.
If one or multiple of the other conditions are checked red, this means that it will be impossible (or very difficult) to immediately start developing the Community Risk Assessment & Prioritization index. However, for these conditions it is possible to start working on an improvement of these conditions to come to a higher maturity level, and increase the eligibility of this country for the toolbox application. In *Chapter 5,* suitable ways to do this will be proposed and discussed.
If most of the checks are green and there are no red checks, it is quite likely that the development of the Community Risk Assessment and Prioritization toolbox for this country will succeed.

If there are a lot of yellow checks, this does not necessarily mean that the tool cannot be developed. It is recommendable to start deploying a governance approach as proposed in *Chapter 5* and *6,* but it could be possible that an initial version of the toolbox can already be developed: many of the green maturity levels represent an ideal situation, whereas many of the yellow maturity levels stand for a less ideal but acceptable or workable alternative.
Another thing to highlight, is that apart from 510 Global taking action to improve the maturity scores of a country on the assessment tool by 'governing' the data ecosystem (*Chapter 5*), also the development of the Community Risk Assessment and Prioritization toolbox will improve the maturity scores on the condition: it is a cyclical process (see *Figure 17*).

FIGURE 17 – Visualization of impact of research steps (own figure)

4.4. CHAPTER CONCLUSION

In this chapter, the following question has been answered:

SQ3: HOW CAN THE DEVELOPED DATA ECOSYSTEMS FRAMEWORK BE APPLIED TO THE CONTEXT OF SCALING UP THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX?

In *Section 4.2,* An evaluation of the relevance of the theoretical data ecosystem success criteria with a group of team members from 510 Global in combination with a desk research led to the insight that the humanitarian data ecosystem around the Community Risk Assessment and Prioritization toolbox is still very much in its infancy. It is emerging, but currently immature. As a result, not all theoretical data ecosystem elements are considered equally relevant in the current phase of the humanitarian data ecosystem.
Some of the 'Data supply' and 'User characteristics' framework elements only become relevant in a later, more mature data ecosystems phase, as their importance increases once more data-sharing starts to take place: the elements privacy, licensing and legal conditions, metadata and data standards are either covered by other research groups (and hence not considered in this study due to overlap) or omitted in the further analysis of this study because are considered of secondary concern. A desk research showed that the governance part of the theoretical ecosystems framework has so far remained very much underdeveloped in the practical context of the Community Risk Assessment and Prioritization toolbox. Governance processes in line with the theoretical governance elements should therefore be developed as such that they positively influence the 'Data supply' and 'User characteristics' elements that require immediate attention for the toolbox development to be enabled: data collection, affordability, data quality, user capabilities and user needs.
To be able to assess and influence those elements, they were operationalized. This has been done in *Section 4.2*. Based on expert input combined with a desk research, a 'data maturity assessment tool' has been created that can be used to assess the data landscape in a country. The assessment tool consists of different conditions, categorized by the elements of the

theoretical framework, that need to be met (at least to a certain extent), to be able to develop the Community Risk Assessment and Prioritization toolbox for a country. They can be rated on their 'maturity level'. The tool does not provide quantitative, clear-cut decision-support, but rather gives an idea of the existing data landscape in a country, and highlights points for attention or problematic issues. The next question is: how can the maturity scores on these conditions be improved? How can an optimal context for developing the Community Risk Assessment and Prioritization toolbox be created? Those questions will be answered in the next chapter, by targeting the possibilities for the development of the governance part of the theoretical data ecosystems framework that, as described, has yet remained underdeveloped.

CHAPTER 5 – A GENERIC GOVERNANCE APPROACH

5.1. INTRODUCTION

The outcome of *Chapter* 4 consists of the operationalization of the theoretical data ecosystems framework developed in *Chapter 3*, in the form of a tool with conditions that can help to assess the eligibility of a country for the application of the Community Risk Assessment and Prioritization toolbox by providing an overview of the existing data landscape. Those conditions have been derived from the ecosystem elements that are considered most important in the initial phase of the toolbox development.

The key in this study however, is not only to assess a country's eligibility for the toolbox application, but also to improve it, as this would enable the actual scaling up of the toolbox. That part is dealt with in this chapter, by focusing on the governance of the humanitarian data ecosystem around the Community Risk Assessment and Prioritization toolbox. In *Chapter 3,* governance was defined as the framework of policies, processes and instruments to realize common goals in the interaction between different public and private sector entities (based on the description of Welle Donker & Van Loenen, 2017). In a data ecosystem, governance can help to facilitate the data supply and assist users to function in the data ecosystem.

As explained, it is expected that a focus on the governance part of the data ecosystems framework, which is currently underdeveloped for the context of the Community Risk Assessment and Prioritization toolbox, plays a key role in scaling up the toolbox. Introducing a governance structure should help to facilitate the roll out of the toolbox in a new country. It is assumed that a humanitarian data ecosystem spans one country, and that the governance is taken care of within this country. The aim of this chapter is hence to figure out what changes and which governance investments in policies, strategies, practices and relationships are necessary to improve the existing conditions related to the data landscape in a country.

In this chapter, the following research question is answered:

SQ4: HOW SHOULD THE GOVERNANCE OF THE DATA ECOSYSTEM AROUND THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX TO FACILITATE THE PROCESS OF SCALING UP THE TOOLBOX BE APPROACHED?

a. *How can the governance criteria for a successful data ecosystem as derived from literature be translated to the context of the Community Risk Assessment and Prioritization toolbox?*
b. *How can these governance criteria be arranged in such a (practical) way that they create an optimal context for the development of the Community Risk Assessment and Prioritization toolbox?*

Sub-question 4a (*Section 5.*2) was answered by translating the theoretical governance concepts derived from literature in such a way that they become of practical relevance in the context of scaling up the Community Risk Assessment and Prioritization toolbox. This step was important, because to answer *sub-question 4b* (*Section 5.2-5.4*), semi-structured interviews with humanitarian data experts were held, to get an idea of how to meet the criteria for successful data ecosystem governance. These interviewees had to be presented with success criteria that were easy for them to interpret and which they could relate to their practical experience. The interview questions posed to the experts followed from the outcome to *sub-question 4a*, and were thus in line with the theoretical framework, but presented in a less conceptual way.
The practical outcome of this chapter is an overview of a proposed arrangements of the different governance criteria to create an optimal context for rolling out the Community Risk Assessment and Prioritization toolbox, based on the empirical findings from the interviews. This specification of the governance should be able to facilitate a positive shift in the data maturity of a country, thereby improving the eligibility to apply the toolbox.
Additionally, there is a theoretical outcome to this chapter. This research step also served as an empirical validation of the theoretical framework developed: applying the governance part of the theoretical framework in practice and testing it on the humanitarian case of scaling up the Community Risk Assessment and Prioritization toolbox, provided the opportunity to evaluate the framework on its suitability for application in practice, and discuss and propose adjustments. This is done in *Section 5.5.*

5.2. PRACTICAL RELEVANCE OF THEORETICAL GOVERNANCE CRITERIA

The specification of the way 510 Global should govern the data ecosystem around the Community Risk Assessment and Prioritization toolbox should be based on the governance criteria from the theoretical framework, as regarding literature these are the governance criteria that need to be present to establish a successful data ecosystem. However, these theoretical criteria are too abstract to directly apply in the humanitarian toolbox context. In this section, the criteria are therefore translated into specific success criteria when pursuing the goal of scaling up the Community Risk Assessment and Prioritization toolbox; when developing the toolbox for a new country.
Table 6 below lists this translation to practice per governance element. The 'description in literature' comes from the outcomes from the systematic literature review presented in *Section 3.4*, and the 'translation to toolbox context' has been done based on logical reasoning from the insights gained in previous sections and discussions with 510 Global team members.

CRITERIA	THEORETICAL AND PRACTICAL APPLICATION
1. VISION, COMMUNICATION AND STIMULATION	**Description in literature** To establish a collaborative environment in which a common goal is provided, to avoid a fragmented approach. **Translation to toolbox context** To be able to apply the toolbox in a country, it generally helps if there is one national data platform on which data is being shared.

CRITERIA	THEORETICAL AND PRACTICAL APPLICATION
DIVISION OF ROLES AND RESPONSIBILITIES	**Description in literature** To assign different roles and responsibilities. • Identify all actors in the ecosystem • Locate relative position of the actors in the ecosystem in terms of roles and responsibilities **Translation to toolbox context** In *Figure 9* (*Section 2.4 – Actor complexity*) the main actors around the Community Risk Assessment and Prioritization toolbox were visualized. The roles in the data ecosystem around the Community Risk Assessment and Prioritization toolbox that need to be assigned are the data providers and the eventual end-users of the toolbox predictions. This means this theoretical success criterion can be translated into two practical success criteria: • There should be an overview of (possible) data providers and the data they possess. • It should be clear which organizations to target as the users of the predictions of the Community Risk Assessment and Prioritization toolbox.
FEEDBACK	**Description in literature** Feedback mechanisms should be present to enable data users to provide feedback to data providers. **Translation to toolbox context** In the context of the prioritization model, it could be very useful for 510 Global to be able to provide feedback to data providers, as to for example the quality or structure of the data they provided. An additional feedback loop can be identified between the toolbox users and 510 Global: how did they experience the use of the toolbox?
LEADERSHIP	**Description in literature** To have a problem owner who stimulates and coordinates open data activities. • E.g. top-down or bottom-up? • E.g. one leader or group of leaders? **Translation to toolbox context** To be able to apply the prioritization model in a country, it is important that there is a local party in a country that is capable and willing to take responsibility for rolling out and managing the Community Risk Assessment and Prioritization toolbox (a local owner / local lead). This party should for example establish and maintain partnerships with data providers to enable the required data collection.
INCENTIVIZATION	**Description in literature** To incentivize key stakeholders to take up an active role and to share data. • E.g. make data sharers feel intrinsically rewarded • E.g. lower the barrier to entry for data-sharing **Translation to toolbox context** All the actors involved in the data ecosystem around the Community Risk Assessment and Prioritization toolbox should be incentivized to take up the role that has been assigned to them: • *Local leader:* When there is a suitable 'local owner' (see 'leadership' criterion) this party should be incentivized to take up this role.

CRITERIA	THEORETICAL AND PRACTICAL APPLICATION
	• *Data providers:* Data might be available, but it is only useful when it is also accessible. This means that data providers should have an incentive to share their data; they should be willing to do so. • *Toolbox users:* Possible users of the Community Risk Assessment and Prioritization toolbox should be incentivized to start using it.
PARTICIPATORY CAPACITY	**Description in literature** Public bodies should have the capacity to be able to participate in the ecosystem – to meet the data demands. • Technical knowledge on certain systems and technologies involved • Data management knowledge (on how to ensure high data quality) • Operational knowledge on how to incorporate data activities into current practices **Translation to toolbox context** Data providers should not only be *willing* to share the data; they should also be *capable* of making their data accessible: they should be data literate (i.e. able to constructively engage in data-related activities).
SUSTAINABILITY: *FINANCING*	**Description in literature** For a data ecosystem to become sustainable, sustainable financing should be arranged. **Translation to toolbox context** In the context of the Community Risk Assessment and Prioritization toolbox, it should be determined which party will be responsible for the financing of the toolbox development in a country.
VALUE CREATION	**Description in literature** For a data ecosystem to become sustainable, value should be generated for the ecosystem stakeholders. **Translation to toolbox context** This criterion will be covered with the 'incentivization' criterion instead, as the two are perceived to be overlapping (incentivizing actors to participate means that some sort of value needs to be created for them).
TRUST AND TRANSPARENCY	**Description** For a data ecosystem to function well, data users and data providers should trust each other and be transparent to one another. **Translation to toolbox context** For the data providers to be transparent about the origin of the data they share and for the data requesters to be transparent to the data providers about what they are aiming to use this shared data for can have a positive influence on data-sharing.

TABLE 6 – Translation of theoretical governance criteria to practical context of research

5.3. SET-UP SEMI-STRUCTURED INTERVIEWS

To come to a generic, country-independent governance approach on how to improve the data landscape in a country to make it eligible for the toolbox application and hence facilitate a scale-up, semi-structured interviews with humanitarian experts from countries or regions with varying data landscapes have been conducted.

In the coming sections, the following topics related to the set-up of these interviews will be discussed:

1. Develop interview questions (5.3.1)
2. Select interviewees (5.3.2)
3. Prepare and hold interviews (5.3.3)
4. Process interviews (5.3.4)

The interview outcomes will be discussed in *Section 5.4.*

5.3.1. Development of interview protocol

The interview questions have been derived from the practical translations of the theoretical data ecosystem concepts as described in *Section 5.2.2.* The order of the concepts was switched in such a way that the interview would be built up in a logical way. The complete interview protocol can be found in *Appendix D*. The first question for each concept – if the concept lends itself to this approach – is to ask the interviewee to what extent a certain concept is present in his/her country on a scale from 1-5. If they give a high score, the next question is to explain which factors contribute to this high score and how can it be improved even further. If they give a low score, the next question will be to explain the barriers that cause this low score, and to come up with ideas on how to improve this low score. It was expected that the use of this structure would allow for the identification of generic best practices to achieve high scores on each governance concept.

5.3.2. Selection of interviewees

From the format for the interview protocol it became clear that to provide valuable answers to the interview questions, the interviewees needed an extensive knowledge of the entire humanitarian data landscape in a country – a much broader overview than just the department that he or she is working at. Therefore, a selection of suitable interviewees was made based on the expected presence of this kind of knowledge and expertise, combined with the aim to come to a set of interviewees from a widely varying set of countries to be able to compare approaches proposed for different data landscapes, and eventually come to a generic approach.

Not many suitable interviewees were identified; if there would be a very large number of people with this kind of expertise, this research would probably have long lost its relevance, as in that case scaling up the Community Risk Assessment and Prioritization toolbox would most likely not be as challenging as it is now. Yet, it was found that people involved in the development of 'INFORM Subnational', a process that very much resembles the development of the Community Risk Assessment and Prioritization toolbox (for a more elaborate description see *Text box 1*), met the criterion of having a broad knowledge of the data landscape in their region. This insight combined with the connections of 510 Global and additional contacts that some of the interviewees recommended to get in touch with ('snowball sampling'), led to the set of interviewees as listed in *Table 7*. In this table, also the interview details, which are further elaborated on in the next section, are specified.

Text box 1 – INFORM Subnational

As described in *Section 2.3 – Technical complexity,* INFORM is a global, open-source risk assessment for humanitarian crises and disasters. They generate risk assessments for countries on a national level ('country profiles'), but during the course of this research they have also started to publish subnational risk indices, visualizing and breaking down risk and its components within a single region or country. The so-called 'INFORM Subnational' uses the same methodology and development process as the global INFORM, but is subnational in resolution (INFORM, 2016). What is interesting, is that "developing an INFORM Subnational model is a locally owned and managed process that is supported by the global INFORM initiative" (INFORM, 2016). As the indicators that serve as an input for the Community Risk Assessment and Prioritization toolbox are based on the INFORM index, it could be stated that 510 Global and INFORM Subnational are pursuing the same objective: collecting subnational data from different data providers on the INFORM indicators, and processing this into a risk index. The current difference between the two, is the administrative level at which the two organizations pursue to collect the data: even within the subnational models, INFORM sometimes still operates at higher administrative levels than 510 Global. To illustrate this: for four of the seven INFORM Subnational models, data is shown on a national level (when certain countries have been merged into one region in the global INFORM index) or on administrative level 1 or 2, whereas 510 Global aims to generate predictions on a granular level of administrative 3, or even higher. Nevertheless, developing INFORM Subnational is a process that shows many similarities with rolling out the Community Risk Assessment part of the Community Risk Assessment and Prioritization toolbox: the two are practically identical. This was confirmed in an interview on the high-level INFORM Subnational methodologies with an INFORM Disaster Risk Consultant (Interviewee 1, personal communication, June 28, 2017). To visualize this, the outcomes of the two different models (INFORM Subnational and Community Risk Assessment Dashboard) are visualized below in *Figure 18* for the INFORM category 'Lack of Coping Capacity'. Both visuals also exist for the 'Vulnerability' and 'Hazards exposure' category, as well as for the overall risk score. As can be seen, even though these figures show the risk scores of different countries, the idea is the same and the scores are calculated based on the same indicators.

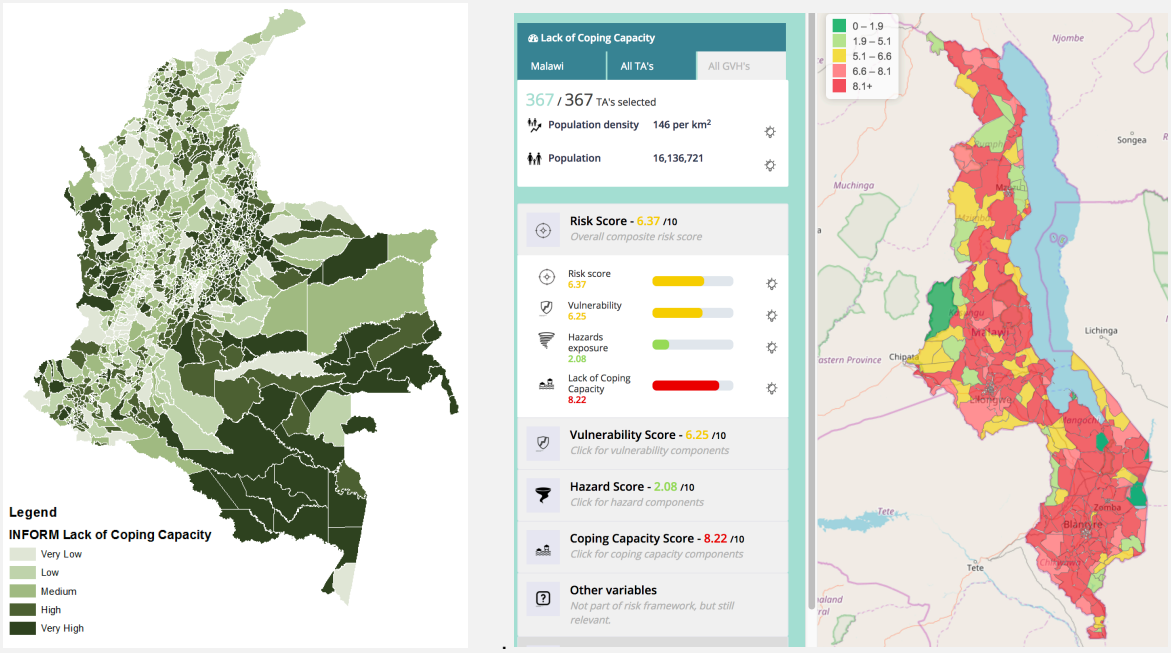


FIGURE 18 – INFORM Subnational dashboard for Colombia (left, INFORM, 2017) and Community Risk Assessment Dashboard for Malawi (right, 510 Global, 2016A): illustration of similarities

In the future, INFORM also plans to target higher administrative levels for the regions for which they have not done this yet – though it will be judged on a case-by-case basis whether this is desired. This depends on the use and scope of the INFORM model, and on the administrative level on which risks are managed and on which authorities operate (Interviewee 1, personal communication, June 28, 2017).

In June 2017, 510 Global was represented at the annual INFORM meeting, in which possibilities to join forces in the future were discussed. So far, the two organizations are targeting different countries, but once their interests in countries start to overlap it would be more efficient to collaborate instead of both carrying out the same work.

INTER-VIEWEE	ORGANI-ZATION	POSITION	RESPONSIBILITIES	REGION	INTERVIEW DATE	INTERVIEW MODE	INTERVIEW DURATION	INTERVIEW LANGUAGE
1	INFORM	Disaster Risk Consultant	I.A. the coordination of the development and roll out of INFORM models	Global	28-06-2017	Skype	45 min	English
2	Malawi Red Cross Society	GIS[3] / data analyst on project base in Malawi	● Setting up a workshop on Data for Sustainable Development Goals ● Visiting and mapping all data producing stakeholders ● Setting up a data team (GIS expert, data analyst and team lead) to build capacity within Malawi Red Cross	Malawi	31-07-2017	Skype	1h30	Dutch
3	UNOCHA	Information Management Analyst	Support to emergency response: GIS mapping / data preparation / supporting operations / pre-disaster data collection / collecting data for emergency response	Philippines	1-08-2017	Skype	1h15	English
4	NLRC	Data Scientist	● Head developer of the Community Risk Assessment & Prioritization toolbox ● Experience as Information Manager for the Philippine Red Cross	Multiple online data landscapes	8-08-2017	In person	1h	Dutch
5	UNICEF (interview together with interviewee 6)	Independent consultant, data and information management and mapping	I.A. the coordination of rolling out INFORM Subnational for Guatemala and Colombia, currently starting the development of INFORM Subnational for Honduras and El Salvador.	Latin America and Caribbean	10-08-2017	Skype	1h15	English
6	UNICEF (interview together with interviewee 5)	Disaster Risk Management and GIS Specialist	I.A. the coordination of rolling out INFORM Subnational for Guatemala and Colombia, currently starting the development of INFORM Subnational for Honduras and El Salvador.	Latin America and Caribbean	10-08-2017	Skype	1h15	English
7	UNOCHA	Lead Analyst for Eastern and Southern Africa	● Head of the Analysis and Information Management Unit of UN Nairobi ● Rolled out INFORM Subnational for the Greater Horn of Africa	Greater Horn of Africa[4]	10-08-2017	Skype	45 min	Dutch
8	UNOCHA	Chief of Information Management Unit	I.A. the coordination of rolling out INFORM Subnational for the Southern Africa Development Community (SADC)	Southern Africa Development Community (SADC)[5]	11-08-2017	Skype	1h	English

TABLE 7 – List of interviewees and interview specifications

[3] Geographic Information System, focusing on spatial data
[4] The Greater Horn of Africa comprises eight countries: Djibouti, Eritrea, Ethiopia, Kenya, Somalia, South Sudan, Sudan and Uganda.
[5] The SADC has 16 member states: Angola, Botswana, Comoros, Democratic Republic of the Congo, Lesotho, Madagascar, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia and Zimbabwe.

Together the interviewees represent five different cases, which are defined by the country or region the expertise of the interviewees is focused on. All are either familiar with the Community Risk Assessment and Prioritization toolbox, or with the development of INFORM Subnational.
Four more possible interviewees were identified and contacted multiple times, but did not respond or were not available on short notice. It was decided not to pursue them or look for additional interviewees. After having interviewed the persons listed in *Table 7,* it was not expected that having more interviews would lead to significant new insights.

5.3.3. PREPARATION AND CARRYING OUT OF INTERVIEWS

The interviewees were sent an e-mail (see *Appendix C*) introducing the topic and asking for their availability for a Skype interview. One of the interviewees wished to see the questions prior to the interview, but for the others the introduction was sufficient. The interviewees whose specific responsibilities and field of expertise were not entirely clear yet, were asked to send some information about this to optimize the interview preparation.
Further preparation consisted of going through the interview protocol to make sure all questions would be understandable for and applicable to the interviewee. For the interviewees who were asked about the development of INFORM Subnational specifically, questions had to be posed in a slightly different way to be able to draw similarities with the process of scaling up the Community Risk Assessment and Prioritization toolbox (also specified in *Appendix D*) and to have the interviewees understand the questions. The content of the questions remained the same.
The interviews were chosen to be *semi-structured,* which means that the interviews consist of different key questions that help to delineate the areas to be explored (as done in the interview protocol, *Appendix D*). However, it is also allowed for the interviewer or interviewee to diverse from these questions to elaborate on an idea or response in more detail. This approach provides flexibility, especially compared to structured interviews, and makes it possible to obtain in-depth, detailed and rich information. Additionally, it facilitates the discovery of information that can be important for the interviewer, but may not have previously been thought of. At the same time, the semi-structure provides guidance, makes it possible to compare different interviews and ensures that all important topics are covered (Britten, 1999; Gill et al., 2008).
All interviews but one were conducted over Skype, due to the interviewer and interviewee not being in the same country (see *Table 7*). The interviews started with a description of the background of the researcher, the study being conducted and a set of questions on the background of the interviewee. Next, all the substantive questions (see *Appendix D*) were posed and discussed, in the way described in *paragraph 5.3.1*. At the end of the interview, the interviewees were given the opportunity to raise topics or issues that had not come forward in the interview, but were considered relevant by the interviewee. Finally, they were asked if they knew of other potential interviewees, and thanked for their help.
Three of the interviews were conducted in Dutch and the rest in English. The Dutch interviews were translated immediately after the interviews took place. The interviews were not recorded, but the typing skills of the interviewer allowed for an almost complete type-along with the interview answers. Sentences that were not documented well, were complemented directly after the interview.

5.3.4. Processing of interview outcomes

After the interviews, the interview transcripts were structured per discussed governance element. When all the interviews had been done, per discussed aspect a document was created with the answers of all interviewees. In this document, the answers were structured in a table to create a clear overview. This allowed for a comparison of all the answers, in which differences and similarities in answers, ideas and trends could be highlighted with colors. This document was used to create the storyline per element to be discussed, which is done in the next section.

5.4. Empirical outcomes: a governance approach

In this section, the main empirical findings resulting from the interviews are presented. The detailed storyline can be found in *Appendix E.*

5.4.1. 510 Global toolbox vs. INFORM Subnational

As described in *Section 5.3.2*, three of the eight interviewees were familiar with the Community Risk Assessment and Prioritization toolbox, and were questioned about the process of scaling it up. The remaining five interviewees were asked about their experience with rolling out INFORM Subnational. As these processes pursue the same objective, are of a very similar nature and expected to comprise identical steps, they are assumed to be interchangeable. Hence no distinction is made between answers from interviewees focusing on scaling up the toolbox, and interviewees focusing on rolling out INFORM Subnational: either two will be further referred to as 'the model'.

The decision not to distinguish between the two was further supported by the fact that the answers of both the toolbox and INFORM interviewees did not contradict, but only complemented each other. The INFORM interviewees knew that the objective was to retrieve generic best-practices from a governance perspective on how to scale up the Community Risk Assessment and Prioritization Index; the toolbox functioning, input and output were explained to them and they all saw the resemblance with the INFORM Subnational model they rolled out.

5.4.2. Genericity of empirical findings

The aim of this research is to come to an approach for rolling out the Community Risk Assessment and Prioritization toolbox that is generically applicable, i.e. not country-specific. Prior to carrying out the interviews it was unclear how generic this approach could become, due to country-specific characteristics possibly requiring different approaches. The answers of the interviewees did illustrate the variety of data landscapes in the cases assessed, in terms of maturity, capacities and attitude. The interviewees repeatedly stressed the presence of these differences. Yet, despite this great variation, all interviewees proposed similar approaches on how to deal with the different governance elements, irrespective of a country's current level of data maturity. The main, generic, empirical findings are presented in *Table 8* below.

5.4.3. Main empirical findings per governance element

In *Table 8* below, the main empirical findings are structured per governance element. The right column shows the numbers of the interviewees (as shown in *Table 7*) that support the

corresponding finding. A complete description of the empirical findings, including illustrations and quotations, can be found in *Appendix E*.

CRITERIA	MAIN EMPIRICAL FINDINGS	INT
• LEADERSHIP • DIVISION OF ROLES AND RESPONSIBILITIES • SUSTAINABLE FINANCING	**Empirical findings: how to arrange or establish these criteria?** The interviewees identified four different roles to be identified and assigned in the process of scaling up the Community Risk Assessment and Prioritization toolbox, of which the first two have a leading role:	
	1. The initiator/coordinator *Role and responsibilities:* • Leading agency that should initiate, coordinate and support the development of the model in a country, and should bring all relevant actors together and incentivize them to facilitate data-sharing and to use the model. • Responsible for the capacity-building of the local lead and the data providers. • Responsible for tailoring the model to the national context and the user needs. • Financially responsible (either through own resources or through funding requests)	*All* *1,5, 6,8* *1,5-8* *1,7*
	Identification: United Nations agencies: • United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA), coordinating humanitarian action in partnership with national and international actors (UNOCHA, 2017). • United Nations Children's Fund (UNICEF), promoting the rights and well-being of children (UNICEF, 2017). • United Nations Development Program (UNDP), working to eradicate poverty and reduce inequalities through the sustainable development of nations (UNDP, 2017).	*All*
	Reason for suitability: • These organizations generally have a structural presence and strong position in countries. • Information management is one of their core activities, which means they should have the capacity (see *participatory capacity* element) • The role fits their organizational responsibility (see *incentivization element),* and they have the ability to bring different organizations together.	*All* *All* *5-8*
	2. The local lead *Role and responsibilities:* • Within a country, the model will be used for national purposes. To create local buy-in, it is therefore important that there is a local party that feels ownership of the model, that understands the value and potential of it, that can locally promote it and stimulate the use, and connect it to existing local initiatives. • This party becomes the local manager after the initiator/coordinator has set out the initial model development (development being the filling of the model with data). • This role is important to ensure the sustainability and continuity of the model.	*1,3,5, 6,8* *1,3,5, 6,8* *3-5*

CRITERIA	MAIN EMPIRICAL FINDINGS	INT

	Identification:
	• A government agency (or national body)
	• Most suitable government agency to take up this role varies per country, depending on the responsibilities and capacities of the different government bodies.
	• The decision of the most suitable local lead within a country is a process. It should be determined in agreement with all government agencies that could potentially fulfill this role (relevant ministries, disaster departments etc.)
	
	Reason for suitability:
	• The model outcomes are expected to be used mainly by government agencies. Also, many of the data owners are governmental organizations. Having another government agency as the local lead could therefore help to involve those parties.
	
	Additional remarks:
	• Five of the eight interviewees identified the role of 'local lead'. The remaining interviewees did not necessarily exclude this role, but they envisioned the responsibilities of the initiator/coordinator and the local lead in one role.
	
	3. Data providers
	Role and responsibilities:
	To share their data so it can serve as an input for the model.
	
	Identification:
	• Getting a good overview of relevant data providers in a country is a time-consuming process
	  ○ Step 1: Try to get a basic preliminary overview of data providers and data available by searching on open data platforms.
	  ▪ But beware that these platforms could be misleading: if there is lots of data, but with a very poor quality, it is still useless.
	  ○ Step 2: From there, start approaching the main data providers in a country (often the National Statistics Office (NSO), a type of disaster department, relevant ministries, a census organization). Organize a workshop with them and ask them to identify key data providers, existing data initiatives, platforms and data collection workgroups.
	
	Reason for suitability:
	Natural role: if organizations own relevant data, they get assigned this role.
	
	4. Model users
	Role and responsibilities:
	To use the model and incorporate the model outputs in their operations.
	
	Identification:
	• There is a wide range of potential users, as the model can be used for many different purposes, like resilience, early warning, early recovery and risk management. Most potential users can be divided
		1,3,4, 5,6,8
		
		5,6
		
		1,3,5, 6,8
		
		All
		
		All
		
		1,2,3, 5,6,8
		
		4,8
		
		1,2,3, 5,6,8
		
		All

CRITERIA	MAIN EMPIRICAL FINDINGS	INT
	into two categories: humanitarian organizations and government agencies. • There is a duplication of roles between the model users and the other identified roles: most of the identified users are also identified as being the most suitable party for taking up one of the other roles. Sometimes government agencies can even have a triple role: local lead, model user and data provider. ○ This duplication of roles provides potential for the incentivization of parties to take up one of the other roles: if a party wants to become a user of the model, they should also be willing to accept this other role. *Reason for suitability:* Natural role: if an organization's activities match the model output, they become a suitable model user.	

INCENTIVIZATION	**Empirical findings: how to arrange or establish these criteria?**
	After having identified suitable parties for the different roles, they must be incentivized to actively take up this role and the corresponding responsibilities. There are different incentives for each role and identified party.
	
	1. Initiator/coordinator
	• According to the interviewees, the incentives for UN agencies to take up the role as initiator/coordinator mostly come naturally. The following quotations illustrate this:
	  ○ "*There is simply not enough time to provide all aid that is necessary, so it is essential to start making prioritizations. It is important to base these prioritizations on hard numbers rather than on assumptions. For UNOCHA, it is just morally the right thing to do to pursue this objective in the humanitarian sector, and to promote and improve evidence-based humanitarian assistance" (Interviewee 7, personal communication, August 10, 2017)*
	  ○ "*it is a fact that in the world of risk management you need information, you need data*" (Interviewee 5 and 6, personal communication, August 10, 2017).
	• However, there must be budget available, and it helps if there is a clear necessity for the model development, like the regular presence of crises in a region.
	
	2. Local lead
	• To become willing to take up the role of local lead, a party must be convinced of the value and sustainability of the model, and of how it can help them to pursue their objectives. They should start to feel responsible for the model development.
	• A way to achieve this, is for the initiator/coordinator to decide on the most suitable local lead in the very beginning of the development process, and to completely involve this party in the model development. It should be avoided to begin with the model development and then in a later stage start looking for the local buy-in.
	  ○ An issue that came up here, is that it becomes more attractive for parties to become involved in the model, once
		5-8
		
		5-7
		
		1,2.3, 5,6,8
		
		1,2.3, 5,6,8
		
		3,5, 6,8

it contains more data. However, to get more data, parties must become involved. This is a vicious circle that needs to be broken (see *Figure 19* below).

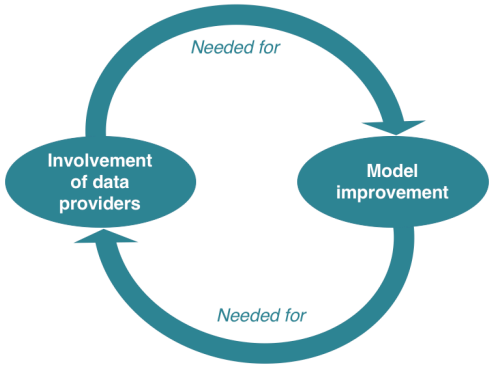


FIGURE 19 – Vicious circle

- To break this vicious circle, it is proposed to take a stepwise approach: start developing the model as far as possible with an organization that is not the most ideal local lead, but is willing to participate or easy to connect to, like the National Red Cross Society in case of 510 Global. Hopefully this leads to model improvements, which then attracts other organizations to start participating. *4*

- Other ways to incentivize a potential local lead to take up its role would be to explain them that it can also help them to meet other objectives, like meeting global agreements like the SENDAI Framework, and to make clear that it has been made relatively easy for them. As illustrated by the following interviewees:
  - *"You must make clear that being the local lead for INFORM Subnational in the end does not cost them work, but it will improve their work, once they have everything embedded in their systems and processes. The challenge is, however, to have them realize this"* (Interviewee 8, personal communication, August 11, 2017).
  - "*Also, we make it easy for them: we hand them the complete model, and they only have to manage and maintain it, which are processes that are quite straightforward and relatively easy. Yes, it requires some work, but eventually you need the information anyway"* (Interviewee 5 and 6, personal communication, August 10, 2017).

3. Data providers

- The willingness to share data varies greatly per data provider, even between data providers within the government. Data owners mainly have incentives *not* to share their data, and these should be removed. The following difficulties were mentioned in the process of requesting data (for a more elaborate explanation, see *Appendix E)*: *All*
  - They say 'of course we will share it' but then do not share it after all *2,3*
  - Having to go through a bureaucratic hassle *2,3*
  - They share data in an unworkable format *3*
  - They are protective of their data due to its privacy sensitivity. *2-6*

CRITERIA	MAIN EMPIRICAL FINDINGS	INT
	○ The person in charge of the data provider has a closed attitude	*3,5,6*
	○ They are afraid that sharing their data reveals unpleasing things about their work	*5,6*
	○ They feel that it will cost theme extra work, whereas there is nothing in it for them (no priority)	*All*
	• The main approach to incentivize data providers would be also to generate a sense of ownership and involve all data providers from the very beginning of the model development.	*All*
	○ By engaging them in a workshop, transparency can be provided and data providers can be convinced of the model value.	
	○ It should be agreed what is in it for them. This should be something concrete, they do not just want to be acknowledged, but they also want for example a specific data analysis they need.	*2,3, 4,6*
	○ *Interviewee 8* illustrates: "*If the data providers eventually want to benefit from the model you create, they must realize they have no other choice than sharing their data with you, it is as simple as that*" (Interviewee 8, personal communication, August 11, 2017).	
	○ Other, specific tips to incentivize data providers can be found in *Appendix E.*	
	• If data providers remain unwilling to share data, it is important to remain flexible: find ways to work around them, keep looking for alternatives, work with what you have and from there improvements can be made. It is an iterative process.	*1,3,5, 6,8*
	4. Model users	
	• Just like for the local leads and the data providers, also for the model users holds that to incentivize them to start using the model, they must be involved in the process of developing the model from the very beginning.	*1,2,3, 6,8*
	○ This is important to have them understand the value of the model, but also to have them give input on the data to include in the model: how can the model be adjusted to the local context? What do the users want to do with it?	*1,2,3, 4,6,8*
	○ This should be done in consultation with the data providers, as the data proposed to be included, should also be available and accessible.	*1,8*
	• User-friendliness is also key in the incentivization of potential users. The presentation of the model output should be tailored to the needs of different users. Additionally, user-friendliness plays a role in meeting the capacity of the users (see *'Participatory capacity'*).	*2,4,5, 6,7,8*
PARTICIPATORY CAPACITY	**Empirical findings: how to arrange or establish these criteria?** Apart from having the incentives to take up their assigned role, the identified parties should also have the capacity to take up their role. If they do not have this capacity and there is no suitable alternative party that does have the capacity, this capacity needs to be built.	
	1. Initiator/coordinator	
	• UN agencies, especially regional hubs, generally have both the organizational capacity (ability to connect organizations and take the lead) and the technical capacity (required to be responsible for the model development) to take up the role of initiator/coordinator.	*5,6 7,8*

CRITERIA	MAIN EMPIRICAL FINDINGS	INT
	2. Local lead • In the process of identifying the most suitable local lead, the presence of data and information management capacities is an important selection criterion. Nevertheless, the local lead should be able to both manage and maintain the model.	*All*
	• Another selection criterion in this process is whether the model outcomes match the activities of a governance agency. However, if there is a perfect match, but this agency does not have the capacity, they are still not considered suitable.	*5,6,3*
	• It is very country-dependent which organization has the required capacity.	*All*
	• If needed, the local lead should be trained on for example excel skills and mapping by the initiator/coordinator. In any case, the initiator/coordinator should keep offering support to the local lead, in the form of technical support or any other consultancy, in case for example issues with the maintenance arise	*1,5, 6,8*
	3. Data providers • Data providers must have the technical capacity to share their data, but they must also have the information management capacities required to adequately collect and store their data.	*All*
	• Within countries these capacities very greatly between data providers, and also between administrative levels: at lower administrative levels, organizations usually have less technical and information management capacities. Information might even be stored paper-based.	*All*
	• The NSO usually has the required capacity, so this is an important party to always involve.	*All*
	• Data providers that do not have the required capacity should be given trainings/workshops by the initiator/coordinator. In these trainings, their technical staff should be addressed.	*2,3,5, 6,7,8*
	• However, training alone is not enough: a shift in organizational culture and in mentality is required to make the capacity-building needed for adequate data-sharing a higher priority.	
	• An additional problem in the Philippines and Malawi is that most qualified information managers get hired by private companies who pay better.	*2,3*
	• To deal with these issues, again the key is to involve data providers from the start in the model development process: Once they see and understand the value of more data-driven operations, they might become more motivated to put effort in improving their capacities. Nevertheless, realizing this can be very difficult. As *Interviewee 3* illustrates: ○ *"You can of course train people, but effectively doing so is easier said than done. Even if the required technology is available and people know how to use excel, they still do not use it in a correct way. They are not sufficiently motivated, or do not have the right mentality. How to change this? That is a very good question: I have no idea"* (Interviewee 3, personal communication, August 1, 2017).	

CRITERIA	MAIN EMPIRICAL FINDINGS	INT
	4. Model users • In the first place, the initiator/coordinator must make sure that the capacities of the model users are met by presenting the model output in a user-friendly way. • Secondly, to ensure that the model outputs will be structurally integrated in an organization's operations, a shift in mentality of the user organizations is required. Again, the best suggested approach related to this shift in mentality is involvement from the start. *Interviewee 5* and *6* illustrate: *"In Colombia, we thought that once we would have developed this tool that, everyone would immediately start using it. Unfortunately, that was not the reality. You must develop a culture around it, around its use and integration in organizational practices* (Interviewee 5 and 6, personal communication, August 10, 2017).	*2,4,5, 6,7,8* *5,6*
VISION, COMMUNICATION AND STIMULATION	**Empirical findings: how to arrange or establish these criteria?** • One of the aspects that is part of this element is 'to avoid a fragmented approach'. Related to this, the interviewees were questioned about the presence of a national open data platform in their country, who set up this platform, and how it could be improved. ○ The interviewees did acknowledge the value of well-organized open data platforms, but none of them was very fond of the quality of the open data platforms in their region. They are often fragmented and it is difficult to find the data you need, especially of high quality. Also, much data is not being shared on these platforms. ○ Other than the capacity-building of data provider, which has already been covered, the interviewees did not have suggestions on how to improve or establish a national data platform. • The importance of the arrangement of this element also followed from the findings on the elements previously discussed. One of the main findings on the incentivization of actors in the ecosystem to take up their role is that it is important to involve them from the start of the model development process, to provide a shared goal. This relates to the importance of the provision of a common goal and the prevention of fragmentation to occur, which means the interviewees basically recommend meeting the 'vision, communication and stimulation'.	*All*
• TRUST AND TRANSPARENCY **• *FEEDBACK***	**Empirical findings: how to arrange or establish these criteria?** • The interviewees generally found it difficult to get transparency of data providers on where they got their data from, and to provide feedback on the (quality of the) data they shared, but the possibilities vary per country and per organization. ○ *Interviewee 2* for example mentioned that getting transparency and providing feedback is rather impossible in Malawi, as it is already sufficiently challenging to get organizations to share their data in the first place, but *Interviewee 3* stated that in the Philippines it is difficult, but not impossible.	

CRITERIA	MAIN EMPIRICAL FINDINGS	INT

	• All interviewees think that there is a lack of transparency from data providers and feedback possibilities, which results in not having insight in the quality of the data.
	  ○ However, this is something that they seem to have accepted: it is difficult and time-consuming enough to enable and stimulate data-sharing in the first place; feedback is of later concern.
	  ○ If certain data cannot be found, the interviewees rather look for alternatives (e.g. proxy-indicators) than putting effort in trying to improve the data of other data providers through feedback mechanisms.
	• These comments all concern transparency of fata providers on their data collection processes. Transparency from the initiator/coordinator towards the other involved parties on the model functioning and output is considered very important.
	• Feedback from the model users to the initiator/coordinator or local lead is also considered very important, but also becomes relevant in a later stage of the model development (once it has already been adopted).	

TABLE 8 – Main empirical findings per governance element

5.4.4. ADDITIONAL FINDINGS

- **On the timeline of the involvement of different roles:** The process of rolling out the model in a country starts with the initiator/coordinator, who needs to involve potential local leads, data providers and model users very early in the development process, and is responsible for building the required capacity and tailoring the model to the national context. During this time, also a suitable local lead that is willing to take up this role must be identified by the initiator/coordinator in a process of mutual agreement.

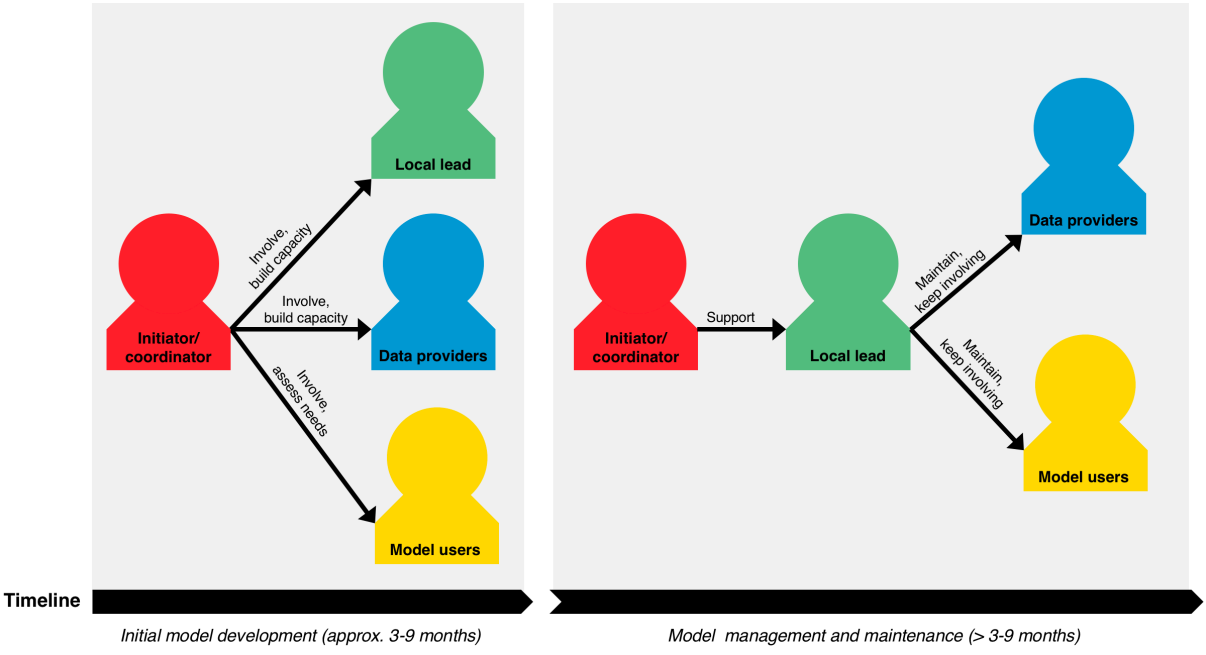


FIGURE 20 – Timeline of the involvement of and relationship between different roles

This initial development of the model takes approximately three to nine months, but it must be noted that this duration varies greatly per country and it may also take longer. The phase ends once a local lead has been identified that has sufficient capacity (or for which sufficient capacity has been built) to take up the local lead responsibilities of managing and maintaining the model. This transition might be more gradually than visualized in *Figure 20*.

During the model management and maintenance phase, the initiator/coordinator should still provide support to the local lead where necessary. The need for this support is expected to decrease over time, as the local lead becomes more and more experienced.

- **On the role of National Red Cross Societies:** Since 510 Global is connected to the Red Cross and their network, it would be easiest for them to have a Red Cross National Society as the initiator/coordinator. However, this party is considered less suitable to take up this role:
  - Their information management capacities are usually not very advanced. UN agencies are ahead of them in terms of these capacities, being it one of their core activities.
  - They are generally less structurally present in countries, because they work on a project basis. Once a project is over they withdraw their resources.
  - This also relates to the visibility of the organization: A Red Cross National Society is active in times of disasters. If a country is less prone to disasters, this results in less visibility for the Red Cross. This is the case in Malawi, whereas in the Philippines the Red Cross is well established due to lots of natural disasters occurring there.

  Nevertheless, National Red Cross Societies could play a role when a stepwise approach as proposed by *Interviewee 4* will be adopted, to break the vicious circle and already start with the model development to be able to attract new partners, including a more suitable initiator/coordinator.

- **On the role of INFORM and 510 Global:** The actors identified by the interviewees all play a key role when rolling out models like INFORM Subnational or the Community Risk Assessment and Prioritization toolbox in a country. However, an important question for 510 Global is how they should be positioned relative to these parties: what is their role in the governance process? To get an idea of how INFORM approaches this, additional questions were posed to *Interviewee 1*, who globally coordinates the development of INFORM models. This led to the following findings on the positioning of INFORM in the process of rolling out INFORM Subnational:
  - The decision of INFORM to start developing subnational models, was driven bottom-up: they received a lot of request from UN agencies with the question whether they could start developing INFORM Subnational for their region. This is still the case: INFORM does not need to take an active role by approaching possible initiators/coordinators, but they wait for organizations to approach them, and this occurs regularly – at least often enough for them not to feel the need to also develop a top-down strategy.

- In the development process, INFORM provides different types of support to the initiator/coordinator:
    - Support at a technical level
    - Support for funding (by means of using their network- INFORM is not a legal entity so they cannot provide funding themselves)
    - Support in the form of providing consultants that can help guiding the process
    - Support in the form of training the initiator/coordinator to work with the model
- What 510 Global could do with this information and how they can use it in the determination of their positioning, will be discussed in *Section 6.4.1 – Recommendations for 510 Global.*

- **On the identification of additional governance elements:** All interviewees were asked whether there were aspects that had remained unaddressed during the interview, of which they thought they would be interesting to address. None of the interviewees identified additional elements to be of interest.

5.5. EMPIRICAL VALIDATION OF GOVERNANCE PART OF THEORETICAL FRAMEWORK

The previous sections described the practical contribution of this research step, but, as described in the introduction of this chapter, this research step also led to an addition to the theoretical insights: it allowed for the empirical validation of the governance part of the theoretical framework of data ecosystem success criteria, which will be discussed in this section.

Chapter 4 already addressed the application in practice of the other two framework parts: the 'Data supply' and 'User characteristics' elements. No difficulties as to the content of these framework elements arose when doing so. There appeared to be a clear demarcation between the two categories, no overlap was thought to be present and no additional conditions were identified that could not be attributed to one of the elements.

This did not hold for the governance part of the framework. During its application in practice in this chapter, more points for discussion emerged.

In general, it was found that the governance elements proved to be useful. They formed a solid foundation for the interview questions, and provided good guidance. All relevant aspects seem to have been covered by the framework elements, the interviewees came up with best-practices or approaches that were in line with the ecosystem elements, and the interviewees did not introduce any additional elements; it was asked to them whether they had the feeling certain important aspects had not been covered, but none of the interviewees felt the need to add something.

Some of the elements and their relationship with the other ecosystem elements however, should be subject to further reflection. Some of the points for reflection are considered to be general, which means that it is proposed to adjust them in the theoretical framework (*Section 5.5.1 – Suggestions for framework adjustments*). Other points for reflection are more case-specific, and should therefore not lead to general adjustments, but rather say something about

the application of the framework in the humanitarian context (*Section 5.5.2 – Evaluation of framework applicability in humanitarian context*).

5.5.1. Suggestions for framework adjustments on the governance side

- **Separate 'Leadership and incentivization':** The 'Leadership and incentivization' element should be split up into two individual elements. They both encompass different aspects and need a separate approach. This was found to be the case when applying this element to the humanitarian context of scaling up the toolbox, but it is expected that it also holds for other data ecosystems.

- **Remove 'Value creation':** The elements 'Incentivization' and 'Value creation' overlap. Value creation is a means to incentivize parties. This was also something that popped up frequently during the interviews: in the process of incentivizing parties to take up the role considered most suitable for them, they need to be convinced of the value that the model has for them. Having the two as separate elements would thus lead to redundancy, which is why 'value creation' should be removed.

- **Merge 'Transparency and feedback' and remove 'Trust':** The elements 'Feedback' and 'Trust and transparency' seemed to be interwoven: by providing feedback and being open to it, parties are being transparent. This was confirmed when carrying out the interviews. Additionally, it was found that 'Transparency and feedback' should be merged into one element, as 'trust' was noticed to cover something else: following the 'Vision, communication and stimulation' element (establish a collaborative environment in which a common goal is provided, to avoid a fragmented approach) was identified by the interviewees as a means to generate trust. Trust is also a concept that cannot be forced into existence. It is therefore suggested to remove 'trust' from the framework, assuming it will develop once effort is being put in meeting the 'Vision, communication and stimulation' element.

- **Replace 'Sustainability' for 'Sustainable financing':** 'Sustainability' is indeed an issue that needs to be fulfilled, but according to the interviewees this is more of a result from other elements. It is not an element that *needs* to be in place to establish a successful data ecosystem, but something that results from a successful ecosystem: it is a characteristic of a successful data ecosystem, not a condition. Nevertheless, 'sustainable financing' is something that can also contribute to this sustainability and to a successful data ecosystem. It is therefore suggested to replace the 'sustainability' element for the 'sustainable financing' element.

These suggestions based on the practical application of the governance part of the theoretical framework have led to the following suggested adjustments to be included in the general framework (the figure outlines highlight the adjusted elements):

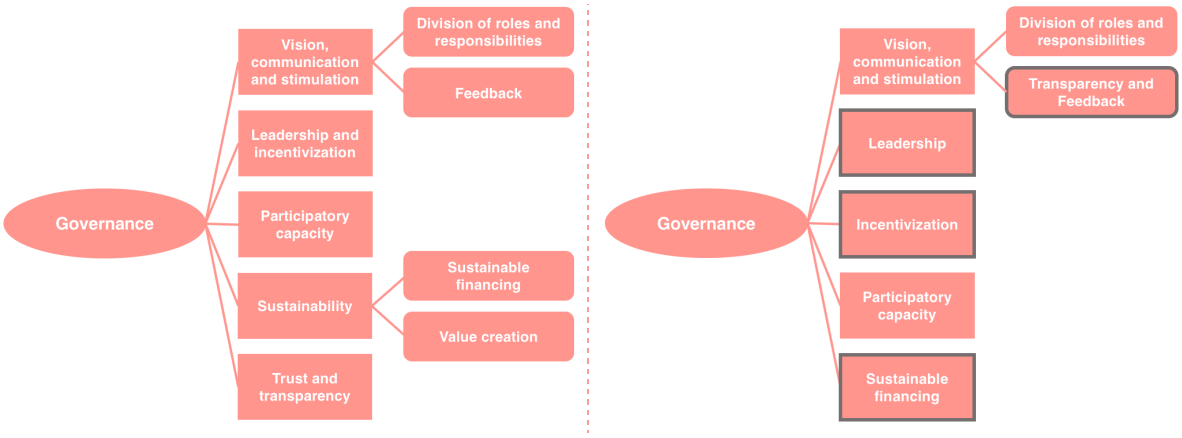


FIGURE 21 – Left: governance elements before empirical validation
Right: governance elements after empirical validation

5.5.2. EVALUATION OF THE APPLICATION OF THE GOVERNANCE PART OF THE FRAMEWORK IN THE HUMANITARIAN DATA ECOSYSTEM

- **Interrelations between different framework parts:** Applying the governance part of the framework in practice, revealed that there are interrelations between the different framework parts (or categories) that are more complex than visualized in the framework (*Figure 15*). Several insights retrieved during the interviews illustrate this.
  The element 'division of roles and responsibilities' was in *Section 5.2.2* translated to the practical context of the Community Risk Assessment and Prioritization toolbox as 'to get an overview of data providers', as the different data providers need to be identified. It was specifically assigned to the role of data provider, in the assumption that the other roles were already covered by other parts of the framework, like 'leadership' and 'user characteristics'. Accordingly, 'incentivization' and 'participatory capacity' were in the initial translation to practice only considered for the data providers. However, this is where some blurry lines appear:
  The elements 'leadership' and 'user characteristics' also relate to different roles. Why would the 'division of roles and responsibilities' then not cover these roles as well? This element gives the impression that it includes the identification and assignment of *all* different roles in the data ecosystem, including the one of user and of leader (in this case the initiator/coordinator and local lead). The problem of the 'leadership' element being a part of the 'division of roles and responsibilities' element can relatively easily be solved by visualizing the elements in a different way. However, when making the assignment of the users part of the 'division of roles and responsibilities' the demarcation between the different framework categories starts to become vague. Yet, similarly it can also be argued that the identification of data providers is part of the 'data supply' part of the framework, relating to data availability.
  Nevertheless, it was determined that the elements 'incentivization' and 'participatory capacity' were relevant to assess for each identified role. This added a new complexity though, because now for the user role, the 'participatory capacity' assessment would overlap with what comprised the 'user capabilities' element: what would be the distinction between the two?

On top of this, the empirical findings showed that actors in the humanitarian data ecosystem (in the process of the model development) can take up multiple roles, which creates an extra issue of the demarcation between the different framework categories: if users can also be data providers can also be local leads and all relate to a different framework category, this results in all categories interrelating.

Lastly, providing feedback and being transparent ('Transparency and feedback') are in the context of the Community Risk Assessment and Prioritization toolbox (or INFORM Subnational) interpreted as relating to the data quality, from the 'data supply' category, which again implies a certain overlap between the categories.

To summarize this, in the humanitarian context, there does not seem to be a clear distinction between the three data ecosystem framework categories. When only assessing the data supply and the user characteristics part there seem to be clear boundaries between the two categories, but when starting to involve the governance elements, many of the elements (and framework categories) start to interrelate. This could be because in this specific context, the 'data supply' and 'user characteristics' categories are related to the content of the Community Risk Assessment and Prioritization toolbox – the object that needs to be developed and implemented –, whereas the 'governance' category addresses the elements to consider when carrying out this development and implementation. It could be an option to position the 'governance' category differently in the framework, outside of the loop with the other two categories, but this would make the framework specific for the situation assessed in this study. Moreover, it would not provide a solution for the interrelation between the categories and elements. This overlap due to the duplication of roles could also be considered an illustration of the context analyzed in this study being too complex to capture in this theoretical framework. To assess whether this is unique for the humanitarian data ecosystem or holds for other ecosystems as well, the framework should be applied to other case studies (see *Section 6.4.2 – Recommendations for future research*).

- **The relevance of certain governance elements changes over time:** In *Section 4.2.1* it was explained that some of the data supply elements only become relevant in a more advanced (mature) phase of a data ecosystem. Based on the empirical findings, it could be stated that this also holds for some of the governance elements: where the discussion of certain elements happened almost naturally (of course the interview protocol was there to steer the questions and provide direction, but still the conversation can naturally lead to the discussion of certain topics), some effort was needed to bring the topics transparency and feedback to table. They seemed to be considered less relevant by the interviewees; they do not seem to fit the phase the humanitarian data ecosystem is now in. It is expected that this element only become relevant in a later, more developed stage of the data ecosystem. In the current humanitarian data ecosystems, there is a stronger focus on 'be happy with what you get', rather than asking for more specifications and providing feedback, possible because the capacities of the data providers do not allow for this (if at all for data sharing in the first place).

  This again illustrates that the humanitarian data ecosystem is still very much in its infancy, and that the concept of data ecosystems is dynamic of nature. To assess the

change in relevance of data ecosystem elements for data ecosystems other than the humanitarian one, the framework should be applied to other case studies as well (see *Section 6.4.2 – Recommendations for future research*).

5.6. CHAPTER CONCLUSION

In this chapter, the following question has been answered:

SQ4: HOW SHOULD THE GOVERNANCE OF THE DATA ECOSYSTEM AROUND THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX TO FACILITATE THE PROCESS OF SCALING UP THE TOOLBOX BE APPROACHED?

To answer this question, eight different humanitarian data experts were interviewed that represent five different regions. These experts all had extensive knowledge on the data landscape in their region, and were familiar with either the Community Risk Assessment and Prioritization toolbox, or the development of INFORM Subnational, a risk index with features comparable to the toolbox. The interviewees were asked about the relevance and presence of the governance elements of the data ecosystem framework in their region, and on how they would propose to arrange them in such a way to create an optimal context for rolling out the Community Risk Assessment and Prioritization toolbox.
The answers of the interviewees illustrated the variety of data landscapes in the cases assessed. Yet, despite this great variation, all interviewees proposed similar approaches on how to deal with the governance elements. It was stressed repeatedly that there are many country-specific characteristics, but the overall takeaways came down to the same core aspects: the same processes to incentivize stakeholders, the same difficulties when creating an overview of data providers and data available, the same processes of identifying the local lead.
The main empirical findings are centered around the identification of four different roles ('division of roles and responsibilities' and 'leadership'), their 'incentivization' and the creation of their 'participatory capacity'. These are summarized in *Figure 22* on the next page.

Moreover:

- For all roles, a shift in mentality and organizational culture is required to move towards more data-driven operations.
- To address this, and other issues brought forward, it is important to involve all relevant stakeholders from the very start of the model development, to create ownership and have them realize the added value. This holds for all roles (but note that the roles do overlap), and is related to the 'Vision, communication and stimulation' element.
- Every case is unique: the generic approach should be tailored to each country-specific situation. Additionally, there are many differences between organizations and administrative levels within one country as well, which illustrates that every assessment should be made on a case-by-case basis.
- It is important to remain flexible during the process, and to focus on having an iterative process. If certain data cannot be found or made available, alternative should be sought for (like proxy-indicators) rather than putting lots of effort into something slow-moving.
- Related to this, is that all interviews stress that the development of the model is a time-consuming process that takes effort and dedication.

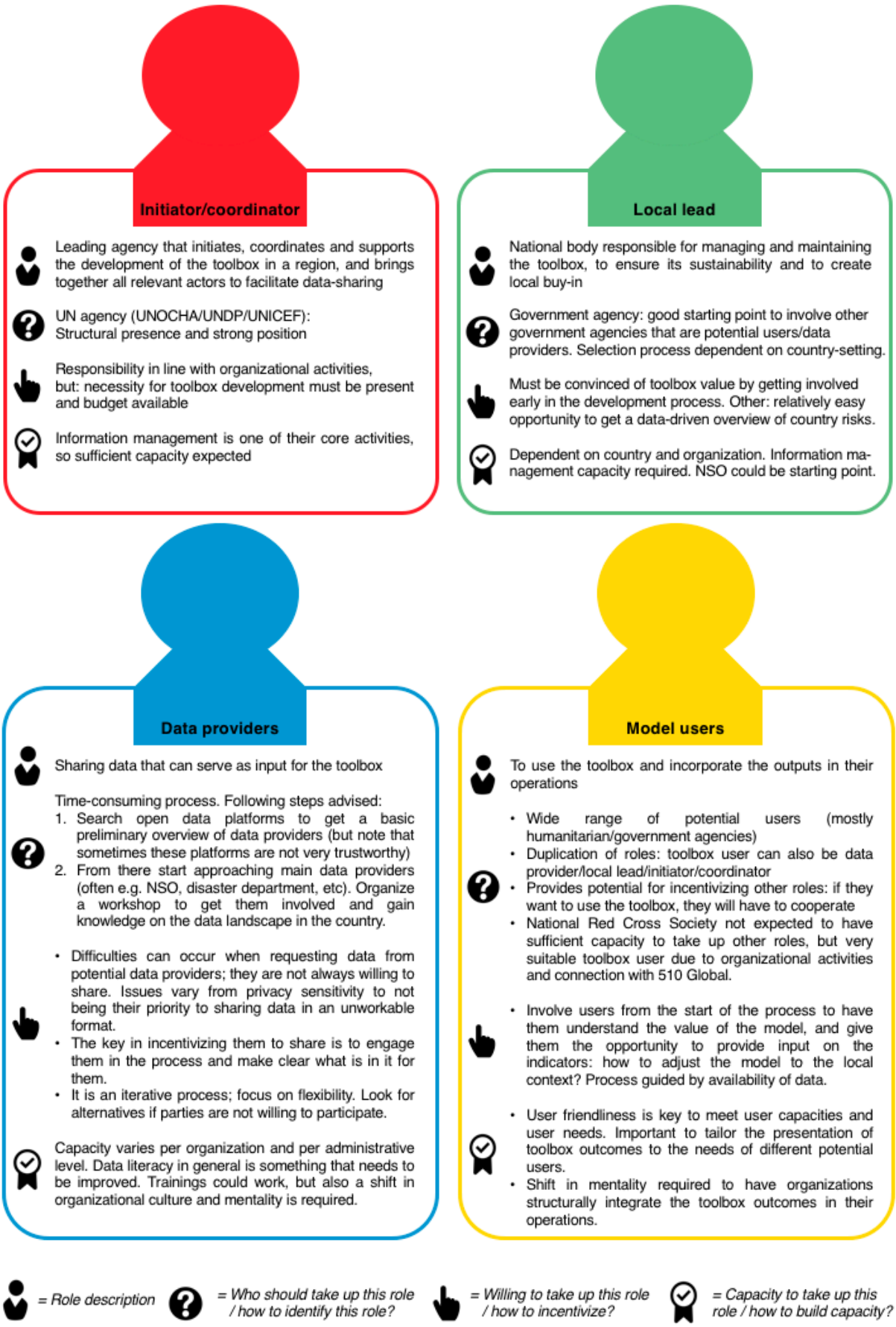


FIGURE 22 – Identified roles and the main empirical findings per role

As to the theoretical contribution of this chapter, the application of the governance part of the framework allowed for an empirical validation that led to the following adjustments:

- To separate ‘Leadership and incentivization’
- To remove ‘Value creation’
- To merge ‘Transparency and feedback’ and remove ‘Trust’
- To replace ‘Sustainability’ for ‘Sustainable financing’:

When incorporating these changes, the empirically validated theoretical framework looks as follows:

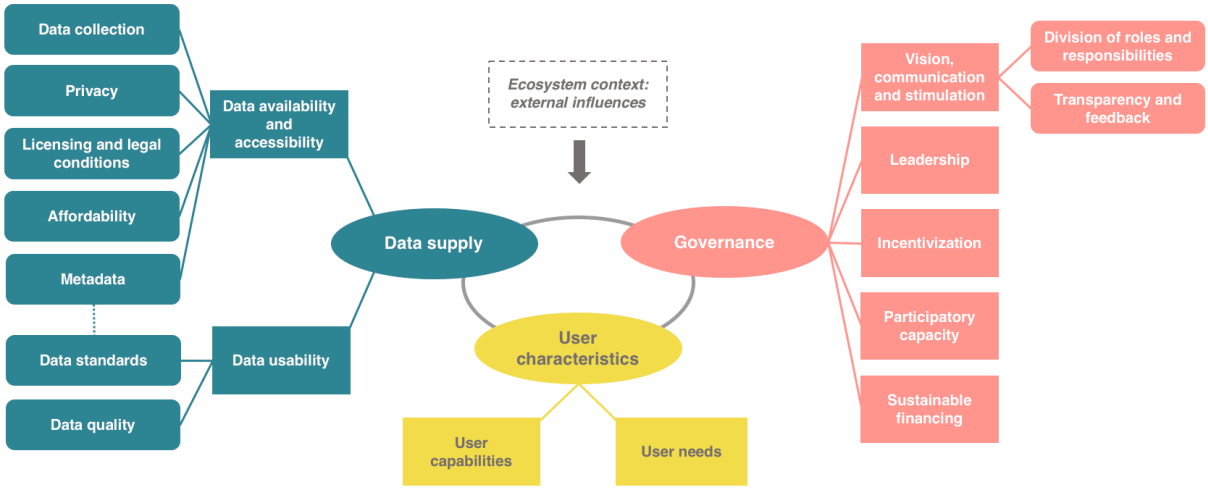


FIGURE 23 – Empirically validated theoretical framework

Furthermore, the application of the governance part of the theoretical framework in practice revealed that there are interrelations between the different framework parts (or categories) that are more complex than visualized in the framework. Also, it was found that the ‘transparency and feedback’ element is not considered highly relevant in the current phase of the humanitarian data ecosystem. It is difficult to get transparency from data providers on their data collection and to provide feedback on the quality of their data; currently it is already challenging enough to get organizations to share their data in the first place. However, it is expected that the importance of this element increases as the ecosystem matures.
To assess whether these findings are unique for the humanitarian data ecosystem or also hold for other data ecosystems, the framework should be tested for other case studies as well (see *Section 6.4.2 – Recommendations for future research*).

CHAPTER 6 – CONCLUSION, DISCUSSION AND RECOMMENDATIONS

In *Section 6.1,* we will give an answer to the main research question of this study, which also reflects the societal contribution. *Section 6.2* describes the additional scientific contribution. Then, in *Section 6.3,* we will discuss and reflect on the research design and the choices made in this study. *Section 6.4* will provide recommendations; both practical recommendations for 510 Global on how to continue with these research outcomes, and recommendations for future research. Lastly, in *Section 6.5* the link between this research and the CoSEM Master's program will be described.

6.1. MAIN CONCLUSION OF RESEARCH

The Community Risk Assessment and Prioritization toolbox is a model in which different types of data on risk components are combined. The outcomes of the model provide a detailed risk assessment for areas and communities in a country, that can help to faster identify priority areas for humanitarian intervention related to natural disasters. Currently the toolbox is under development for three different countries: the Philippines, Malawi and Nepal. If it can be taken to scale across the humanitarian system by making it applicable for many countries, it has the potential to lead to improved, more evidence-based humanitarian decision-making and aid provision.
With this research, the following main question has been answered:

WHAT DOES A GENERIC GOVERNANCE STRUCTURE THAT FACILITATES THE PROCESS OF SCALING UP THE COMMUNITY RISK ASSESSMENT AND PRIORITIZATION TOOLBOX LOOK LIKE?

We showed that the process of scaling up the Community Risk Assessment and Prioritization toolbox can be approached as a socio-technical system, involving both complex physical-technical systems and networks of interdependent actors. The technical complexity around the toolbox is characterized by the need for a rapid collection of disaster-specific hazard and exposure data, and a timely collection of pre-disaster secondary data on vulnerable areas from low administrative levels. A high level of organizational fragmentation, a high number of (potential) parties involved and a lack of coordination contribute to the actor complexity.
A perspective that can be used to deal with this complexity and understand interrelationships between stakeholders is the 'data ecosystems' approach: the people and technologies collecting, handling, and using the data and the interactions between them. A systematic literature review on the criteria for establishing a successful data ecosystem has resulted in the creation of a framework consisting of these criteria. In the framework, three categories of criteria are distinguished: data supply, user characteristics and governance. The framework,

that has been validated in an expert session with 17 open data researchers, has provided insight in the factors that should be considered in the process of scaling up the Community Risk Assessment and Prioritization toolbox.

Firstly, it has helped to show that the humanitarian data ecosystem around the Community Risk Assessment and Prioritization toolbox is still very much in an emerging phase.

There is a strong focus on aspects related to the elements 'data collection', 'affordability' and 'data quality' of the data supply part of the framework, as well as on the 'user capabilities' and the 'user needs' of the user characteristics part of the framework. These are the critical elements that need to be in place for a country to be eligible for the development of the toolbox in the first place. Other elements become relevant in later stages of the toolbox development; their importance increases as the ecosystem evolves and more data-sharing starts to take place. Based on these critical elements, a 'data maturity assessment tool' has been developed, consisting of conditions to be fulfilled – at least to a certain extent – for a country to be eligible for the development of the Community Risk Assessment and Prioritization toolbox. This tool can be used to assess the data landscape in a country, and to get an idea of points of attention or problematic issues.

Secondly, the theoretical data ecosystems framework has been used to create a generic governance structure to realize an optimal context for the toolbox development in a country, by facilitating a positive shift in the data maturity level. The theoretical governance elements formed the basis for a set of interviews conducted with eight humanitarian data experts from regions with widely varying data landscapes. The interviewees provided their insights on how these elements should be arranged in the context of rolling out a model like the Community Risk Assessment and Prioritization in their region, thereby facilitating a scale up. Despite the variety of data landscapes in the cases they represented, they all proposed similar approaches:

CRITERIA	GENERIC APPROACH
• **LEADERSHIP** • **DIVISION OF ROLES AND RESPONSIBILITIES** • **SUSTAINABLE FINANCING**	In the process of developing the Community Risk Assessment and Prioritization toolbox for a country, four different roles should be identified and assigned, of which the first two have a leading role: 1. *The initiator/coordinator:* Leading agency that should initiate, coordinate and support the development of the model in a country, and should bring all relevant actors together and incentivize them to facilitate data-sharing and to use the toolbox. This party is also responsible for the financing of the toolbox, for tailoring the model to the national context and user needs, and for the capacity-building of the local lead and the data providers. The most suitable organization to take up this role is a UN agency, due to their strong and structural position in a country, and their advanced information management capacities. 2. *The local lead:* National body responsible for managing and maintaining (e.g. including new features in and updating the data in) the toolbox, to ensure its sustainability and to create local buy-in. The most suitable organization to take up this role is a government agency, as this provides a good starting point to involve other government agencies that are potential users or data providers. It is dependent on the country setting which specific government agency should take up this role. All potential agencies should be involved in the selection probes to determine the most suitable party.

CRITERIA	GENERIC APPROACH
	3. *Data providers:* Organizations sharing data that can serve as an input for the toolbox. The identification of data providers is a time-consuming process. The following steps are recommended to take: 1. Search open data platforms to get a basic preliminary overview of data providers (but note that sometimes these platforms are not very trustworthy) 2. From there start approaching main data providers (often e.g. NSO, disaster department, etc.). Organize a workshop to get them involved and gain knowledge on the data landscape in the country. 4. *Toolbox users:* Organizations that use the toolbox and incorporate the output in their operations. There is a wide range of potential users for the toolbox, mostly humanitarian and government agencies. Most perceived toolbox users have a double – or triple – role and have also been identified as being the most suitable organization to take up one of the other roles. This provides potential for the incentivization of other roles: if they want to use the toolbox, they must cooperate.
INCENTIVIZATION	Organizations should be involved in the development of the toolbox from the very beginning, to get a sense of ownership and start to feel responsible for the model development. They should be convinced of the value of the model, of 'what is in it for them', and they should be given the opportunity to provide input.
PARTICIPATORY CAPACITY	In the identification process of a suitable lead, capacity is an important criterion, as the elected party needs to have sufficient technical and information management capacity to manage and maintain the toolbox. For the data providers, capacity should be built by means of trainings and workshops, whereas the key for the toolbox users is to try to meet their capacities as best as possible by presenting the toolbox outcomes in a user-friendly way. For the data providers and toolbox users, additionally a shift in organizational culture and mentality is required to structurally start moving towards more data-driven operations. To facilitate this shift, again the main *incentivization* approach of involving parties from the start is suggested.
VISION, COMMUNICATION AND STIMULATION	No suggestions have been provided on how to establish or improve national open data platforms, though, when of a high quality, they are considered very valuable in the process of scaling up the toolbox. This is therefore a suggestion for future research. This element is also covered by the involvement of parties to create a common goal and avoid a fragmented approach.
TRANSPARENCY AND FEEDBACK	Transparency of data providers on their data collection processes and feedback to data providers on the quality of their data is not considered a priority in this phase of the emerging data ecosystem around the toolbox. To get data providers to share their data in the first place is already challenging enough. There should be a stronger focus on flexibility and looking for alternatives when failing to retrieve the required data, rather than putting effort in trying to improve data providers' data collection processes.

TABLE 9 – Proposed generic governance structure

These are the approaches that should be adopted when rolling out the Community Risk Assessment and Prioritization toolbox in a new country, thereby facilitating a scale up, which contributes to an enhanced data-driven, evidence-based decision-making in humanitarian aid

6.2. SCIENTIFIC CONTRIBUTION

No scientific studies and very few other studies have so far addressed the humanitarian data ecosystem and the possibilities for its successful establishment and governance. By taking a data ecosystem approach to analyze the governance around scaling up the Community Risk Assessment and Prioritization toolbox, thereby focusing on the 'humanitarian data ecosystem', this research fills a gap in existing literature on the development of and governance in the humanitarian data ecosystem.

Furthermore, by means of a systematic literature review, all existing literature on elements to consider when establishing successful data ecosystems in general, was merged into one framework of data ecosystem success criteria. In this framework, the success criteria are divided into three categories: 1) Data supply, relating to the provision of data as open data, 2) Governance, being the framework of policies, processes and instruments to realize common goals in the interaction between entities, and 3) User characteristics, encompassing both the user needs and the user capabilities. This developed theoretical framework was validated in two ways. In the first place, it was presented to and discussed with a group of Open Data researchers at the Delft University of Technology (expert validation). In the second place, it was case study-tested by practically applying it to the research problem of scaling up the Community Risk Assessment and Prioritization toolbox (empirical validation).

The generation and validation of this theoretical framework through a combination of a systematic literature review, expert validation and case study-testing is therefore another scientific contribution of this study.

6.3. DISCUSSION

This section will reflect on the research design and the choices made in this research.

6.3.1. DISCUSSING THE DATA ECOSYSTEM APPROACH AND LIMITATIONS

In this study, using a data ecosystem approach has been an important design choice. We found that in general, this approach proved to be a useful foundation for this study. The framework categories and criteria provided solid guidance, and brought structure to the study. Moreover, because the framework was validated both by experts and by applying it in practice, we expect that the framework allowed us to cover the whole spectrum of issues to be considered when developing a data ecosystem and do not think that important features have been overlooked.

Nevertheless, there are aspects related to the use of the data ecosystem approach in this study that we would like to reflect upon.

1. **Open data ecosystems vs. the humanitarian data ecosystem**

In the first place, we saw in *Table 1* in *Section 3.3 – Systematic literature review methodology,* that the majority (14/17) of the articles analyzed in the systematic literature review, including the article of Welle Donker & Van Loenen (2017) used to structure the systematic literature review, was centered around *open* data ecosystems. We proposed to remain a critical attitude

towards aspects resulting from the systematic literature review that would specifically characterize open data, because for the humanitarian sector, the data ecosystem in general rather than the open data ecosystem had to be assessed: much of the humanitarian data is not open, so solely focusing on open data would not be the way to go in this study. Hence, we expected that it would not be possible to translate the theoretical framework to the humanitarian data ecosystem one-on-one.
However, in retrospect, we can conclude that these concerns were unnecessary. The application of the framework that was based on open data ecosystem success criteria appeared not to be a problem, as we found that to function well, also the humanitarian data ecosystem should eventually become an open data ecosystem. The success criteria from literature are hence definitely applicable in the humanitarian data ecosystem as well, the only thing is that some of the elements only become relevant in a later ecosystem stage. We therefore propose to view the humanitarian data ecosystem as an immature version of the open data ecosystems assessed in the literature review, rather than stating that the open data ecosystems perspective was not useful to be applied in the humanitarian context. Yet, at the same time this illustrates a limitation of the use of the theoretical framework: the framework is static, whereas it is used to describe a dynamic situation.
As described before, it would be interesting to conduct further research of what this transition in data ecosystem maturity entails for data ecosystems other than the humanitarian one (see *Section 6.4.2 – Recommendations for future research).*

2. Alternative approaches

Looking back on this study, another question we can pose ourselves is: why the data ecosystem approach in the first place? When determining on a suitable theoretical framework for investigation, there was one other perspective that we considered, namely the perspective of 'data collaboratives'. Susha, Janssen & Verhulst (2017) 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" (p.2691). Especially the fact that data collaboratives are focused on 'societal challenges' appealed to us in relation to this study. Nevertheless, we found that the definition of data collaboratives is somewhat narrower than a data ecosystem. Data collaboratives address a specific, shorter term problem to be solved, and it is a collaboration between a well-defined set of actors. Data collaboratives can also be established within a data ecosystem. We therefore considered the data ecosystem approach to be more applicable to this research problem, in which a sustainable ecosystem with a network of actors needed to be developed.

6.3.2. Discussing the selection and role of the interviewees

Another important design decision has been the selection of the interviewees. The aim of this research was to come to generic governance approach for scaling up the Community Risk Assessment and Prioritization toolbox – generic being not-country specific. In *Section 5.3.2.* we already discussed that it was difficult to identify interviewees that could contribute to this aim. The interviewees were required to have an overview of the inter-organizational data landscape in their country, and not many people were found to fulfill this criterion (which is also one of the issues causing this study to be relevant).
But despite the challenge to find a set of suitable interviewees, we succeeded and we reflect positively on the set of interviewees selected ultimately. They represented a great variety of

cases, were all very knowledgeable on the data ecosystem success criteria they were questioned about, and provided similar approaches to deal with the rolling out of models like INFORM Subnational and the Community Risk Assessment and Prioritization toolbox in a country. This allowed for the creation of one generic approach, so we can say that the selected set of interviewees did serve the purpose.
Nevertheless, we think that on some specific aspects, some bias may have been introduced by the interviewees, strengthened by the fact that our set of interviewees was limited in the sense that for most regions we only interviewed one humanitarian data expert.
To illustrate this: the interviewees did not prioritize the presence of the 'feedback and transparency' criterion due to their current focus of getting data providers to share their data in the first place. However, a 510 Global team member very recently participated in a workshop with local stakeholders in Malawi, and one of the things that popped up during this workshop was that the attendants felt the need for better feedback mechanisms. This contradicts the interview outcomes. It illustrates that people have different perspectives and that a limitation of this study is that this is not entirely accounted for. We interpreted the consensus of the interviewees in their answers as an increased probability that their answers were indeed accurate and that they would hold in general, but with a set of only eight interviewees, which is not large in terms of sample sizes, this of course does not necessarily need to be the case.
As to the actual carrying out of the interviews, we found that the interview protocol was very useful and that it was important to use it properly: some of the interviewees liked to talk and easily diverted to a different topic if they were not given direction. Sticking to the interview protocol in these cases helped them to get back on track.

6.3.3. Discussing the value of data-driven aid prioritization

The last aspect to reflect on is a very high-level point of discussion on the general correctness of the use of data-driven prioritization models like the Community Risk Assessment and Prioritization toolbox, around which this this study is centered. In this study, we claim that the data-driven prioritization of aid areas is more effective and efficient than solely relying on field assessments, which is the idea behind the development of the toolbox. We would like to remark that we think that this is indeed true, that we think that a shift towards more evidence-based decision-making should indeed be facilitated, and that data-driven models are a very valuable addition to the current field assessment methodologies, but that there can also be a danger in relying too much on these kinds of models. They give the impression that everything can be measured and based on hard numbers, but this is not the case in the humanitarian sector, when working with aggregated scores on 'risks' and 'vulnerabilities'. Because how for example to interpret the 'mean vulnerability'? Human assessment will be necessary to provide meaning to risk and vulnerability scales and scores like this, and to give insight in the context. Additionally, the aggregation of many indicators can lead to a distorted picture of the situation, that a human eye would assess differently.
Of course, there are also downsides to assessment teams going into the fields, as they are also prone to certain biases which can lead to unreliable assessments. However, in these cases everyone seems to account for these biases, whereas the inclusion of data-driven models leads to the idea that the generated numbers are completely accurate, and that is the pitfall we wanted to highlight here. To conclude, we think there is a good motivation for using data-driven prioritization models like the Community Risk Assessment and Prioritization

toolbox, but to ensure they are interpreted correctly and placed in the right context, these models should structurally be combined with human insights.

6.4. Recommendations

6.4.1. Recommendations for 510 Global

The generic governance structure we developed in this study shows how the process of rolling out the Community Risk Assessment and Prioritization toolbox should eventually be approached. It in fact envisions what the optimal situation should look like. This process ideally starts with a UN agency (the initiator/coordinator) initiating the toolbox development for a country and involving the other roles. In *Section 5.4.4 – Additional findings,* we described that for INFORM Subnational the start of this process has so far never been an issue, as UN agencies tend to approach INFORM with the question whether they can take the lead in developing INFORM Subnational for their region. Hence, for INFORM, the identification of a suitable initiator/coordinator is a process that occurs naturally.

However, this is not the case for 510 Global and their toolbox. For the countries for which they are currently developing the toolbox, they are the party taking up the role of the initiator/coordinator; they are the sole party taking the responsibility for the toolbox development. This is for example the case for Malawi, where 510 Global is working on the capacity-building of potential data providers and toolbox users and making clear what is the added value of the model to them by involving them in workshops. Due to this process of rolling out the toolbox being very time consuming and resource intensive, this is not a scalable solution that can be widely applied. The question therefore is, how can 510 Global make the transition from this current situation where they carry out the role of initiator/coordinator and there is no local lead, to the better scalable and more sustainable division of roles that resulted from the governance structure developed in this study (see *Figure 24*)?

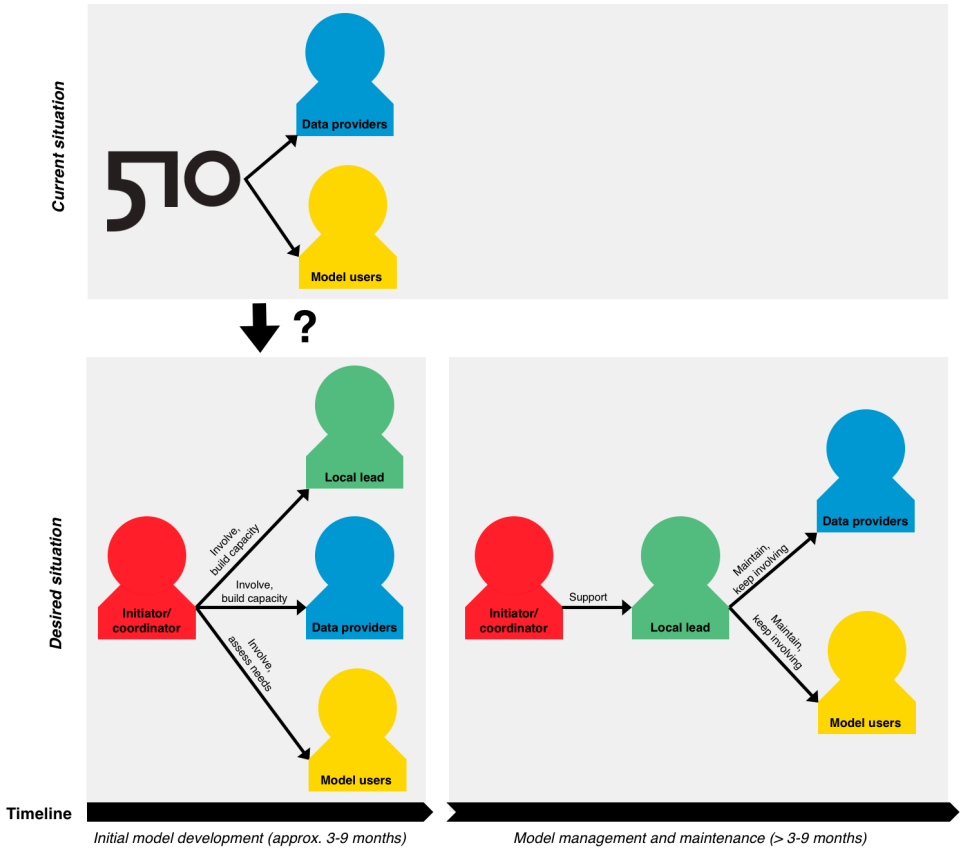


Figure 24 – Visualization of desired transition from current situation to desired situation

To realize this transition, we advise 510 Global the following as regards their position relative to the other stakeholders and their (future) role in the process of rolling out the toolbox in a country:

1. **Explore possibilities to better collaborate with INFORM**

In this study, we showed that the process of developing INFORM Subnational shows very much resemblance to the process of rolling out the Community Risk Assessment and Prioritization toolbox in a country. Both models are based on the INFORM dimensions and aim for the highest achievable level of granularity. It would therefore be efficient to join forces and discuss countries to be targeted. We expect that this is interesting for both parties, as both INFORM and 510 Global benefit from each other's model development. Developing both INFORM Subnational and the Community Risk Assessment and Prioritization toolbox for a country would be an unnecessary duplication of efforts.
Therefore, the possibilities for a closer collaboration with INFORM should be assessed by entering the dialogue with them, fueled by the research findings that both INFORM Subnational and the Community Risk Assessment and Prioritization toolbox pursue the same objectives, and that both development processes resemble each other.

2. **Use INFORM as an entry to directly approach UN agencies**

The fact that it occurs regularly that UN agencies send requests to INFORM for the development of INFORM Subnational in their region, indicates that the brand 'INFORM' is already sufficiently well established in the humanitarian sector – at least among UN agencies – to be a driver of their operations. As UN agencies were identified as being the most suitable initiator/coordinator and they know where to find INFORM, we recommend 510 Global to discuss with INFORM the possibilities of INFORM serving as an entry for 510 Global to directly approach UN agencies to ask whether they want to become responsible for the toolbox development.
Concretely, this means that for a country for which 510 Global wants to develop the toolbox, INFORM could invite the UN agency they have best connections with to a round-the-table discussion together with them and 510 Global, in which 510 Global explains to this UN agency why the toolbox development would be valuable for them ('what is in it for them'), and INFORM serves as a mediator. Also, the finances of the toolbox development should be discussed: the initiator/coordinator should become financially responsible, but to convince them to participate it can help to for example already outline possibilities for funds and grants. INFORM can play a role here, by explaining how other UN agencies did this with the development of INFORM Subnational.
If this would succeed, the timeline would look as visualized in *Figure 25*, in which also the next recommendation is included.

3. **Anticipate long-term involvement and develop a business model accordingly**

As can be seen in *Figure 25*, we expect that in practice also during the initial model development 510 Global keeps playing a role, as they are the facilitator of the toolbox. They should be available to provide support to both the initiator/coordinator, this support being of a similar nature as the support that INFORM offers to the initiators/coordinators that approach them, as specified in *Section 5.4.4 – Additional findings* (support at a technical level, for funding

ideas, in the form of providing consultants that can help guiding the process and in the form of training the initiator/coordinator to work with the model). 510 Global should ideally train the initiator/coordinator as such that they will be able to provide the required assistance to the local lead, but if the local lead needs additional help, this is also where 510 Global would come in. This supporting role of 510 Global should diminish over time, as the initiator/coordinator and local lead become more experienced.
Additional research should be done on a suitable business model for 510 Global as the service provider and facilitator in this process. This will be further discussed in *Section 6.4.2 – Recommendations for future research.*

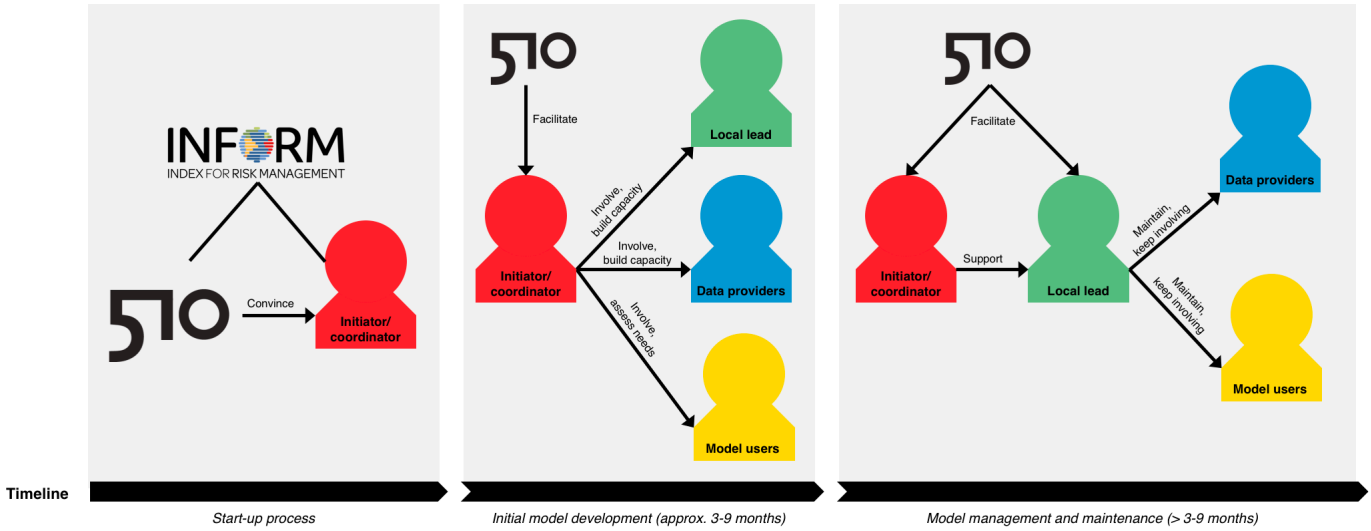


FIGURE 25 – Timeline of the involvement of and relationship between different roles, adjusted for the recommendation of directly approaching a suitable initiator/coordinator via INFORM

4. Adopt a stepwise approach and link this to donor

If the second recommendation does not turn out and 510 Global cannot directly find a suitable initiator/coordinator, 510 Global should adopt the 'stepwise approach' as proposed by one of the interviewees (and described in *Section 5.4.3 – Main empirical findings)*: to use the easiest entry of 510 Global, namely the National Red Cross Society, and from there start developing the toolbox, expand the connections in the country and in a later stage try to incentivize a more suitable party to take over the leading role. This would be a gradual, stepwise transition of the current situation to the desired situation as visualized in *Figure 24.* This stepwise approach is very resource-intensive for 510 Global, as National Red Cross Societies will generally need a lot of support due to their limited capacities.
An additional issue here is that National Red Cross Societies do not only have limited capacities, but also limited resources. This means that in this case there would be a problem with the financing of the toolbox development. A suggestion on how to deal with this, is to link this stepwise approach to a financial donor. If financial donors make the development of the Community Risk Assessment and Prioritization toolbox a condition for National Red Cross Societies to apply for their funding, National Red Cross Societies will have a financial incentive to become engaged in the toolbox development. Part of the funding that becomes available through this approach should be used for the required capacity-building of the National Red Cross Society in question.

In a recent meeting with 'ECHO', the aid department of the European Union who have budget for humanitarian funding, it appeared that they are interested in becoming more involved in the projects they support financially, and would be open to ideas like this. Moreover, moving towards a more evidence-based humanitarian decision-making and thereby adopting a standardized approach for each country is also in line with their objectives. This possibility should be further investigated, by discussing it with financial donors National Red Cross Societies regularly appeal for.

This stepwise approach will be more suitable to adopt in countries of which the National Red Cross Society is a partner society of the Netherlands Red Cross. In those countries, the Netherlands Red Cross already has a formal role and is involved in projects with that national society. Consequently, the toolbox development can be linked to other ongoing projects, and it will be easier to explain and show the value of it.

6.4.2. Recommendations for future research

On the Community Risk Assessment and Prioritization toolbox

- This study focused on the development, roll out and scale up of the Community Risk Assessment and Prioritization toolbox: on the data ecosystem required to get the model running. However, to have the model structurally integrated into humanitarian practices and decision-making, also other aspects need to be considered, that have not been assessed in this study. For example, more research could be done on the adaptation of the toolbox. Interviewees mention that a shift in organizational culture and in mentality is required to have the stakeholders incorporate the toolbox in their operations, but, apart from involving them in the process, how can this be established?

- Also, the finance around the toolbox requires more attention. In this study, only the party responsible for the financing has been identified (the initiator/coordinator), but further details on how the realization of this sustainable financing can be best approach should be studied in a future research. Additionally, currently a long-term business model for 510 Global is lacking: if they continue to actively operate as a facilitator and service provider of the toolbox, what will their business case be? This is something that should be further investigated.

- As mentioned in the previous section, also the possibilities to collaborate with financial donors and have them play a formal role in the development of the toolbox should be further assessed.

- The interviewees did not provide useful insights in the establishment and improvement of national open data platforms, whereas they did consider them to have the potential to be valuable. Currently however, their quality is often questionable, and data is scattered across multiple initiatives. Future research could address these issues.

On the theoretical framework of data ecosystem success criteria

- As described in *Section 5.5.2.,* the theoretical framework should be further validated by means of applying it to other case studies, to see if it is also a useful for the analysis

of other data ecosystems, and if the application on other data ecosystems reveals additional suggestions for improvement.

- When doing so, it would be interesting to assess whether the interrelation between the different framework parts and the duplication of roles that was found to be present in the humanitarian data ecosystem also holds for other data ecosystems, or if this complexity specifically characterizes the humanitarian data ecosystem.
- Additionally, attention should be paid to the evolvement and dynamics of the maturity of other data ecosystems: what do the transitions in the relevance of different elements look like, and how are these transitions realized? This is something the emerging humanitarian data ecosystem can learn from.

6.5. Reflection on link with CoSEM program

The CoSEM Master's program addresses innovations in complex socio-technical environments. To come to successful innovations, it is important not just to design the technological system behind it, but also to consider the social context of the innovation that enables its implementation. Moreover, design in complex socio-technical systems often has an international character, which is why it is important to account for cultural differences, for example.

The link between the CoSEM Master's program and this research relies in the fact that the development and roll out of the Community Risk Assessment and Prioritization toolbox is a typical example of an innovation in a complex socio-technical system, that crosses borders. In this study, it became clear that a solid technological solution alone is not sufficient to succeed; dealing with the complex actor environment in which the toolbox needs to be embedded and is at least as important.

REFERENCES

510 Global. (2016A). *A Priority Index for Humanitarian Aid After a Typhoon.* November 9, 2016. Retrieved from https://510.global/philippines-typhoon-haima-priority-index/#fn-5-1.

510 Global. (2016B). *Personalised Humanitarian Aid Through Big Data: an initiative of the Dutch Red Cross.* Retrieved from http://510.global/.

510 Global. (2017A). *Community Risk Assessment and Prioritization toolbox.* Retrieved from https://profiles.510.global/#!/ .

510 Global. (2017B). *Mission & Vision.* Retrieved from https://www.510.global/video/.

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard Business Review, 84,* pp. 98-107.

Attard, J., Orlandi, F., Auer, S. (2016). *Data Value Networks: Enabling a New Data Ecosystem.* 2016 IEEE/WIC/ACM International Conference on Web Intelligence.

Backx, M. (2003). *Gebouwen Redden Levens. Toegankelijkheidseisen van Gebouwgegevens in Het Kader van de Openbare Orde en Veiligheid. [Buildings Save Lives. Accessibility Requirements for Buildings in the Context of the Public Order and Safety].* M.Sc. thesis, Delft University of Technology.

Barthélemy, F. (2016). *The Belgian open data ecosystem and innovation through open data.* M.Sc. thesis, Louvain School of Management, Université catholique de Louvain.

Benini, A. (2015). *The use of Data Envelopment Analysis to calculate priority scores in needs assessments.* A note for ACAPS.

Berens, J., Raymond, N., Shimshon, G., Verhulst, S., and Bernholz, L. (2016). The Humanitarian Data Ecosystem: the Case for Collective Responsibility. Retrieved from https://pacscenter.stanford.edu/wp-content/uploads/2017/08/humanitarian_data_ecosystem.pdf.

Britten, N. (1999). Qualitative interviews in healthcare. In C. Pope and N. Mays (Eds), *Qualitative research in health care* (pp. 11-19). London: BMJ Books.

Bulte, E. (2017). *Estimating Post-Earthquake Aid Priority Areas: A data-driven approach towards aid needs modelling.* M.Sc. thesis, Geographical Information Management and Applications, UU WUR UT TUD.

Centre for Research on the Epidemiology of Disasters (CRED) & The United Nations Office for Disaster Risk Reduction (UNISDR). (2015). *The Human Cost of Weather Related Disasters.* Retrieved from http://reliefweb.int/sites/reliefweb.int/files/resources/COP21_WeatherDisastersReport_2015_FINAL.pdf.

Chan, J., Bateman, L and Olafsson, G. (2016). A people & purpose approach to humanitarian data information security and privacy*. Procedia Engineering 159,* pp. 3-5.

Chaumba, J. & Van Geene, J. (2003). Participatory capacity building in NGOs. *Participatory Learning and Action (PLA) Notes, 48,* pp. 69-75.

Craglia, M., Annoni, A., Smith, R.S. and Smits, P. (2002). *Spatial Data Infrastructures: Country Reports.* GINIE: Geographic Information Network in Europe. Retrieved from http://www.ec-gis.org/ginie/doc/SDI_final_en.pdf .

Crowe, A. (2013). *A paucity of privacy: Humanitarian, development organisations need beneficiary data protection policies.* Privacy International. Retrieved from https://www.privacyinternational.org/node/240 .

Data-Pop Alliance. (2016). *What is Data Literacy?* Retrieved from http://datapopalliance.org/item/what-is-data-literacy.

Davies, T. (2012). *Open Data: Infrastructures and ecosystems.* ACM Web Science Conference 2012, Evanston, Illinois, USA.

Dawes, S.S., Vidiasova, L. and Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly, 33,* pp. 15-27.

De Bruijn, H. and Herder, P.M. (2009). System and actor perspectives on socio-technical systems. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 39*(5), pp. 981–992.

Gill, P., Stewart, K., Treasure, E. and Chadwick, B. (2008). Methods of data collection in qualitative research: interviews and focus groups. *British Dental Journal, 204,* pp. 291-295.

Global Pulse. (2015). *Big Data for Development and Humanitarian Action: Towards Responsible Governance.* Retrieved from http://unglobalpulse.org/sites/default/files/Big_Data_for_Development_and_Humanitarian_Action_Report_Final_0.pdf.

Grant, M.J. and Booth, A. (2009). A typology of reviews: an analysis of 14 reviews types and associated methodologies. *Health Information and Libraries Journal, 26*, pp. 91-108.

Hadzic, M. and Sidhu, A. (2008). Digital health ecosystems. *2nd IEEE International Conference on Digital Ecosystems and Technologies.*

Harke, J. and De Leeuw, S. (2015). Enhancing Sustainability in Managing Inventory Prepositioning Networks for Disaster Relief Through a Simulation Game. In M. Klumpp et al. (Eds), *Humanitarian Logistics and Sustainability* (pp. 215-232). Springer International Publishing.

Harrison, T.M., Pardo, T.A. and Cook, M. (2012). Creating Open Government Ecosystems: A Research and Development Agenda. *Future Internet 2012, 4*, pp. 900-928.

Heimstädt, M., Saunderson, F. and Heath, T. (2014). *Conceptualizing Open Data ecosystems: A timeline analysis of Open Data development in the UK.* Discussion paper, Freie Universität Berlin, School of Business & Economics: Management, No. 2014/12.

Herder, P.M. and Stikkelman, R.M. (2004). Methanol-based industrial cluster design: a study of design options and the design process. *Industrial & Engineering Chemistry Research, 42*, pp. 3879–3885.

Hilhorst, D. (2013). Disaster, conflict and society in crises: everyday politics of crisis response: Routledge.

Humanitarian Country Team. (2015). *Inter-Agency Common Feedback Project.* Retrieved from http://reliefweb.int/sites/reliefweb.int/files/resources/Nepal%20Earthquake%20Inter-agency%20Common%20Feedback%20Project%20Proposal%20v3.pdf .

ICCO, NLRC, HOT and Akvo. (2016). *Technical Proposal: Developing data schema for inclusion of site-specific infrastructural and sub-national socio-economic data into the Exposure Database for Tanzania and Bangladesh through a participatory and collaborative approach.* Internal document.

ICRC. (2013). *The Movement.* Retrieved from https://www.icrc.org/eng/who-we-are/movement/overview-the-movement.htm .

ICRC. (2017I). *The International Red Cross and Red Crescent Movement.* Retrieved from https://www.icrc.org/en/who-we-are/movement.

ICRC. (2017II). *What we do.* Retrieved from https://www.icrc.org/en/what-we-do.

IFRC. (2017I). *The International Federation of Red Cross and Red Crescent Societies (IFRC).* Retrieved from http://www.ifrc.org/en/who-we-are/the-movement/ifrc/.

IFRC. (2017II). *Types of disasters: Definition of hazard.* Retrieved from

http://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/definition-of-hazard/.

Immonen, A., Palviainen, M. and Ovaska, E. (2014). *Requirements of an Open Data Based Business Ecosystem. IEEE, 2,* pp. 88-103.

INFORM. (2016). *Methodology.* Retrieved from http://www.inform-index.org/InDepth/Methodology.

INFORM. (2017). *INFORM Colombia.* Retrieved from http://www.inform-index.org/Subnational/Colombia.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management, 29*(4), pp. 258–268.

Jetzek, T. (2017). *Innovation in the Open Data Ecosystem: Exploring the Role of Real Options Thinking and Multi-sided Platforms for Sustainable Value Generation through Open Data.* In Analytics, Innovation, and Excellence-Driven Enterprise Sustainability, Part of the series Palgrave Studies in Democracy, Innovation, and Entrepreneurship for Growth, pp. 137-168.

Khan, K.S., Kunz, R., Kleijnen, J. And Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine, 96*(3), pp. 118-121.

Kok, B. and Van Loenen, B. (2005). How to Assess the Success of National Spatial Data Infrastructures? *Computers, Environment and Urban Systems, 29*(6), pp. 699-717.

Lee, D. (2014). *Building an Open Data Ecosystem – An Irish Experience.* ICEGOV2014, Guimaraes, Portugal.

Leidig, M. and Teeuw, R.M. (2015). Quantifying and Mapping Global Data Poverty. *PLoS ONE, 10*(11).

Leidig, M, Teeuw, R.M. and Gibson, A.D. (2016). Data poverty: A global evaluation for 2009 to 2013 - implications for sustainable development and disaster risk reduction. *International Journal of Applied Earth Observation and Geoinformation, 50,* pp.1-9.

Macharis, C. and Crompvoets, J. (2014). A stakeholder-based assessment framework applied to evaluate development scenarios for the spatial data infrastructure for Flanders. *Computers, Environment and Urban Systems, 46*, pp. 45-56.

Mans, U., Berens, J. and Shimshon, G. (2015). *The New Humanitarian Data Ecosystem: Challenges and Opportunities to Increase Trust and Impact.* World Humanitarian Summit White Paper. Retrieved from http://www.peaceinformaticslab.org/uploads/1/5/6/9/15692298/20150813_whs_document_leiden_university_final.pdf.

Monaghan, A and Lycett, M. (2013). *Big Data and Humanitarian Supply Networks: Can Big Data Give Voice to the Voiceless?* Paper presented at the 2013 Global Humanitarian Technology Conference (GHTC), IEEE.

Moore, J.F. (1996). *The Death of Competition – Leadership and Strategy in the Age of Business Ecosystems.* New York: Harper Business.

Nardi, B. and O'Day, V.L. (1999). *Information Ecologies: Using Technology with Heart.* MIT Press: Cambridge, MA, USA.

Netherlands Red Cross. (2017). *Project Proposal: Open source Community Risk Assessment and Prioritization toolbox.* Internal document.

OpenData Barometer. (2017). *The Open Data Barometer.* Retrieved from http://opendatabarometer.org/?_year=2016&indicator=ODB.

Open Data Charter. (2017). *Open data measurement: an update.* Retrieved from https://medium.com/@opendatacharter/open-data-measurement-an-update-b4079f7870c6.

Open Data Watch. (2016). *Open Data Inventory 2016.* Retrieved from http://odin.opendatawatch.com.

Open Knowledge Network. (2017). *Tracking the State of Open Government Data.* Retrieved from https://index.okfn.org/.

Organisation for Economic Co-operation and Development (OECD). (2017). *Open Government Data.* Retrieved from http://www.oecd.org/gov/digital-government/open-government-data.htm.

Ortuño, M.T. et al. (2013). Decision Aid Models and Systems for Humanitarian Logistics. A Survey. *Decision Aid Models for Disaster Management and Emergencies,* Volume 7 of the series *Atlantis Computational Intelligence Systems,* pp. 17-44.

Payne, K., Florance, P. and Shain, S. (2012). The Role of Data Repositories in Humanitarian Information Management and Crisis Mapping. *Journal of Map & Geography Libraries, 8*(2), pp. 118-133.

Pollock, R. (2011). *Building the (Open) Data Ecosystem.* Retrieved from https://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/ .

Parrish, C. (2016). *Better Information for a Better Response – The basics of humanitarian transparency.* Development Initiatives. Retrieved from http://www.globalhumanitarianassistance.org/wp-content/uploads/2016/05/Better-information-for-a-better-response.pdf .

Parsons, M.A., Godøy, Ø., LeDrew, E., De Bruin, T.F., Danis, B., Tomlinson, S. & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science, 37*(6), pp. 555-569.

Ponte, D. (2015). Enabling an Open Data Ecosystem. *ECIS 2015 Research-in-Progress Paper* 55; Presented at the Thenty-Third European Conference on Information Systems (ECIS), Münster, Germany.

Raymond, N. and Al Achkar, Z. (2016). *Data preparedness: connecting data, decision-making and humanitarian response.* Harvard Humanitarian Initiative: signal program on human security and technology - standards and ethics series 01. Retrieved from http://hhi.harvard.edu/sites/default/files/publications/data_preparedness_update.pdf.

Raymond, N., Al Achkar, Z., Verhulst, S. And Berens, J. (2016). *Building data responsibility into humanitarian action.* OCHA Policy and Studies Series 018. Retrieved from http://www.unocha.org/sites/unocha/files/Building%20data%20responsibility%20into%20humanitarian%20action.pdf.

Raymond, N. A. and Card, B.L. (2015). *Applying Humanitarian Principles to Current Uses of Information Communication Technologies: Gaps in Doctrine and Challenges to Practice.* Harvard Humanitarian Initiative, signal program on human security and technology. Retrieved from http://hhi.harvard.edu/sites/default/files/publications/signal_program_humanitarian_principles_white_paper.pdf.

Stojic, M. (2013). *Disaster Management: Staying Ahead of the Storm.* Earth Imaging Journal, December 18, 2013. Retrieved from http://eijournal.com/print/articles/disaster-management.

Susha, I., Janssen, M. and Verhulst, S. (2017). *Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development.* Proceedings of the 50th Hawaii International Conference on System Sciences, Waikoloa, HI.

TU Delft Global Initiative. (2017). *Story nr. 11- Bartel Van de Walle: The impact of data on disaster relief.* Retrieved from http://globalstories.tudelft.nl/story/bartel-van-der-walle/.

Twigg, J. (2004). *Disaster risk reduction: Mitigation and preparedness in development and emergency programming.* London: Humanitarian Practice Network.

UNICEF. (2017). *About UNICEF.* Retrieved from https://www.unicef.org/about-us .

UNISDR. (2017). *Terminology.* Retrieved from https://www.unisdr.org/we/inform/terminology.

UN/ISDR & UN/OCHA. (2008). *Disaster Preparedness for Effective Response: Guidance and Indicator Package for Implementing Priority Five of the Hyogo Framework.* United Nations secretariat of the International Strategy for Disaster Reduction (UN/ISDR) and the United Nations Office for Coordination of Humanitarian Affairs (UN/OCHA), Geneva, Switzerland.

United Nations Development Program (UNDP). (2017). *A world of development experience.* Retrieved from http://www.undp.org/content/undp/en/home/operations/about_us.html.

United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA). (2013). *Tools and Services for Disaster Response.* Retrieved from http://www.unocha.org/publications/asiadisasterresponse/ToolsAndServicesForDisasterResponse.html#Information Management Services.

United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA). (2017). *Who We Are.* Retrieved from http://www.unocha.org/about-us/who-we-are .

Van den Homberg, M. (2016). *Data for Disaster Management: Mind the Gap.* Retrieved from http://www.mei.edu/content/map/data-disaster-management-mind-gap.

Van den Homberg, M., Monné, R. and Spruit, M. (2017 – to be published). Bridging the information gap by optimizing data selection using cost and quality to cover information needs of disaster responders. CAGEO.

Van den Homberg, M. and Van der Veen, M. (2017). *Data Preparedness for National Societies.* The Netherlands Red Cross, Internal Document.

Van den Homberg, M., Van der Veen, M. and Visser, J. (2016). *Open Source Community Risk Assessment and Prioritization Toolbox as part of Data Preparedness.* 510.global, Internal Document.

Van den Homberg, M., Visser, J. and Van der Veen, M. (2017). *Unpacking Data Preparedness from a humanitarian decision-making perspective: toward an assessment framework at subnational level.* Proceedings of the 14th ISCRAM conference, Albi, France.

Van Lint, S. (2016). *Sense-making of the Netherlands Red Cross Priority Index Model. Case: Typhoon Haiyan, Philippines.* M.Sc. thesis, Disaster and Conflict Studies, Wageningen University.

Van Schalkwyk, F., Willmers, M. and McNaughton, M. (2016). Viscous Open Data: The Roles of Intermediaries in an Open Data Ecosystem. *Information Technology for Development, 22*(1), pp. 68-83.

Van de Walle, B. and Comes, T. (2014). Risk Accelerators in Disasters: Insights from the Typhoon Haiyan Response on Humanitarian Information Management and Decision Support. In M. Jarke et al. (Eds), *Advanced Information Systems Engineering, Vol. 8484* (pp. 12-23). Springer International Publishing.

Van de Walle, B. and Comes, T. (2015). On the Nature of Information Management in Complex and Natural Disasters. *Procedia Engineering 107,* pp. 403-411.

Van de Walle, B. and Comes, T. (2016). Information Systems for Humanitarian Logistics. In G. Kovas, K. Spens and I. Haavisto (Eds.), *Supply Chain Management for Humanitarians: Tools for practice* (pp. 257-284). Kogan Page.

Vitoriano, B., Montero, J., and Ruan, D. (2013) Decision aid models for disaster management

and emergencies. In J. Lu & J. Montero (Series Ed.): Vol. 7 (pp. 333): Atlantis Press.

Wang, Y., Gu, J. and Zhou, Z. (2014). Complex Adaptive Socio-Technical Systems Theory View of Ubiquitous Computing Systems Research. In R. Hervás et al. (Eds), *UCAmI 2014, LNCS 8867* (pp. 216-223). Springer International Publishing.

Welle Donker, F. and Van Loenen, B. (2017). How to assess the success of the open data ecosystem? *International Journal Of Digital Earth, 10*(3), pp. 284-306.

Wiener, M., Sommer, F.T., Ives, Z.G., Poldrack, R.A. and Litt, B. (2016). Enabling an Open Data Ecosystem for the Neurosciences. *Neuron, 92*, pp. 617-621.

Whipkey, K. and Verity, A. (2015). *Guidance for incorporating Big Data into Humanitarian Operations.* Document created as part of the Digital Humanitarian Network's community of interest on Decision Makers Needs.

Zuiderwijk, A., Janssen, M. and Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity, 19,* pp. 17-44.

APPENDICES

APPENDIX A – LITERATURE SEARCH ON THE 'HUMANITARIAN DATA ECOSYSTEM'

Searching scientific databases for the term 'humanitarian data ecosystem' resulted in zero hits. On Google Search, searching for "humanitarian data ecosystem" resulted in 50 hits. Six different documents in which the term popped up were found. 30 of the search hits included the same link in which was referred to *document 1* in the list below, and the remaining hits redundantly referred to the other five documents. The six documents in which the term was used are briefly described below, as well as their perceived usefulness as to the achievement of this research objective.

As can be seen, four authors (Al Achkar, Berens, Raymond and Verhulst) were involved in more than one of the documents listed below. Additionally, the author of *Document 5* also contributed to *Document 3*.

Doc. 1 Berens, J., Raymond, N., Shimshon, G., Verhulst, S., and Bernholz, L. (2016). The Humanitarian Data Ecosystem: the Case for Collective Responsibility. Retrieved from https://pacscenter.stanford.edu/wp-content/uploads/2017/08/humanitarian_data_ecosystem.pdf.

- Strong focus on data responsibility and ethics.
- Description of actor complexity in the humanitarian data ecosystem, in line with *Section 2.2.2 – Actor complexity.*
- Stress the need for the development of governance structures for the humanitarian data ecosystem, but only discuss this in relation to the responsible use of data; it is argued that governance processes for the responsible use of data should be created.
- Conclusion: can be used as background material, but not useful to generate insight in success criteria for the development of a humanitarian data ecosystem, needed for this research.

Doc. 2 Mans, U., Berens, J. and Shimshon, G. (2015). *The New Humanitarian Data Ecosystem: Challenges and Opportunities to Increase Trust and Impact.* World Humanitarian Summit White Paper. Retrieved from http://www.peaceinformaticslab.org/uploads/1/5/6/9/15692298/20150813_whs_document_leiden_university_final.pdf

- Provides an overview of how new data streams impact humanitarian affairs and takes up a practical approach of how data-driven innovation can take shape in the coming years.
- Describe developments in the 'humanitarian data ecosystem' and use the term, but do not define it.
- Additionally, the authors take up a rather practical approach without a clear research design.
- Conclusion: valuable background material, but not useful to generate insight in success criteria for the development of a humanitarian data ecosystem, needed for this research.

Doc. 3 Raymond, N. and Al Achkar, Z. (2016). *Data preparedness: connecting data, decision-making and humanitarian response.* Harvard Humanitarian Initiative: signal program on human security and technology - standards and ethics series 01. Retrieved from http://hhi.harvard.edu/sites/default/files/publications/data_preparedness_update.pdf.

- Document of the Harvard Humanitarian Initiative.
- *"This paper seeks to provide a blueprint for how the concept of data preparedness may be put into practice by members of the humanitarian data ecosystem"* (p.3).
- Only refers to the term, and the definition of it by Raymond et al. (2016) (see *Document 4*), but no further analysis of it.
- Conclusion: can be used as background material, but not useful to generate insight in success criteria for the development of a humanitarian data ecosystem, needed for this research.

Doc. 4 Raymond, N., Al Achkar, Z., Verhulst, S. And Berens, J. (2016). *Building data responsibility into humanitarian action.* OCHA Policy and Studies Series 018. Retrieved from http://www.unocha.org/sites/unocha/files/Building%20data%20responsibility%20into%20humanitarian%20action.pdf.

- Publication developed by OCHA Policy Development and Studies Branch (PDSB), Policy Analysis and Innovation Section in partnership with the Harvard Humanitarian Initiative, NYU GovLab and Leiden University Centre for Innovation.
- Define the 'humanitarian data ecosystem': *"humanitarian actors, their partners and affected communities are producing, capturing and accessing growing volumes of digital data about vulnerable populations. This group of data producers, users and consumers together make up a diverse and complex 'humanitarian data ecosystem'"* (p. 3).
- The authors use the term, but do not further analyze the characteristics of this data ecosystem; they describe it in relation to the responsible use of data. Their approach is: the responsible use of data is the collective responsibility of the entire humanitarian data ecosystem.
- Conclusion: can be used as background material, but not useful to generate insight in success criteria for the development of a humanitarian data ecosystem, needed for this research.

Doc. 5 Sabou, J. (2016). *The Humanitarian "Data" Ecosystem: Intersection of Innovation & Disaster Relief.* Retrieved from http://www.leidensafetyandsecurityblog.nl/articles/the-humanitarian-data-ecosystem-intersection-of-innovation-disaster-relief.

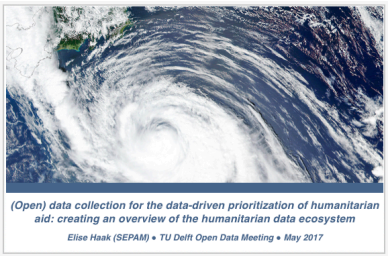
- Short blogpost on the increase of the importance of data in the humanitarian sector.
- Stresses the need for central governance of the "humanitarian – 'digital data' – ecosystem", as there is currently a lack of.
- No insights on how to achieve this.

- Conclusion: can be used as background material, but not useful to generate insight in success criteria for the development of a humanitarian data ecosystem, needed for this research.

Doc. 6 Terp, S. (2015). *Evolution of the Humanitarian Data Ecosystem.* Slides for presentation given at AAAI Spring 2015 on humanitarian linked open data. Retrieved from https://www.slideshare.net/bodacea/evolution-of-the-humanitarian-data-ecosystem.

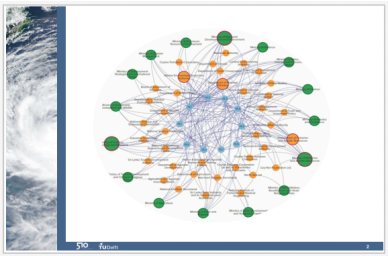
- Presentation slides on data in the humanitarian sector.
- No specific definition or characterization of the humanitarian data ecosystem; only use of the term.
- Conclusion: not useful to generate insight in success criteria for the development of a humanitarian data ecosystem, needed for this research.

APPENDIX B – PRESENTATION SLIDES EXPERT VALIDATION SESSION

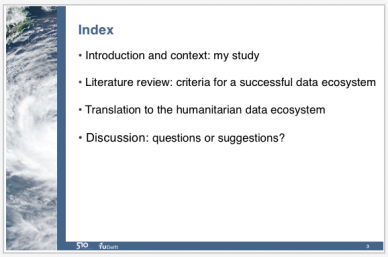


1

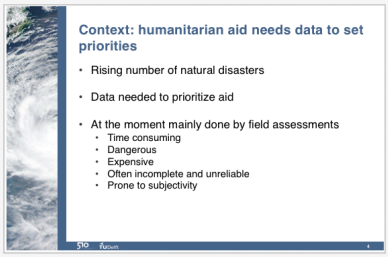
2



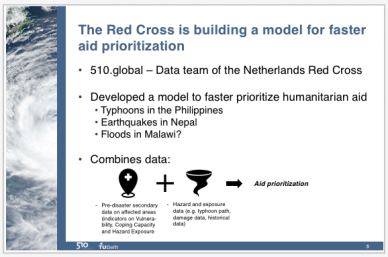
3



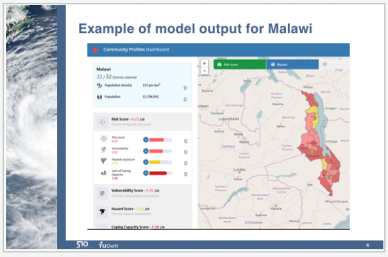
4



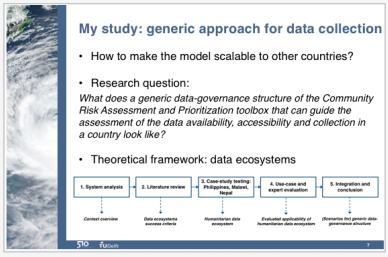
5



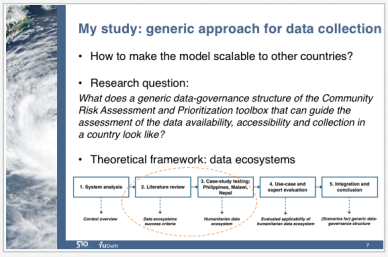
6



7



8



9

(Continue on next page)

Literature review: approach

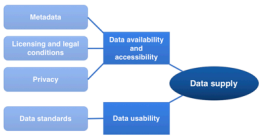
- Scopus, Web of Science, Google Scholar, reference list of relevant articles
- Data ecosystem, open data ecosystem, government ecosystem, criteria ecosystem, data infrastructure
- Focus on criteria for a successful data ecosystem / essential elements of a data ecosystem
- Set of 17 relevant articles, mostly on '*open* data ecosystems'

10

Literature review: Criteria for a successful data ecosystem

11

Key elements of data supply

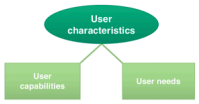


12

Key elements of data governance

13

User characteristics



14

Take into account external influences

15

Literature review evaluation

1. Suggestions for additional literature?
2. Is this overview of elements complete?
3. Other aspects missing/overlapping?
   - Privacy only mentioned in 4/17 articles
   - Importance of trust not mentioned at all
   - Limited attention given to importance of user characteristics

16

How to translate this framework to a humanitarian data ecosystem?

- Not just open data: importance of data collection neglected
- Dynamic environment
- Complex and fragmented actor environment
- Importance of trust and transparency
- Limited financial means
- Other?

17

The Humanitarian Data Ecosystem?

18

Discussion, questions and suggestions

- What do you think of the list of criteria for a successful data ecosystem I derived from literature?
  - Do you miss certain criteria?
  - Do you think certain criteria are irrelevant / overlapping?
  - Do you have suggestions for additional literature?
- How can this framework be translated to the humanitarian data ecosystem?
  - Are there other characteristics of the humanitarian data ecosystem I should take into account?
  - Which of the criteria would, and which would not apply?

19

APPENDIX C – INTERVIEW INVITATION E-MAIL

Dear Sir/Madam,

My name is Elise Haak and I am a Master's student Systems Engineering, Policy Analysis & Management at the Delft University of Technology in the Netherlands. I am currently writing my thesis for the '510' data team of the Netherlands Red Cross, who aim to make humanitarian aid more efficient through the smart use of (big) data.

Meaningful insights based on expertise and experience would be of great value for my research project. I would therefore like to ask you, could I ask you some questions in a Skype interview somewhere in the coming weeks, before August 12?

To give you an idea of what this interview will be about, I will provide you with some background information on my study:

- *Background research:*
  - The 510 data team of the Netherlands Red Cross has developed a data-driven model that can help to prioritize humanitarian aid in the direct aftermath of a natural disaster, before the official damage counts have been done. They call this model the 'Community Risk Assessment & Prioritization toolbox'.
  - In the model, different types of data are combined: static secondary data on the affected area (e.g. poverty, building materials, number of hospitals, sanitation) are combined with specific data on the disaster (e.g. wind speed, rain fall).
  - The aim is to get this model running within 12-24 hours after a disaster has struck. This means that the data on the affected area must already have been collected prior to the disaster.

- *Research topic:*
  This prioritization model is currently being developed for different countries and disasters (e.g. typhoons and earthquakes in the Philippines and floods in Malawi). To scale-up the model and be able to apply it in other countries as well, sufficient data is necessary to keep the model running, preferably data from low administrative levels: municipalities or communities. It is not always clear if this necessary data is available in a country, and if and how it can be retrieved. With my research, I am aiming to provide clarity about this process. There are two questions I am trying to answer:
  1. *How can 510 assess whether a country is eligible for the toolbox: is sufficient data available and accessible? If not →*
  2. *How can the context for the toolbox application be improved: how can the data availability and accessibility be improved, and who should take the lead in this process?*

- *Interview goal*:
  I would like to use this interview to get insight in the availability and accessibility of the data necessary for the prioritization model to function well in your area of expertise, the organizations (data providers) involved, and possible ways to improve this availability and accessibility. Using these insights and best-practices, the chance of a successful application of the prioritization model in a country can be increased.

If you have any further questions, do not hesitate to contact me. Could you please let me know if and when you are available? I am looking forward to hearing from you.

Kind regards,

Elise Haak

Graduate Researcher at 510.global
The Netherlands Red Cross
e-mail: XXXX // phone: XXXX // Skype: XXXX

APPENDIX D – INTERVIEW PROTOCOL AND QUESTIONS

"A research on scaling up the Community Risk Assessment and Prioritization toolbox"

1. Prior to interview

- Identify respondents
- Do preliminary interview to try out the protocol
- Send mail to interviewees with explanation of my research and the reason for my interview request (see Appendix C)

2. Instructions to interviewer

- Be clear with your questions
- Determine and share the time limit for the interview
- Stay focused: know what your goal is and work towards this goal
    - *Goal:* set of governance guidelines for the 510 data team of the Netherlands Red Cross on how to optimize the context for applying the Community Risk Assessment and Prioritization toolbox in a country: what would be the optimal context that allows for the improvement of the *maturity scores* on the conditions in the *data readiness tool* I developed?
    - *Assumption:* The factors that can contribute to / need to be in place to come to this optimal context are the *governance elements* of the *Data Ecosystems Framework* I developed based on a thorough literature study.
- Be flexible and open to the interviewee
- Ask for further clarification when there is something you do not understand
- Remain critical (e.g. towards inconsistencies)
- Summarize long answers: 'So if I understand correctly, you say that…'

3. Introduction

Note: to avoid confusing the interviewee with the official but long name of the 'Community Risk Assessment and Prioritization toolbox' I will use the simpler term 'prioritization model' to refer to the toolbox. Summarize this introduction where necessary.

a. Research

- *Background interviewer:* Master student Systems Engineering at the Delft University of Technology, conducting research for the 510 data team of the Netherlands Red Cross.
- *Background research:*
    - 510 has developed a data-driven model that can help to prioritize humanitarian aid in the direct aftermath of a natural disaster, before the official damage counts have been done.
    - In the model, different types of data are combined: static secondary data on the affected area in line with the INFORM indicators (e.g. poverty, building materials, number of hospitals, sanitation) are combined with specific data on the disaster (e.g. wind speed, rain fall).
    - The aim is to get this model running within 12-24 hours after a disaster has struck.

This means that the data on the affected area must already have been collected prior to the disaster.

- *Research topic:* This prioritization model has currently been developed for three different countries and disasters (typhoons in the Philippines, earthquakes in Nepal, floods in Malawi). My research is about scaling up the model to other countries. To scale up the model and be able to apply it in other countries as well, sufficient data is necessary to keep the model running, preferably data from low administrative levels: municipalities or communities. It is not always clear if this necessary data is available in a country, and if and how it can be retrieved. With my research, I am aiming to provide clarity about this process. There are two questions I am trying to answer:
    - *How can 510 assess whether a country is eligible for the model: is sufficient data available and accessible? What is the data landscape?*
    - *How can the context for the model application be improved / how can the optimal context for applying the model in a country be established: how can the data availability and accessibility be improved?*

  → *final goal: scaling up the model to other countries*

- *Interview goal interviewee not involved in rolling out INFORM Subnational:* I would like to use this interview to get insight in the availability and accessibility of the data necessary for the prioritization model to function well in your country, the organizations involved, and possible ways to improve this availability and accessibility, to increase the chance of a successful application of the prioritization model in a country: how should / could the model be embedded in your country?
  So in short, I would like to retrieve best practices for the creation of an overview of data providers and data sets, and to enhance data-sharing in a country.
- *Interview goal interviewee involved in rolling out INFORM Subnational:* I would like to use this interview to get insight in your approach and best practices in rolling out Inform Subnational in your country / region. I think this insight can be very valuable to the 510 team in their process of scaling up the prioritization model (which is also strongly relies on INFORM indicators), as this is a process comparable to rolling out INFORM subnational.

Question	Answer
1. Have you heard of this model, called the *Community Risk Assessment & Prioritization toolbox*?	
2. Do you have any further questions about the functioning of this model, or my research?	

b. Practical matters

- I will use this interview to provide practical recommendations to 510 on scaling up their

prioritization model.

- The interview will take about 1 – 1,5 hr.
- Do you have any further practical questions before we start?

4. The interview

a. Background of interviewees

- *(For interviewer) Goal of questions:* To get background information on the interviewee and his/her position.

Question	Answer
1. What is your current position?	
2. What are your responsibilities in this position?	
3. When did you start with this position?	
4. What did you do before / do you have other experience in this field / in other countries?	

b. Substantive questions

- *(For Interviewer) Goal of questions:*
  - To get an idea of the data-sharing context and related governance structure in the country of the interviewee, guided by the governance elements in my developed *Data Ecosystems Framework.*
  - To be able to retrieve practical recommendations out of the field experience of the interviewee, I will focus on best practices and ways to overcome the impediments to the establishment of these elements.

- *Introduction:*

(For interviewer) To structure my interview questions, I make use of insights I gained during a literature study on 'Data Ecosystems', which are networks consisting of multiple actors in which data is being shared. For a country to be eligible for the application of the prioritization model, such a data ecosystem needs to be established or improved.

Interviewer: According to the literature study I conducted, several governance aspects need to be present or can contribute to an optimal context for the application of the prioritization model. I would like to go through these aspects with you, to see whether and to what extent they are present in your country, how there are arranged, and how they can be improved. In short: to know what works and what doesn't. Overall goal: to retrieve 'best practices' that could also be applied in other countries.

Note: Overview for interviewer. Chosen not to show this to interviewee: this would make it too abstract. The interviewee must be able to grasp the practical relevance of the questions.

Criterion to be present	Description
Vision, communication and stimulation	A collaborative, interactive environment should be established and cooperation between stakeholders should be stimulated
Division of roles and responsibilities	There should be a clear division of the roles and responsibilities of the actors in an ecosystem
Feedback	The data ecosystem should include feedback mechanisms to enable data users to provide feedback to data providers
Leadership and incentivization	Activities in an ecosystem should be stimulated, incentivized and coordinated (either top-down or bottom-up, depending on the context)
Participatory capacity	Data supply should match data demand in an ecosystem; public bodies require certain capacities to be able to participate in an ecosystem
Sustainability: financing and value creation	In order for a data ecosystem to become sustainable, sustainable financing should be arranged and value should be generated for the ecosystem stakeholders
Trust and transparency	Data providers and data users should mutually trust and be transparent to each other

- *Core questions related to ecosystem elements:*
    - **Element: Leadership (combined with incentivization)**

Variable 1	*(For interviewer) To have a problem owner who stimulates and coordinates open data activities.* ▪ E.g. One leader or group of leaders?
Translation to toolbox context	To be able to apply the prioritization model / develop INFORM Subnational in a country, it is important that there is a local party in a country that is capable and willing to take responsibility for developing and managing the model; to take ownership. This party should for example establish/maintain partnerships with data providers to enable the required data collection.
Question	

	1. (*Interviewee not involved in rolling out INFORM Subnational*) Do you have an idea which organization (or organizations), in your country, could be a suitable local lead / local owner / manager of the prioritization model, and why? OR *(Interviewee involved in rolling out INFORM Subnational)* Which organizations took this responsibility when developing Inform Subnational in your country / region?
Answer	
Question	2. (*Interviewee not involved in rolling out INFORM Subnational*) Do you think this organization would be willing to take up this role; would they have an interest in it? A. If yes, why would this be interesting for them? B. If not, why not? a. Could you think of a way to remove this barrier / to incentivize this organization to become the local lead? OR (*Interviewee involved in rolling out INFORM Subnational*) Was this organization approached by INFORM (top-down), or did they approach INFIROM (bottom-up)? A. If they approached INFORM: what were their incentives to take on this responsibility / why were they willing to do so? B. If INFORM approached them: how did you select this organization, and incentivized them to take on this responsibility?
Answer	

- **Element: Vision, communication and stimulation - division of roles and responsibilities – on Data Providers**

Variable 2	*(For interviewer) To assign different roles and responsibilities.* ○ Identify all actors in the ecosystem ○ Locate relative position of the actors in the ecosystem in terms of roles and responsibilities ▪ Are there any accountability mechanisms present? (also a way to incentivize?) ▪ Is all the work divided clearly?
Translation to toolbox context	One of the aspects that determines the effort necessary to apply the prioritization model / develop INFORM Subnational in a country, is the extent to which an overview of (possible) data providers exists. (*related to: identification of actors in the ecosystem*).
Question	3. To what extent do you have an overview or idea of parties that possess possibly relevant data / possible data providers, on a scale from 1-5? A. If 4 or 5: a. Why this score; which factors contribute to this success / what are the best practices in getting such an overview of data providers? b. How much of an effort had to be made to get this overview? c. How could this overview be even further improved? Do these providers e.g. also possess data on low administrative levels? B. If 1, 2 or 3: a. Why this score; what are the difficulties in obtaining such an overview of data providers? b. Do you have an idea of how these difficulties could be overcome, and how to generate an overview of data providers?
Answer	

- **Element: Incentivization**

Variable 3	*(For interviewer) To incentivize key stakeholders to take up an active role and to share data.*

	○ E.g. make data sharers feel intrinsically rewarded ○ E.g. lower the barrier to entry for data-sharing
Translation to toolbox context	Data might be available, but it is only useful when it is also accessible. This means that data providers should have an incentive to share their data; they should be willing to do so.
Question	4. How willing to share their data would you say data providers, like local governments, are generally, on a scale from 1-5? A. If 4 or 5: a. Why this score; which factors contribute to this success / what are the best practices in getting data providers to share their data? b. How could data-sharing between parties be even further enhanced? B. If 1, 2 or 3: a. Why this score; what are the difficulties in having data providers to share their data? b. Do you have an idea of how to incentivize them to share their data?
Answer	

- **Participatory capacity**

Variable 4	*Public bodies should have the capability to be able to participate in the ecosystem – to meet the data demands.* ▪ E.g. technical knowledge on certain systems and technologies involved ▪ Data management knowledge (on how to ensure high data quality) ▪ Operational knowledge on how to incorporate data activities into current practices
Translation to toolbox context	

	Data providers should not only be *willing* to share the data, they should also be *capable* of making their data accessible: they should be data literate (i.e. *able to constructively engage in data-related activities*).
Question	5. How would you rate the general *data literacy* of data providers in your country, on a scale from 1-5, in terms of being able to share the data they possess? A. If 4 or 5: a. Why this score; which factors contribute to this success / what are the best practices in getting data providers to become data literate? b. How could this data literacy be even further increased? B. If 1, 2 or 3: a. Why this score; what are the difficulties in getting data providers to become data literate? b. Do you have an idea of how to get them to become more data literate?
Answer	

- **Governance Element: Vision, communication and stimulation - division of roles and responsibilities – on Toolbox Users**
- **User characteristics Elements: User needs and User capabilities**

Variable 2	*(For interviewer) To assign different roles and responsibilities.* ○ Identify all actors in the ecosystem ○ Locate relative position of the actors in the ecosystem in terms of roles and responsibilities ▪ Are there any accountability mechanisms present? (also a way to incentivize?) ▪ Is all the work divided clearly?
Condition 12	*The toolbox users should be capable of using the toolbox predictions*
Condition 13	*The toolbox output should match the user's information needs*
Question	6. Who would be the users of the toolbox predictions / INFORM Subnational?

Answer	
	7. Do you think these users are *capable* of incorporating the model predictions in their actions? Are they used to working with certain data services and tooling? A. If yes, how did they get this capability? B. If not, how could this capability be improved, or how could the predictions be altered in such a way that they meet their capabilities?
Answer	
	8. Do you think the toolbox output meets the user's information needs?
Answer	

- **Element: Vision, communication and stimulation**

Variable 5	*(For interviewer) To establish a collaborative environment in which a common goal is provided, to avoid a fragmented approach.*
Translation to toolbox context	To be able to apply the toolbox in a country, it generally helps if there is one national data platform on which data is being shared / To be able to develop INFORM Subnational, I can imagine it generally helps if there is one national data platform on which data is being shared.
Question	9. Does such a platform exist in your country / region?
Answer	
Question	10. *If platform is present:* Who has set up this platform, and why?

Answer	
	11. *If platform is present:* How would you rate the quality of this platform (in terms of data that can be found on the platform and quality of this data) on a scale from 1-5? a. If 4 or 5: ○ Why this score; which factors contribute to this success? ○ How could this platform be even further improved? b. If 1, 2 or 3: ○ Why this score? ○ Do you have an idea of how to improve the quality of this platform?
Answer	
Question	12. *If platform is not present:* Why don't you think it doesn't exist?
Answer	
Question	13. Do you have an idea how such a platform could be set up, and by whom? Would this be feasible?
Answer	

- **Element: Vision, communication and stimulation: feedback mechanisms**
- **Element: Trust & transparency**

Variable 6	*(For interviewer) Feedback mechanisms should be present to enable data users to provide feedback to data providers.*
Variable 7	*(For interviewer) Data users and data providers should trust each other and be transparent to one another*

Translation to toolbox context	• In the context of the prioritization model, it could be very useful to be able to provide feedback to data providers, as to for example the quality or structure of the data they provided. • The presence of trust & transparency between data providers and data users could have a positive influence on data-sharing.
Question	14. Are data providers generally transparent about how they get their data? A. If not, a. Why not? b. Do you have an idea how you can get them to become more transparent?
Answer	
Question	15. Having experience with dealing with data providers, is it currently possible to provide feedback to data providers on the (quality/structure of the) data they provided? B. If yes, a. How is this being done? b. What are the best-practices / how could other countries introduce these mechanisms? C. If not, a. Why not? b. (How) do you think this barrier could be overcome?
Answer	

- **OTHER: related to country assessment tool**

Condition 8	*The data should be collected at the lowest possible administrative level*
Question	

	16. Do you have tips or best practices to share on how data could be collected at low administrative levels?
Answer	

Condition 6	*It must be clear to what extent damage data is counted and how this is done*
Question	17. Do you know how damage data is counted in your country? 18. Do you think these damage counts are generally reliable? 19. How could they be improved?
Answer	

d. Closing the interview

- Are there additional aspects you would like to highlight? Or questions you expected, but we did not discuss?
- Do you have recommendations for other people I should get in touch with?
- Would you be interested in receiving my final deliverable, once I am done?
- Thank you for your time!

5. Finishing up

- How did the interview go?
- How could you improve your next interview?
- Process all the insights you gained in a structured manner

APPENDIX E – STORYLINE CREATED BASED ON EMPIRICAL FINDINGS

During the interviews, it became clear that there are several roles to be identified or assigned in the roll out of the Community Risk Assessment and Prioritization toolbox/INFORM Subnational. When discussing these roles, for each role it appeared to be relevant to assess the 'capacity' and 'incentivization' elements: when determining the suitability of a party to take up a certain role, these are the aspects that must be considered. A party needs to be both capable of and willing to take up an assigned role.
The interview outcomes related to these elements will therefore be discussed per identified role. Per role, three different aspects will be considered:

1. *What would be a suitable party to take up this role/how to identify this role?*
2. *Would this party be willing to take up this role?*
    - *If not: how to incentivize this party?*
3. *Does this party have the capacity to take up this role?*
    - *If not: how to build this capacity?*

By following this structure, the following elements of the theoretical data ecosystems framework will be covered:

- 'Division of roles and responsibilities'
- 'Leadership and incentivization'
- 'Participatory capacity'

The remaining elements that have been assessed ('vision, communication and stimulation', 'feedback', and 'trust and transparency') will be discussed separately.

Elements: Leadership – Incentivization – Participatory capacity

The first (substantive) interview question considered the 'leadership' element of the theoretical framework, of which the translation to practical context was the following:

To be able to apply the Community Risk Assessment & Prioritization model/roll out INFORM Subnational in a country, it is important that there is a local party or coalition of parties that takes responsibility for developing and managing the model: the local lead.

The first interesting remark related to this element mentioned by the four interviewees that were closely related to the roll out of INFORM Subnational is that they distinguish between two different leadership roles:

1. *The initiator/coordinator:* leading agency that initiates, coordinates and supports the development of INFORM Subnational in a region, and brings all the right actors together to facilitate data-sharing.
2. *Local lead:* national body responsible for managing and maintaining the model, to ensure its sustainability.

1. The initiator/coordinator

As described, the initiator/coordinator should be the leading agency that initiates, coordinates and supports the development of INFORM Subnational (or Community Risk Assessment and Prioritization toolbox – when learning from the INFORM approach in this context) in a region, and brings all relevant actors together to facilitate data-sharing. After the model has been developed, this party may become less involved, but should still be available to offer support to the local lead.

What would be a suitable party to take up this role/how to identify this role?

All interviewees identify United Nations (UN) agencies as the suitable party to initiate and coordinate the development of INFORM Subnational/the Community Risk Assessment and Prioritization toolbox in a country or region, due to their structural presence and strong position in the different case regions, and their information management related capacities. Different UN agencies mentioned were:

- United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA)
- United Nations Children's Fund (UNICEF)
- United Nations Development Programme (UNDP)

UNOCHA is "the part of the United Nations Secretariat responsible for bringing together humanitarian actors to ensure a coherent response to emergencies" (UNOCHA, 2017). Their mission consists of mobilizing and coordinating "humanitarian action in partnership with national and international actors to alleviate human suffering in disasters and emergencies, advocating the rights of people in need, promoting preparedness and prevention and facilitating sustainable solutions" (UNOCHA, 2017). UNICEF promotes the rights and wellbeing of children (UNICEF, 2017) and the UNDP works to eradicate poverty and reduce inequalities through the sustainable development of nations, by helping countries to develop policies, leadership skills, partnering abilities, and institutional capabilities (UNDP, 2017).

In relation to the Community Risk Assessment and Prioritization toolbox specifically, National Red Cross Societies were mentioned as theoretically being the ideal initiator/coordinator, as this would be the easiest connection for 510 Global. However, also two barriers were identified to assigning National Red Cross Societies as the initiator/coordinator.

In the first place, Interviewee 2, GIS/data analyst on a project base in Malawi, highlighted the following:

"Malawi Red Cross is not as well positioned as UN agencies in Malawi, because they only work on a project basis: they participate in a project at a certain location, but once the project is finished, they remove their staff. It is difficult to request data from government organizations saying you are from the Malawi Red Cross. UNICEF and UNDP for example, have a much stronger position in the country. They do not work on a project basis, but are structurally present. This has resulted in the government respecting these organizations: requesting data from government organizations on behalf or UNICEF or UNDP is a lot easier" (Interviewee 2, personal communication, July 31, 2017).

This image issue of a National Red Cross Society when requesting data was acknowledged by Interviewee 4, who has experience working as an information manager for the Philippine Red Cross. He mentioned that the Philippine Red Cross is very well and wide-spread positioned in the country, with lots of branches and a very strong presence due to the frequent occurrence of natural disasters – as opposed to Malawi –, and in times of disasters is when

the Red Cross becomes visible and known. Nevertheless, they are not always respected as such.
He added that even though a National Red Cross Society is the easiest way in for the 510 Global to distribute their toolbox, in the long run, to establish and maintain the entire data collaborative around the toolbox in such a way that it could also serve as a basis for other data initiatives, it would be better to have a government or UN agency as the initiator/coordinator/local lead (*note: he does not distinguish between the different leadership roles*), that has the ability to bring different organizations and data providers together (Interviewee 4, personal communication, August 8, 2017).
This relates to the second barrier to assigning National Red Cross Societies as the initiator/coordinator: they not always have the capacity to do so. The Malawi Red Cross for example, already lacks the leadership skills to coordinate activities within the organization, let alone manage different organizations, nor do they have an Information Management team. According to Interviewee 4, it is not as bad at the Philippine Red Cross, but does hold that their IT capacity is underdeveloped (though developing).
This is also one of the reasons that all interviewees did identify UN agencies as a suitable initiator/coordinator: information management is their core activity, whereas this is not the case for the Red Cross. That is why regional or national UN agencies generally outperform National Red Cross Societies (or regional Red Cross hubs) when it comes to information management capacities, and a reason they are considered a more suitable initiator/coordinator of the data-related activities concerning the roll out of the Community Risk Assessment and Prioritization toolbox/INFORM Subnational.

Would this party be willing to take up this role/how to incentivize them?

Interviewee 1, global coordinator of the development of INFORM models, including INFORM Subnational, described that the process for identifying the initiator/coordinator for INFORM Subnational mostly occurs naturally. It is driven from bottom-up: potential initiators/coordinators (mainly UN agencies) approach INFORM whether they can start developing INFORM Subnational for their region. They want to use it to support regional cooperation between humanitarian and development actors in managing risk and building resilience across the region, which is their organizational responsibility.
Apparently INFORM is already that well established in the sector – at least in those regions where this process occurred bottom-up – that humanitarian actors know it and acknowledge the value. This could be a reason for 510 to investigate the possibilities to collaborate more closely with INFORM, so they could use the 'INFORM brand' to create a stronger position for themselves.
Interviewee 7, Lead Analyst UNOCHA Eastern and Southern Africa and the coordinator of the INFORM Subnational activities for the Greater Horn of Africa, confirms that the incentive from UNOCHA to take up the role of initiator/coordinator goes back to the core values of the organization. He said that "*there is simply not enough time to provide all aid that is necessary, so it is essential to start making prioritizations. It is important to base these prioritizations on hard numbers rather than on assumptions. It is just morally the right thing to do to pursue this objective in the humanitarian sector, and to promote and improve evidence-based humanitarian assistance"* (Interviewee 7, personal communication, August 10, 2017). Interviewee 5 and 6 from UNICEF, who coordinate all INFORM Subnational activities in Latin America and the Caribbean, stated something similar, saying that it is their natural

organizational role, and that "*it is a fact that in the world of risk management you need information, you need data*" (Interviewee 5 and 6, personal communication, August 10, 2017). This was also acknowledged by Interviewee 8, coordinator of INFORM Subnational for the SADC Region in Africa.
However, this incentive relating to the organization's core values is not always sufficient for UN agencies to come to action. Interviewee 7 explained that his team has sent a briefing to other regional UNOCHA hubs to start developing INFORM Subnational for their areas (other regions in Africa) already quite a while ago, but this has not happened yet. This has two reasons. In the first place, OCHA is currently in a budget crisis. There is not a lot of money available and people are getting fired, which means it is not the right time to start with the very labor intensive process of rolling out INFORM Subnational.
In the second place, regional hubs must feel the necessity to start developing such an initiative. In East Africa, there are many disasters occurring, and many humanitarian needs, but this is less the case for Southern Africa, where there are issues of structural underdevelopment rather than real disaster situations. There is more data preparedness work than response work, and this usually seems to get less priority: the necessity for developing INFORM Subnational is less present, and this could be another reason why it has not been given priority yet.
The importance of seeing the necessity for the model development when it comes to the incentives of organizations to take up the initiator/coordinator role was also mentioned by Interviewee 5 and 6.

Does this party have the capacity to take up this role/how to build this capacity?

As explained, UN agencies – and more specifically the regional UN offices that would become responsible – are expected to have the required leadership and technical capacities to become the initiator/coordinator of rolling out the Community Risk Assessment and Prioritization toolbox/ INFORM Subnational (Interviewee 5, 6, 7 and 8).

2. The local lead

As explained, five of the eight interviewees identified a second leadership role to be of great importance: the one of 'local lead', or 'local owner', to create local buy-in (Interviewee 1, 3, 5, 6 and 8). This party should be responsible for managing and maintaining the model (either the Community Risk Assessment and Prioritization toolbox or INFORM Subnational). For example, connections with data providers and model users need to be managed and maintained.
The remaining interviewees did not necessarily exclude this role, but they envisioned the responsibilities of the initiator/coordinator and local lead in one role. The reasons for the other interviewees to separate the two roles, were the following.
Interviewee 5 and 6 repeatedly emphasized the importance of involving local parties from the very beginning of the process of rolling out INFORM Subnational. They mentioned that the tool is not a UN tool, but a tool that will be used locally, for national purposes. It is therefore important that there is a local party that feels ownership of the tool, that understands the value and potential of the tool, that becomes the local manager after the initiator/coordinator has set out the initial development and that ensures local support. This way, the model can become sustainable: "*Involving a local party from the start is a time-consuming process, but it is the only way to ensure the sustainability of the model"* (Interviewee 5 and 6, personal

communication, August 10, 2017). This is also acknowledged by Interviewee 3, Information Management Analyst for UNOCHA Philippines, who states that you want to have a local owner, like a government agency, leading the model, to ensure its continuity.
Apart from generating a sense of ownership and creating local buy-in, another responsibility for the local lead would be to maintain the tool; on an annual basis, the tool needs to be updated and revised.
Interviewee 8, Chief of the Information Management Unit of UNOCHA in the SADC region, adds another reason why there should be a local lead. Eventually, once the product is finished, the outcomes of the model should be presented to and implemented by all relevant partners in a country. According to Interviewee 8, it is much more powerful if this is done by a renowned national actor like a government department, instead of by an international actor like the UN. On developing INFORM Subnational for a country, he states that "*if we do not have the local buy-in from day one, we cannot proceed. We really need the local buy-in from a local owner like the National Statistics Office (NSO), they should be the advocates of the model*" (Interviewee 8, personal communication, August 11, 2017).

What would be a suitable party to take up this role/how to identify this role?

Six of the interviewees (Interviewee 1, 3, 4, 5, 6 and 8) identified 'a government agency' (or 'national body') as being the most suitable local lead (the remaining interviewees did not specify a local lead and only identified a UN organization as the most suitable initiator/coordinator, so they did not contradict this). The interviewees argue that the model outcomes are expected to be used mainly by government partners, and usually many of the data providers are also governmental organizations. For the success of the model it is very important that these parties participate, contribute to and use the model. Having a government agency as a local lead of the model therefore helps to involve those other government actors as well.
The most suitable government agency to take up this role varies per country; it is dependent on the setting in a country and on the responsibilities and capacities of the different government bodies, which will be discussed later. Deciding on the most suitable local lead is a process. Interviewee 5 and 6 explained that as a first step in the development of INFORM Subnational for El Salvador, their government is now deciding: what would be the best organization to position the tool in terms of management, promotion and maintenance? To make this decision, they have had an introductory meeting with all relevant government bodies. Based on this introduction they are entering the dialogue on who is the most suitable local owner.
Examples of suitable government agencies that are mentioned are:

EXAMPLE OF SUITABLE GOVERNMENT AGENCY	MENTIONED BY	FOR REGION
• Office of Civil Defense • Department of Social Welfare and Development (DSWD)	Interviewee 3	Philippines
• A Ministry of Planning • An Office for Disaster Management • The National Statistics Office (NSO)	Interviewee 5 and 6	Latin America and Caribbean
• NSO	Interviewee 8	SADC

TABLE 10 – Overview of suitable government agencies proposed by interviewees

Would this party be willing to take up this role/how to incentivize them?

As well as for the initiator/coordinator, to be willing to take up the role of local lead, it also holds that a party must feel the necessity for the model development (either the Community Risk Assessment and Prioritization toolbox or INFORM Subnational). In addition, the potential local lead must be convinced of the value and sustainability of this specific model: what will be the added value when they start to invest in it? This is an important aspect mentioned by all interviewees. As Interviewee 8 describes: "*You must make clear that being the local lead for INFORM Subnational in the end does not cost them work, but it will improve their work, once they have everything embedded in their systems and processes. The challenge is, however, to have them realize this"* (Interviewee 8, personal communication, August 11, 2017).

Interviewee 2 mentioned about Malawi that the local parties there have already seen so many data initiatives passing, that their attitude towards new data initiatives has become skeptical. They need to be assured of the unique value of the model, and that it is an initiative that will last.

Interviewee 3, 5, 6 and 8 all state that this is something that will automatically happen over time: as the model becomes richer and richer, includes more data and starts to generate better outcomes, government parties will automatically become interested in becoming the local lead. However, these same interviewees stress that it is important to create local ownership from the very beginning, and to involve a government body from the start of the process of developing the model. Thus, a local government party is necessary to be able to develop the model, but to get this party involved, it helps if the model has already been developed so the value can be shown. This is a clear example of a vicious circle that needs to be broken, by being able to demonstrate (potential) local leads the value of the model already in a very early phase of the model development.

Interviewee 4 proposes a solution to work around this vicious circle, namely adopting a stepwise approach: "*Seeing the Red Cross National Society is currently our best entry into a country, it could be valuable to start building what we can with the Red Cross as our local lead: start building the Community Risk Assessment and Prioritization toolbox with the data we can then already gather. If this starts to work, it automatically becomes interesting for other organizations to become involved and then the data collaborative gets bigger and bigger, which leads to the resulting model improvements. In this improved situation, it is more likely that another, more suitable, local owner starts to see the added value of the model, becomes interested in a wide application of the model and becomes willing to take up this role."* (Interviewee 4, personal communication, August 8, 2017).

Another important aspect in the incentivization of potential local leads is the aspect stressed repeatedly by Interviewee 5 and 6: to involve them from the very beginning. This way, they can be convinced by the value of the model in an early stage by clearly explaining to them how it works and how it can help them in pursuing their objectives. Additionally, they start to feel responsible for developing the model. It should be avoided to create a tool and then later start looking for the local buy-in. In Colombia, the rolling out of INFORM Subnational was mainly driven by the work of independent consultants. Interviewee 5 and 6 felt that this led to a limited sense of ownership from the government in the process, and this was a problem.

The importance of the generation of local involvement is also highlighted by Interviewee 2. He says that the strength and uniqueness of the Community Risk Assessment and Prioritization toolbox is the spatial data aspect. He hopes that the organization of a workshop in Malawi with

all relevant stakeholders helps to convince potential local leads about the power of the model, and thinks that this could also be a means to create local ownership in other countries.
Furthermore, according to Interviewee 5 and 6, an incentive for local leads, especially governments, to take up this role could also be that it could help them to meet goals established in global agreements, like for example the Sendai Framework for Disaster Risk Reduction that entered force in 2015. Interviewee 5 and 6 use this to convince potential local leads that they can help them to move forward; to give them the means to progress, in a way that is very convenient for them: "*Also, we make it easy for them: we hand them the complete model, and they only have to manage and maintain it, which are processes that are quite straightforward and relatively easy. Yes, it requires some work, but eventually you need the information anyway"* (Interviewee 5 and 6, personal communication, August 10, 2017).
Lastly, Interviewee 5 and 6 also think that an incentive for a government agency to become the local lead could be to show off to other organizations or governments. They think that eventually every government wants to show that they are understanding the risks in their country, and wants to give a good example, especially if it is doable.

Does this party have the capacity to take up this role/how to build this capacity?

In the process of determining which government agency would be a suitable local lead, this is an important criterion. The reason why in two of the interviews the NSO has come forward as a suitable local lead (see *Table 65*), is because, besides being a government agency, they generally have the capacity: collecting and processing data is their core activity. Additionally, collecting data is in their best interest.
Another criterion to determine on a suitable local lead would be the affinity with/interest in the model outcomes, as explained before. That is why Interviewee 5 and 6 proposed an Office for Disaster Management, and how Interviewee 3 came up with the Office of Civil Defense, who in the Philippines are the chair for disaster risk reduction and emergency response and the ones doing pre-disaster assessment planning. Position-wise they would be a very suitable local lead, but they have a very limited technical capacity and hence they would not be able to carry out the tasks of the local lead. In the Philippines, the DSWD has a lot of technical and GIS employees that also do predictive analyses themselves. Also, their activities match with the outcomes of the Community Risk Assessment and Prioritization toolbox/INFORM Subnational, and they are therefore considered a much more suitable local lead.
This illustrates how the decision on a suitable local lead may differ per country, based on the structure of the government and a combination of the responsibilities and capacities of the different bodies.
If needed, the local lead should be trained (capacity-building) on for example basic excel skills and mapping by the initiator/coordinator (Interviewee 8, personal communication, August 11, 2017). In any case, the initiator/coordinator should keep offering support to the local lead, in the form of technical support or any other consultancy, in case for example issues with the maintenance arise (Interviewee 1, 5, 6 and 8).

To summarize

To create local buy-in, ensure the sustainability of the model and link it to existing data initiatives, it is important that there is a local party that feels ownership, that understands the value and potential of the tool and can locally promote it. The party taking up this role should

be a government agency, as it is expected that such a national body can count on most support from the other stakeholders. It will be country-dependent of which government agency the activities and capacities match best. This should be decided on in a process with all relevant stakeholders.

Elements: Division of roles and responsibilities – Incentivization – Participatory capacity

3. The data providers

The second element to be addressed in the interview was the 'division of roles and responsibilities', of which the identification of all actors in the data ecosystem is an important aspect. In this part of the interview, the 'division of roles and responsibilities' element was translated specifically to the identification of data providers:

One of the aspects that determines the effort necessary to apply the prioritization model/develop INFORM Subnational in a country, is the extent to which an overview of (possible) data providers exists.

What would be a suitable party to take up this role/how to identify this role?

The first question concerning this topic posed to the interviewees, was to what extent they have an overview of relevant data providers in their country/region. If they already had quite a complete overview, they were asked to explain how they got to this overview, and if they would adopt this same approach if they would have to do it again. If they did not have a good overview yet, they were asked about the difficulties, and about how they would recommend to get to this overview.

All the interviewees stressed that getting a good overview of relevant data providers in a country takes a lot of time and effort. It is a time-consuming process during which you slowly start to figure out what are the best data providers. Data is usually scattered across many different organizations, and initially it is not clear what these organizations are.

Interviewee 2 rates the overview he has of relevant data providers in Malawi a four out of five, but he highlights the investments he and his team have made to come to this overview: two field trips, one of three weeks and one of three months, and a remote online data collection that took months, were necessary. Interviewee 3 rates the overview he has of the data landscape in the Philippines a four to five, describing he slowly gathered this overview during his years of experience in the sector.

Interviewee 2 and Interviewee 3 both recommend the same approach when starting to create an overview of relevant data providers in a country:

1. Start with a search for existing open data platforms in the country or with open data available on the 'Humanitarian Data Exchange' (HDX), the humanitarian open data platform, to get a basic preliminary overview. Of course, there could be a lot more data providers that are just not openly sharing their data, but at least this gives an indication.

2. From there you can start approaching the main data providers in a country (NSO, Department of Surveys, some disaster department, ministries). Then you ask them about the most important data providers in the country, about data initiatives, platforms and data collection working groups. From there you continue.

However, on this first step of using open data platforms as an indicator for the richness of data in a country, Interviewee 4 mentioned that this is not always helpful: open data portals do not always truly represent the quality and amount of data available. In Malawi for example, he illustrates, there are quite a lot of known open data portals, but experience learns that they almost always disappoint: much of the data is not useful. It is old, there is no national coverage, it is not of a high level of granularity, etcetera. However, this is difficult to find out. To realize this, you need to deep-dive into the data, which takes time. So, he warns, at first sight these kind of data portals might give the impression that there is a lot of data preparedness in a country, but this can be very misleading. This is also acknowledged by Interviewee 8.
Additionally, Interviewee 4 would like to take the second step one step further. He proposes to go into the country and start a workshop with all relevant parties together, in which the first topic would be to explain what is meant by 'data' and what data sources are considered suitable. Based on this workshop, you could start making an inventory of the data providers and the data they have available.
This suggestion is in line with the approach that the coordinators of the INFORM Subnational activities in Latin America and the SADC region in Africa already adopt. Even though they do start with a first assessment of the data in a country that is publicly available, Interviewee 5 and 6 always kick-off the development of INFORM Subnational with the organization of a workshop in which there will be a discussion on the conceptual framework of INFORM. In this workshop, key stakeholders and representatives of different sectors are invited to ask for their opinion and suggestions, and they also use this meeting to together identify relevant data providers. Interviewee 8 described the same approach for the development of INFORM Subnational for countries in the SADC region: they start with a workshop with key data providers, like the NSO, the United Nations Population Fund (UNFPA, in charge of census and population numbers), relevant ministries like the Ministry of Health and UNICEF, and from there they proceed. Interviewee 1 confirmed that this is the approach all INFORM Subnational coordinators usually adopt.
All in all, all interviewees described similar approaches: start with a search on online data platforms to get an initial idea (but be aware that the data on these platforms can be misleading), from there start approaching the main data providers (which usually include the NSO, an organization collecting population census data, a disaster department and relevant ministries), and then have a workshop with them to identify other possibly relevant data providers.

Would this party be willing to take up this role/how to incentivize them?

The 'incentivization' element was translated to the practical context of this study as follows:

Data might be available, but it is only useful when it is also accessible. This means that data providers should have an incentive to share their data; they should be willing to do so.

The questions concerning this topic posed to the interviewees, were to what extent they would say data providers in their region are willing to share their data, what are the difficulties the interviewees encountered, and what are the incentives of the data providers to share their data (or how the interviewees thought they could be incentivized). Two of the interviewees, Interviewee 4 and 7, only have experience with the collection of open data, so they could not answer these questions.
All remaining interviewees highlight that the willingness to share data varies greatly per data provider, and even within the government. The following difficulties were encountered by the different interviewees in the process of requesting data from data providers:

- **Saying 'yes', doing 'no':** Interviewee 3 (UNOCHA Philippines) and Interviewee 2 (Malawi Red Cross) both explain that there is a difference between data providers saying they will share the data with you, and doing it. Interviewee 2 stated: "*When you talk to people face-to-face, they all say they will of course share their data. But in the end, you do not get anything. You follow up and they will say 'I forgot, I will do it later', and you end up in this time-consuming process of requesting data and not getting it*" (Interviewee 2, personal communication, July 31, 2017).
- **Bureaucratic hassle:** On this same issue, Interviewee 3 describes: "*If you organize a forum and you invite a lot of partners that are data providers, they all say: 'yes we have the data and yes we will share it'. But saying it is something different than doing it. So, you need to ask again, and then suddenly they are not willing to share it anymore. Then you write an official letter, and they reply to you by sending you the printed, scanned copy of the data. Of course, you cannot use that, as it will cost a lot of time to code it digitally, so you request an excel version. And then suddenly they have protection issues and you need to contact their supervisor about the data. With the supervisor, the process of requesting data starts all over again. In the end, they do give you the data, but it takes a lot of time and effort: it can take months, or sometimes even years to go through all the bureaucratic hassle"* (Interviewee 3, personal communication, August 1, 2017). This illustration also includes another issue, namely:
- **Wrong format:** Sometimes data providers do share their data, but in a format that is unusable (Interviewee 3).
- **Sensitivity of data**: Sometimes data providers are hesitant to share their data/very protective of their data because a dataset contains sensitive personal information on for example people in conflict zones (Interviewee 2, 3, 5 and 6). A related issue is that data providers might be afraid you will use their data commercially (Interviewee 8). In Malawi, UNICEF was hesitant about trusting Open Street Map, as they were afraid the information on their maps was used for commercial purposed (Interviewee 2).
- **Closed attitude of data provider:** The willingness of data providers to share data is also dependent on the attitude towards data-sharing of the person in charge. It can be as simple as a personality issue (Interviewee 3, 5 and 6). Interviewee 3 illustrates this stating that in the Philippines, some organizations (also government agencies) have the backwards mentality of 'information is power', so they want to keep the information to themselves. This can be especially frustrating in a country like the Philippines where so much data is being produced, but not being shared.
- **Something to hide:** Sometimes government agencies or departments do not find it convenient to publish or share certain data, because this would reveal that certain

developments they are responsible of are not going in the direction they had planned for (Interviewee 5 and 6).

- **No priority:** Data providers often feel that sharing data costs them extra work, whereas there is nothing in it for them (all).

The main approach to, despite these difficulties, incentivize the data providers to share their data, resembles the suggested approach to incentivize a party to become the local lead: generate a sense of ownership, and involve the parties from the very beginning of the model development process. To be able to come to agreements with data providers on sharing their data, they need to start feeling engaged in the process. This was mentioned by all interviewees.

By involving them and organizing for example a workshop with them, you get the chance to start a dialogue. You get to be transparent and show and convince the data providers of the added value of open data and of the quality of the model, and explain them how they can contribute.

In addition, it should be explained to and agreed with them what is in it for them. According to Interviewee 2, 3, 5 and 6, this should be something very concrete. Interviewee 3 mentions that UNOCHA in the Philippines always acknowledges the organization they got their data from. However, this is often not enough: data providers do not just want the credits, but they also want to have something that is useful to them, like a specific data analysis that they need. Interviewee 8 adds that this must be approached in a straightforward manner: "*If the data providers eventually want to benefit from the model you create, they must realize they have no other choice than sharing their data with you, it is as simple as that*" (Interviewee 8, personal communication, August 11, 2017). This will be further elaborated on when discussing the 'model user' role.

Related to the close involvement of data providers from the start, Interviewee 2 mentioned about Malawi that when requesting data from data providers, it is necessary to meet people face-to-face. He says that nine out of ten times sending an e-mail will not work: they will just tell you they did not receive it.

Another tip he gives is to assign one spokesperson per organization that is your point of contact for the data-sharing, and make clear to this person that this is his or her responsibility. This gives you the opportunity to personally address people when you did not get your data, and hopefully gives them an extra incentive to try harder.

Moreover, he thinks it helps to frame things in a certain way: in Malawi, it is more powerful to call a workshop an 'NSO workshop' than a 'Red Cross Workshop'. With an 'NSO Workshop' title you can count on more support, as this organization has a much stronger position in the country than the Red Cross.

Lastly, he says that in Malawi he experienced that money can be the key to a lot of things. He noticed that providing something like a free lunch is one of the most effective ways to incentivize local people to cooperate. Additionally, in Malawi it is usual to incentivize stakeholders to participate in a workshop by providing them a daily allowance.

A note to be here is that the Red Cross never pays for data.

As to the issue of privacy sensitive data, Interviewee 3 provides the solution of making agreements on anonymizing all the data, to remove the protection issues. Before sharing the data, the data providers can remove everything from a dataset that allows for the identification of persons or families. Additionally, close involvement from the beginning of data providers

could help them to become convinced of the fact that data will for example not at all be used for commercial ends.
Lastly, Interviewee 1, 3, 5, 6 and 8 stress that it is important to keep in mind that collecting data and developing the model is an iterative process, and that it is important to remain flexible. If a certain data provider is determined about not sharing their data, it is mostly not worth the effort to keep trying. In that case, a way to work around this should be found: look for an alternative data provider, or an alternative indicator to include in the model. Model indicators should be adapted to the availability and accessibility of the data. Then, at some point, the model development can start with the data that has been collected, and additional data and indicators can be added once possible.
The most suitable approach to collect data varies per country and per organization. It should be assessed on a case-by-case basis which approaches can be applied best (Interviewee 1).

Does this party have the capacity to take up this role/how to build this capacity?

The 'participatory capacity' element was translated to the practical context of this study as follows:

Data providers should not only be willing to share the data, they should also be capable of making their data accessible: they should be data literate (i.e. able to constructively engage in data-related activities).

The questions concerning this topic posed to the interviewees, were to what extent they would say data providers in their region are data literate (in terms of being able to share the data they possess), what are the difficulties and what would be ways to build this capacity; to get the data providers to become more data literate.

Perceived capacity

The findings on the perceived data literacy of data providers per country/region are summarized in *Table 11* below. Especially the interviewees specialized in one particular country (Interviewee 2 and 3), provided an extensive description of the data literacy in their country. All interviewees involved in rolling out INFORM Subnational, coordinate multiple countries (Interviewee 1, 5, 6, 7 and 8) They did not give such an extensive answer per country, as they are all involved in the coordination of over eight countries, and they all stressed that the data literacy varies greatly per country. This marks the variety of cases. Additionally, they mentioned that the data literacy within a country varies a lot per organization and per administrative level, which is in line with the two main findings from Interviewee 2 and 3.

COUNTRY/REGION	DATA LITERACY OF DATA PROVIDERS
Malawi (Interviewee 2)	Average: **+** Organizations do more and more see the value of open data. They try to figure out what value open data could have for them and how they could share data. **–** There are a lot of differences between organizations (UNICEF, UNDP, CARE[6] and MSF[7] are very data literate; they have a large GIS department. The Department of Survey is also quite good, but the Malawi Red Cross is dramatic: data is just non-existent there. **–** The is a limited number of concrete data initiatives. The data initiatives that do exist are very much fragmented/scattered; they do not work together. **–** Everyone relies on the government to take the lead in the development of data initiatives, but government agencies do not cooperate and do not have the capacity. All their data is stored locally on drives, and it happens regularly that hard drives or computers crash and all the data is gone.
The Philippines (Interviewee 3)	Quite high: **+** The data environment in the Philippines is considered very good. It is ahead of other countries. For OCHA in other countries, information managers struggle with the data from their operations. To them the Philippines serve as an example. **+** Progress is being made in the development of standards for, for example, the identification of administrative boundaries. This could make it easier to merge data from different agencies. Not all organizations use this structurally yet, but there is some improvement. Yet, currently there are still structural problems with the comparability of data. **–** There is an overload of information without clarity on which data to use for which purpose. There is no coordination between organizations. **–** Lots of differences between organizations. It for example differs per organization if they are working with simple excels or with very high-tech software. **–** Great variations in the data literacy between different administrative levels. At national offices, the data literacy can be very high, but if you go to the provincial levels, data would still be stored paper-based. If you request data from these agencies they would send it to you through fax, or just give you a big pile of folders.
SADC Region (Interviewee 8)	• Varies greatly per country, and lots of differences between organizations and administrative levels. Many organizations store data both digitally and paper-based. • In Zimbabwe and Zambia there is quite a lot data available online. Their NSOs are quite advanced, especially relative to other countries in the SADC region, and they see the value of data-sharing. • But in Angola and Congo for example, there are a lot of difficulties in the publication of data: their governments are not accepting the open data policy. • Also, there is a great difference between coverage within countries. Mainly Zambia is a good example for the rest of the region. Some of their assessments are done using tablets, which makes it easier for them to do data collection and leads to a very high coverage and a good real-time assessment. In the SADC region, Zambia has the highest capacity: there is quite some IT capacity – using GPS for data collection, mapping software, and software to design questionnaires. It is easier for countries to share data when their data collection is more digitized.
Others	No specific nuance, otherwise than varying greatly per country, organization and administrative level.

TABLE 11 – Remarks on data literacy made by the interviewees

To summarize, in every country the data literacy of data providers could be improved. Which specific organizations need capacity-building should be assessed on a case-by-case basis as

[6] CARE (Cooperative for Assistance and Relief Everywhere) is an international humanitarian agency delivering emergency relief and carrying out long-term international development projects to fight poverty and educate people.

[7] MSF (*Médecins Sans Frontières)* is an international, independent, medical humanitarian organization.

this varies greatly per country and per organization, but organizations at lower administrative levels generally have less (or none) Information Management (and hence data-sharing) capacities.
All interviewees mention that the NSO usually has technical staff in their offices that knows how to work with data, so that it is important to at least always involve them.

Capacity-building
On how to build the required capacity, Interviewee 5 and 6 point out that when having contact with a data provider, it needs to be made sure that there is contact with their technical staff that is responsible for data management. On a case-by-case basis it must then be determined which capacities need to be strengthened: some could for example be advanced in mapping, but having a limited capacity to construct the model. If it is unclear who has these capacities, the team that should have them should be identified. A training involving all technical staff (of e.g. multiple government agencies) can then help to build the required capacities. During the time of the interview, Interviewee 5 and 6 were for example preparing a training for all technical staff from the government in Honduras, to go through the entire process of constructing the model, managing the different indicators and data and using the model (which role will be discussed next).
Interviewee 8 describes a similar process for the SADC Region in Africa, where they recently organized a big workshop to train all the information managers of involved agencies like NSOs and government departments on topics like data mapping and database maintenance. Also in the Philippines, there is an 'Information Management Technical Working Group' with UNOCHA and government agencies, and at a national level lots of efforts are being put in training people too.
This capacity-building would be the responsibility of the initiator/coordinator.
However, Interviewee 2, 3, 5, 6, 7 and 8 also stress that this training process can be challenging, and is not always as straightforward as it sounds: if an organization almost starts from scratch when it comes to technical, data and information management capacity, a shift in organizational culture and in mentality is required to make this a higher priority. Realizing this can be very difficult, and the interviewees do not have concrete suggestions on how to do this. As Interviewee 3 noted: "*You can of course train people, but effectively doing so is easier said than done. Even if the required technology is available and people know how to use excel, they still do not use it in a correct way. They are not sufficiently motivated, or do not have the right mentality. How to change this? That is a very good question: I have no idea"* (Interviewee 3, personal communication, August 1, 2017).
Additionally, he points out the problem that in the Philippines most good, qualified Information Managers that are trained well, get hired by private companies who pay better than government agencies or humanitarian organizations. Interviewee 2 states the same about Malawi. When he compared organizations in Malawi with a high data literacy with organizations with a low data literacy, he noticed that the organizations with a high data literacy either have staff that is not local, or they pay them very well so they can attract good, qualified staff.
To conclude, building capacity to have data providers become more data literate is a challenging process. Providing training and workshops could be a means to strengthen capacities, but to make organizational operations more data-driven, also a shift in organizational culture and mentality is required. The interviewees do not have concrete ideas about how to facilitate this shift, apart from involving agencies from the very beginning in the

model development. Once they see and understand the value of more data-driven operations, they might become more motivated to put effort in improving their capacities.
Apart from the required shift in mentality, also money can play a key role in not losing people that are qualified to organizations that pay better.

Elements: Division of roles and responsibilities – Incentivization – Participatory capacity (Governance) – User needs – User capabilities (User characteristics)

4. The model output users

Another role to be defined in the ecosystem around scaling up the Community Risk Assessment and Prioritization toolbox, are the toolbox users. In the developed data ecosystems framework, the 'users' in the data ecosystem are captured with the 'user needs' and 'user capabilities' elements. However, for the data ecosystem around the toolbox, it could be argued that the discussion of the toolbox users is also part of the governance part of the framework: it is a role that needs to be assigned, and for which the same questions need to be answered as for the other roles discussed: who are the potential users, are they willing to use the toolbox (if not, how can they be incentivized) and are the potential users capable of using the toolbox (if not, how can this capacity be built)?

What would be a suitable party to take up this role/how to identify this role?
Would this party be willing to take up this role/how to incentivize them?
Does this party have the capacity to take up this role/how to build this capacity?

Duplication of roles
All interviewees mentioned a very wide range of potential users of the Community Risk Assessment and Prioritization toolbox/INFORM Subnational, as the models can be used for many different purposes, like resilience, early warning, early recovery, and risk management. Essentially, every agency involved in those activities could be a potential user. Yet, the most likely (or suitable) potential users mentioned, are mostly also organizations that have been proposed as the most suitable party to fulfill one of the other roles (broadly to be divided into humanitarian organizations and government agencies). This makes sense, since one of the reasons for the interviewees to consider this party suitable, is that the model content matches with or contributes to their responsibilities (the other reason being are they capable, which will be discussed later). Interviewee 3 and 8 for example bring up UNOCHA and UNDP as potentially being interested in becoming a user of the Community Risk Assessment and Prioritization toolbox or INFORM Subnational model respectively, whereas these organizations were also proposed as being a suitable initiator/coordinator of the development of the models. Interviewee 3 additionally suggested the two government departments that he also identified as the most suitable local needs to become users of the model (Office of Civil Defense and the Department of Social Welfare and Development, due to their activities and responsibilities relating to disaster management and risk reduction). The experience of Interviewee 5 and 6 with rolling out INFORM Subnational in Latin America is that government agencies with these kind of activities, for which the type of agency or department varies per country, are generally thought of as suitable potential users. Interviewee 5 and 6 explicitly stress that sometimes

these government agencies even take up a triple role: the one of local lead, the one of model user, and the one of data provider, as these agencies are also often in possession of relevant data to feed into the model. This provides opportunities to involve organizations in the process and incentivize them: if a data provider or local lead is also a potential user of the model, it should be relatively easy to convince them that 'there is something in it' for them. If they want to become a user of the model, they should also be willing to accept taking up the role of data provider or local lead. This could be used in the incentivization of other roles.

Different roles require different capacities

So, the suitability of parties to take up a certain role is partly determined by their willingness to do so, which relates to their activities matching with the use and content of the model. A second aspect that contributes to an organization's suitability, is their capacity. This second factor makes that there are also parties that are identified as potential users, who are not suggested as a local lead or data provider.

The National Red Cross Society for example, was by most interviewees dismissed as a suitable local lead, partly due to their usually limited data and information management capacities. However, seeing their activities, the output of the Community Risk Assessment and Prioritization toolbox can be very valuable to them. Adding this to the fact that 510 Global has close Red Cross connection makes them a likely potential user (Interviewee 2 and 4).

This is explained by Interviewee 5 and 6, who say that the different roles require different capacities. Being the user requires less advanced technical capacities than being the local lead or data provider. The users do need to be familiarized with the model and be explained how they can use it in their processes, but they do not need knowledge on how it is developed. Interviewee 5 and 6 make clear that it is consequently also possible that one organization takes up different roles, but that the responsibility for these roles lies at different departments within that organization, based on their capacities and responsibilities.

User-friendliness is key

Moreover, for the user role, 'willingness to use' (or incentivization) and 'capacity' are very much interrelated: their willingness to use is determined by the ease with which they can use the model. The less capacities necessary, the better: user-friendliness can incentivize them to start using the model output. The importance of user-friendliness in ensuring the use of the model is stressed by many of the interviewees: Interviewee 2, 4, 5, 6, 7 and 8 all mention it.

User-friendliness can be created in different ways, but it comes down to presenting the model output to the users in a way that is appealing to them. Interviewee 2 thinks that users of the model generally want a combination of visualization and explanation. Both the Community Risk Assessment and Prioritization toolbox and INFORM Subnational currently already have a publicly available dashboard in which the model outputs are visualized in a way that is, according to Interviewee 2, quite straightforward and easy to interpret. However, it could be possible that the methodology behind the visualization is not entirely clear to a user, and in that case, it should be possible for a user to request additional assistance.

Interviewee 5, 6, 7 and 8 add that the presentation of the model should be tailored to the needs of the different potential users, if you want to appeal to for example different sectors and government bodies. Different users might want different types of presentations: some want to see the maps, some the infographics, and others a two-page narrative, as long as it is translated into something that is accessible for them. According to Interviewee 5, 6, 7 and 8,

who were all involved in rolling out INFORM Subnational, it is the responsibility of the initiator/coordinator to make sure this is taken care of.

Involve from the start to create sense of ownership and meet user needs

Apart from the user-friendliness of the model outputs, another way to incentivize users, proposed by Interviewee 1, 2, 5, 6 and 8, is once more to involve them in the development of the model from the start. If they get a sense of ownership and see the value of the model, they are more likely to start using it. This is a returning aspect considered important for each role, which could also relate to the fact that they are overlapping.

This involvement could be realized by the organization of meetings or workshops with possible users (ministries, humanitarian organizations), to discuss how the model outputs could be valuable to them, and to have them have their say in the process. This approach is already being adopted when rolling out INFORM Subnational for a country or region.

These workshops could also be a means to meet the user needs on a content-level. When developing INFORM Subnational, the model always must be adjusted to the local context. This decision on which indicators to include in the model is made in consultation with possible users. It is important to get their ideas, input and suggestions to get an idea of which indicators would be relevant to them. Interviewee 4, main developer of the Community Risk Assessment and Prioritization toolbox, acknowledges that suggestions by the user should always be considered. However, he also highlights that the inclusion of indicators suggested by users must also be feasible and value-adding.

The workshop to include model users can be merged with the meeting to include possible local leads and data providers. This is something that is unavoidable due to the overlapping roles, but also important: the inclusion of indicators is merely driven by the user needs, but also by the availability and accessibility of data. It is efficient if the two can be discussed simultaneously (Interviewee 1 and 8). Another aspect to take account of in this decision-process is the retention period of the indicators. Interviewee 1, 3, 5, 6, 7 and 8 all stress the importance to choose indicators that do not have to be updated very regularly, so they can be maintained for a longer period.

Challenging: shift in mentality required

Lastly, according to Interviewee 5 and 6, ensuring the model will eventually be used is the biggest challenge in the model development because, just as with the data providers, a shift in mentality needs to be established to get users to incorporate the model in their operations. They explained: *"In Colombia we thought that once we would have developed this tool that, everyone would immediately start using it. Unfortunately, that was not the reality. You must develop a culture around it, around its use and integration in organizational practices* (Interviewee 5 and 6, personal communication, August 10, 2017).

To summarize

Most of the identified suitable users of the model outfit, are parties that have also been identified as most suitable in one – or multiple – of the other roles. Additionally, National Red Cross Societies were identified as potential users, whereas they were not mentioned suitable for any other of the roles due to their (usually) limited information management capacities. This illustrates that different roles require different capacities.

To meet the capacities of the potential users and their needs, it is important that the model outcomes are presented in a user-friendly way, preferably in a way tailored to the needs of a specific user.
Additionally, users should, just like the other roles, be involved from the start of the model development, both to get a sense of ownership (necessary to ensure they will start using it and facilitate the required shift in mentality) and to give them the opportunity to provide their input on which data should be include in the model to have their needs be met. Whether this will be possible also depends on the availability and accessibility of the data requested by them.

Elements: Vision, communication and stimulation

The 'Vision, communication and stimulation' element relates to the establishment of a collaborative environment in which a common goal is provided, to avoid a fragmented approach.
This was translated to the practical context of the Community Risk Assessment and Prioritization toolbox as follows:

To be able to apply the toolbox in a country, it generally helps if there is one national data platform on which data is being shared.

In hindsight, however, after having analyzed the interview questions already discussed, this translation appeared not to be entirely accurate: as some of the interviewees pointed out, the data available on open data platforms not always truly represents the data available in that country: it might be misleading. Additionally, interviewees did not provide answers to the questions (*Is such a platform present? If not, if and how should it be realized? If yes, who has set it up and how can the quality be improved?)* that 510 Global would not be able to find out themselves: they explained the existing open data platforms in their region, but did not have suggestions on how to improve these platforms or how to set them up other than capacity-building of data providers, which has already been covered in the previous section. This led to the realization that, indeed apart from the capacity-building part, setting up a local data initiative is something which is not directly in the power of 510 Global to govern.
Therefore, the answers to the questions related to this element will not further be discussed.
So, to conclude: during its application, the translation to practice of this element was not considered very useful or accurate. Nevertheless, the content of the theoretical element (to establish a collaborative environment and avoid a fragmented approach) cannot be dismissed, and its importance has only been emphasized in the empirical outcomes: one of the main findings from the previous section was that to incentivize organizations to take up the role that is expected to suit them best, it is key to involve them early in the development process to give them a sense of ownership and have them realize the value of the model. This very much relates to the importance of the provision of a common goal and the prevention of fragmentation to occur, which means the interviewees basically recommend meeting the 'vision, communication and stimulation'.

Elements: Trust and transparency – Feedback

These elements were assessed together, as they are considered related.

They were translated to practice as follows:

Trust and transparency: *For the data providers to be transparent about the origin of the data they share and for the data requesters to be transparent to the data providers about what they are aiming to use this shared data for can have a positive influence on data-sharing.*

Feedback: *In the context of the Community Risk Assessment and Prioritization model, it could be useful to be able to provide feedback to data providers, as to for example the quality or structure of the data they provided.*

Firstly, the interviewees were asked whether data providers are generally transparent about how they get their data, and if they had ideas about getting them to become more transparent. The concept of trust was thought to have been covered already in previous interview questions. Secondly, questions posed on the possibility to provide feedback to data providers on the data they shared: is this currently possible, how is it being done and how could this be improved?
The general answer to these questions was that it is difficult to get transparency of data providers on where they got their data from, and to provide feedback on the (quality of the) data they shared. It varies of course per organization (and on top of that, per administrative level). Some organizations, like national organizations like the NSO or a Census bureau, do give you more specifications when you ask for it, but mostly this does not happen. And if they are being transparent, their specifications are difficult to validate. Interviewee 3 adds that for disaster data, it is even more difficult to create transparency: data is collected in the field, but by the time this data is verified, it is not relevant anymore as new data has already been collected or the situation has changed.
On the question to Interviewee 2 whether in Malawi it is possible to provide feedback to data providers he stated: "*No, this is not possible. It is already enough of a challenge to get them to share data in the first place, so we are happy with everything we get"* (Interviewee 2, personal communication, July 31, 2017).
He adds to this that it is important to be flexible in the process of collecting data, something that has already been stressed before. If you cannot find a specific indicator you are looking for, you should look for proxy-indicators that is accessible that can be used to approach this original indicator.
Interviewee 3 is a bit more nuanced in his answers, saying that it is difficult to provide feedback, but not impossible. He illustrated this claim by telling that he once experienced that a government agency suddenly stopped sharing data with UNOCHA, after a UNOCHA employee had told them in an unsubtle manner that there were some discrepancies in their data. He therefore stated that when providing feedback, it is important to remain sensitive: say it in a proper way, through a proper means, like a feedback forum that has been set up for this cause.
All in all, the interviewees agree that there generally is a lack of transparency, which results in not having insight in the quality of the data. However, this is something that they seem to have accepted. Their general tendency is perfectly expressed in the quote of Interviewee 2, on being happy with everything they get, and in his further explanation on the importance to remain flexible.

Accordingly, the ‘Trust and transparency’ and ‘Feedback’ element do not seem to fit the phase the humanitarian data ecosystem is currently in. It also seems that trust, transparency and the possibility to provide feedback are aspects that could be developed (or even automatically develop) after following the main finding from previous questions of the interviewees, namely to involve all relevant stakeholders from the start in the development of the model and to start the dialogue with them (which comes down to following the ‘Vision, communication and stimulation’ element). In hindsight, ‘Trust and transparency’ and ‘Feedback’ could have better been identified as elements of ‘secondary importance’, that become relevant once a data ecosystem is already further developed (see *Figure 16* in *Section 4.2.1*)

APPENDIX F – SCIENTIFIC ARTICLE

Developing a successful data ecosystem: a case study on the data-driven prioritization of humanitarian aid

Elise Haak
Delft University of Technology

ABSTRACT

The incidence of natural disasters worldwide is increasing. As a result, there is a growing number of people in need, whereas limited resources to help these people are available. It is therefore important to effectively and efficiently prioritize the most vulnerable people in the preparedness phase, and the most affected people in the response phase of humanitarian action. Data-driven models have the potential to help in doing so. However, to be able to apply these models in a country, a certain level of data preparedness is required. Therefore, there is a need to understand how to facilitate, stimulate and coordinate data-sharing between humanitarian actors, to achieve this level of data preparedness on a large scale. This paper aims to create this understanding based on a data ecosystem perspective, by providing insight in the success criteria for establishing a 'humanitarian data ecosystem'. As no scientific literature exists on this topic, as a first step, a general framework with data ecosystem governance success criteria is developed by means of a systematic literature review. Subsequently, the applicability of this framework in the humanitarian sector is assessed through a case study on the 'Community Risk Assessment and Prioritization toolbox developed by the Netherlands Red Cross. It is found that mainly the governance of the humanitarian data ecosystem requires attention, and that the data ecosystem approach provides a suitable framework for assessing the criteria to be addressed when aiming to establish a successful humanitarian data ecosystem.

Keywords

Humanitarian data, natural disasters, data-driven aid prioritization, data ecosystem, governance

INTRODUCTION

The humanitarian sector worldwide is under pressure: there is an increasing demand for humanitarian aid and emergency response services, as both the number and length of humanitarian crises grow (510 Global, 2016; Monaghan & Lycett, 2013; Whipkey & Verity, 2015). One of the causes of this increase is the rising number of disasters triggered by natural hazards, due to environmental degradation and climate change. This results in a fast-growing number of people in need (Hilhorst, 2013). Due to scarcity of budget and resources for disaster response, humanitarian actors cannot reach all these people. The prioritization of vulnerable and affected people is therefore at the core of all humanitarian interventions, and equally important in both the preparedness (pre-disaster) and the response phase (post-disaster) of humanitarian action (Van den Homberg, Visser & Van der Veen, 2017). Obtaining reliable and objective information for decision-making related to this prioritization process is considered very challenging (Van de Walle & Comes, 2014). Factors contributing to this are for example aid workers operating under stress and time pressure and being presented with uncertain information, which could lead to decision-making biases (Comes, 2017), and the complex organizational humanitarian setting in which information is scattered across many different organizations and sectors (Van den Homberg, Monné & Spruit,

2017; O'Brien 2017).

A trend that provides potential to deal with this challenge, and offers humanitarian actors ways to be become more effective and efficient, is the rapidly changing information environment, which leads to a fast increase in availability of data, and provides potential for a shift towards more evidence-based humanitarian decision-making. Paper-based processes get digitized increasingly and trends around crowdsourcing, social media, collaborative digital spaces and mobile services lead to a significant increase in availability of data. This also affects the humanitarian sector. Humanitarian donors for example, push the organizations they sponsor to open and share their data. In addition, major global agreements that recently came into effect, like for example the Paris Climate Agreement, the Sustainable Development Goals, and the Sendai framework for Disaster Risk Reduction, all stimulate countries to collect data on different indicators that are also potentially valuable for humanitarian organizations (Berens et al., 2016; Van den Homberg, Visser & Van der Veen, 2017). As Sabou (2016) describes, "humanitarian organizations now seek better ways to collect, store and use digital data to improve their collective responses to large scale crises".

As a result, a growing number of studies is dedicated to mathematical models and data-analytic techniques that can help in the humanitarian decision-making processes when trying to prepare for or respond to the consequences of a disaster (Ortuño et al., 2013). Data-driven priority indices have grown popular for effectively and efficiently identifying communities most vulnerable for or affected by disasters and numerous data and machine learning tools have been developed to aid prioritization (Benini, 2015).

However, humanitarian actors miss timely, reliable and sufficient granular data as well as often the skills and tools to analyze the data needed for a transparent and structured prioritization (Van den Homberg, Visser & Van der Veen, 2017). Data is collected, stored and used by an uncoordinated, diverse group of actors with varying degrees of expertise. This lack of alignment causes a range of inefficiencies (Berens et al., 2016). In addition, the collection, collation and analysis of data often only starts once a disaster hits, and to be able to apply these prioritization models, a certain level of 'data preparedness' is required (Van den Homberg, Visser & Van der Veen, 2017). Data preparedness is defined by Raymond & Al Achkar (2017) as "the ability of organizations to be ready to responsibly and effectively deploy data tools before a disaster strikes" (p.3).

Consequently, there is a need to understand how to facilitate, stimulate and coordinate data-sharing between humanitarian actors to increase the level of data preparedness in a country and hence be able to apply data-driven models that can help to come to a more effective and efficient prioritization of humanitarian aid. To create this understanding, this paper suggests taking a 'data ecosystem' approach, thereby referring to "the people and technologies collecting, handling and using the data and the interactions between them" (Parsons et al., 2011: p.557). By applying this approach to the humanitarian sector, in this paper we aim to provide an overview of criteria for the development of a successful humanitarian data ecosystem, combining a systematic literature review with case study insights.

This paper is structured as follows. First, the data ecosystem approach will be elaborated on and the research design will be described. Then, the outcomes of the systematic literature review will be presented, after which they are tested on a case study around the data-driven prioritization of humanitarian aid. Lastly, the findings are discussed and concluded.

A DATA ECOSYSTEM APPROACH

Academics in information intensive, socio-technical contexts have applied the ecosystems perspective to get an idea of the diverse interrelationships between data users, data providers, data itself, institutions and material infrastructure (Harrison, Pardo & Cook, 2012), as the perspective can help to deal with complexity (Heimstädt, Saunderson & Heath, 2014). According to Harrison, Pardo & Cook (2012), the approach can be used to outline existing conditions and develop desired conditions, stating that "its users often aim to provoke new thinking about the conditions and requirements necessary to actively cultivate development of an ecosystem to achieve a set of specific and desirable goals" (p.905). This is what is needed in the case of the humanitarian data ecosystem. Other potentially useful applications as described in literature are that data ecosystems can be used "as a means for decision-making and planning" (Zuiderwijk, Janssen & Davis, 2014: p.18), to locate the "relative positions of the actors in the ecosystem (data providers, sources, resources and users)" (Van Schalkwyk, Willmers & McNaughton, 2016: p.72), and to "facilitate access to sharing and using data" (Welle Donker & Van Loenen, 2017: p.286).

Berens et al. (2016) already propose the use of a data ecosystem perspective on the use of digital data in the humanitarian sector: "that of a complex data ecosystem comprised of a variety of actors that are touched by flows of digital data due to data sharing and data-related service delivery, and hence become part of a phenomenon that stretches beyond their organizational boundaries" (p.5). The perspective allows for a cross-organizational

understanding of data use, rather than only assessing internal data life cycles (Berens et al., 2016; Raymond et al., 2016). Raymond et al. (2016) describe the 'humanitarian data ecosystem' as the network of humanitarian organizations, affiliated and affected communities who are producing, collecting and analyzing digital data.

RESEARCH DESIGN

Scientific knowledge gap

No academic literature so far addresses the humanitarian data ecosystem, or the success criteria for the establishment of a humanitarian data ecosystem: Searches on Scopus, Web of Science and Google Scholar for 'humanitarian data ecosystem' or 'humanitarian' and 'data ecosystem' at the time of writing this paper, resulted in zero hits. The term does pop up in some grey literature, but in none of those documents the development of a successful humanitarian data ecosystem is addressed. Berens at al. (2016) and Raymond et al. (2016) for example, discuss the humanitarian data ecosystem from the angle of the responsible use of data and argue that the data ecosystem perspective can help to address the issue of a lack of centralized governance in the humanitarian sector. Sabou (2016) shares this view in a column, claiming that "the existing humanitarian - 'digital data' - ecosystem is essentially a collection of ungoverned pilot programs, and a new lens is needed to understand how we can meet the increasing demand for humanitarian aid in disaster onsets with relevant and ethical humanitarian innovation." This underscores the relevance of this study.

Systematic literature review and expert validation

We developed a framework for the success criteria for the development of a data ecosystem based on literature on data ecosystems in general. A systematic literature review has been carried out, in which the outcomes from publications addressing the establishment of data ecosystems were combined. A systematic literature review seeks to put together and synthesize all known knowledge on a topic area (Grant & Booth, 2009). This allowed us to come to a set of important aspects to consider when aiming to set up a successful data ecosystem.

Suitable literature was found by means of a thorough search among journal and conference articles, books, reports and other informative documents. Scopus, Web of Science and Google Scholar and Google Search were consulted as online databases. The following (combinations of) keywords were used to search these databases:

- "Data ecosystem"
- "Open data ecosystem"
- "Government ecosystem"
- "Criteria" + "data ecosystem"
- "Element" + "data ecosystem"
- "Successful" + "data ecosystem"
- "Establish" + "data ecosystem"
- "Data infrastructure"

The search of relevant articles was an iterative process. By inserting (combinations of) these keywords, a multitude of articles was found. The found articles were judged on relevance by reading the titles and summaries, and by scanning the text. When an initial selection was made, an additional search for literature took place based on the references used in these selected articles, which resulted in an increase of articles with potential. The new articles were scanned in the same manner. In total, 17 publications were found to be relevant in this systematic literature review. These publications are listed in *Table 1*.

The findings from the systematic literature review, which will be presented in the next section, have been validated with a group of 17 open data researchers from Delft University of Technology (Faculty of Technology, Policy and Management and Faculty of Architecture and the Built Environment), by means of an open discussion following a presentation of the outcomes.

	AUTHOR	FOCUS	OBJECTIVE	DOMAIN OF AUTHORS	METHODOLOGY OF STUDY
1.	Attard, Orlandi & Auer (2016)	Economic data ecosystem	To project their vision of generating a new Economic Data Ecosystem that has the Web of Data as its core.	Enterprise Information Systems	- Model design by means of a literature and practice review - Model evaluation by means of a survey
2.	Barthélemy (2016)	Open data ecosystem	To provide an overview of the Belgian open data ecosystem	School of Management: MSc in Business Engineering	Master thesis: - Literature review - Surveys/interviews - Case study
3.	Davies (2012)	Open data infrastructures and ecosystem	To highlight some of the interventions that may be necessary to support realization of impact from open data initiatives.	Open Data & Open Government	Argumentative text based on domain description
4.	Dawes, Vidiasova & Parkhimovich (2016)	Open government data ecosystem	To develop a preliminary ecosystem model for planning and designing Open Government Data programs	E-governance / Open Government	Literature review combined with findings from two empirical case studies
5.	Harrison, Pardo & Cook (2012)	Open government ecosystem	To create a research and development agenda with questions essential to the development of Open Government Ecosystems.	IT in Government / Open Government	- Literature review serves as framework - To consider the outcomes of a workshop with a multi-sector expert group
6.	Heimstädt, Saunderson & Heath (2014)	Open data ecosystem	To identify a set of structural business ecosystem properties.	Open (Government) Data	- Review of popular Open Data definitions and business ecosystem theories applied to empirical data using a timeline analysis - Work informed by combination of discourse analysis and in-depth interviews
7.	Immonen, Palviainen & Ovaska (2014)	Open data based business ecosystem	To define the requirements of an open data based business ecosystem (an open data ecosystem from the business viewpoint)	Service / Software Engineering	- State-of-the-art knowledge explored from the literature - State of the practice on data-based business in the industry collected through interviews
8.	Jetzek (2017)	Open data ecosystem	To explore the possibilities for sustainable value generation in the Open Data Ecosystem	IT Management: Big Data & Open Data	Theoretical review combined with case study (interviews and participation in an open data initiative in Denmark)
9.	Lee (2014)	Open data ecosystem	To specify a series of specific elements critical for building an Open Data Ecosystem	Linked & Open Data	Literature review combined with liaising with stakeholders
10.	Macharis & Crompvoets (2014)	Spatial data infrastructure	To evaluate development scenarios for the spatial data infrastructure for Flanders	Supply Chain Management / Public Governance / Spatial Data Infrastructures	Multi Actor Multi Criteria Analysis (MAMCA) applied to a case study
11.	Parsons et al. (2011)	Science data ecosystem	To success several short- and long-term strategies to facilitate a socio-technical evolution in the overall science data ecosystem	Science Data	Case study
12.	Pollock (2011)	Open data ecosystem	To stress the importance of data cycles with feedback loops when building open data ecosystems	Open Knowledge	Blogpost
13.	Ponte (2015)	Open data ecosystem	To provide an overview of the issues to be addressed when enabling an open data ecosystem	Organization Science & ICT	Survey of the open data market
14.	Van Schalkwyk, Willmers & McNaughton (2016)	Open data ecosystem	To consider the supply, demand and use of open data, as well as the roles of intermediaries, using an ecosystem approach	Open data / Open ICT Ecosystems & Scholarly Communication	Case study
15.	Welle Donker & Van Loenen (2017)	Open data ecosystem	To develop an open data assessment framework based on three output indicators as conditions for a successful open data ecosystem	Open Data / Information Infrastructures	- Framework analysis (domain description) - Framework development - Case study
16.	Wiener et al. (2016)	Open data ecosystem	To enable an open data ecosystem for the neurosciences	Neuroscience & Data-sharing	Workshop with actors involved
17.	Zuiderwijk, Janssen & Davis (2014)	Open data ecosystem	To provide an overview of essential elements of open data ecosystems for enabling easy publication and use of open data	Open Data / ICT & Governance	-Literature review -Scenario analysis

Table 1: List of articles reviewed (objectives and methodologies have been retrieved from the articles)

Case study testing

As can be seen in *Table 1,* 14 of the 16 articles are centered around open data ecosystems and many of the authors have a background in open (government) data. As it is unclear whether the criteria that hold for the successful establishment of open data ecosystems specifically can also be translated one-on-one to the humanitarian data ecosystem, in which a difficulty is that much of the data is not shared openly (and sometimes not even stored digitally), the framework of success criteria for the development of a data ecosystem that followed from literature was also empirically validated. It was analyzed on its application in the humanitarian sector by testing it on a case study: the 'Community Risk Assessment and Prioritization toolbox', developed by the data team of the Netherlands Red Cross. This is an alternative data-driven solution that preemptively gathers and combines relevant data on risk components (like population density, poverty, building material), to provide a detailed risk assessment for areas and communities in a country, that helps to faster identify priority areas for humanitarian intervention related to natural disasters. The output is an easy-to-understand dashboard in which colors on a map visualize the risk or predicted damage for a specific area. Currently, this model is under development for the Philippines, Malawi and Nepal (510 Global, 2017). To be able to develop this model for a country, a humanitarian data ecosystem needs to be created in which data from different administrative levels can be collected and users are stimulated to make use of the model.

In the first place, the outcomes of the systematic literature review were presented to a group of seven people involved in the development of the Community Risk Assessment and Prioritization toolbox, to assess whether additional criteria could be identified that were relevant specifically for the humanitarian sector.

Secondly, to check for the practical relevance of the outcomes, interviews with eight humanitarian data experts were held. These experts represented five different regions (Greater Horn of Africa[8], Latin America and Caribbean, Malawi, the Philippines and the Southern Africa Development Community[9]), to assess a variety of data landscapes and to be able to distinguish between country-specific and generic answers. They all had knowledge on the broader data landscape in their country, and experience with collecting data similar (or identic) to the input data for the Community Risk Assessment and Prioritization toolbox on low administrative levels, which is necessary for the toolbox to generate accurate outcomes.

FINDINGS

Theoretical findings: general data ecosystem success criteria

The systematic literature review showed that there is a diverse set of aspects that needs to be considered when establishing a successful data ecosystem. Three different categories can be distinguished: 1) Data supply, relating to the provision of data as open data, 2) Governance, being the framework of policies, processes and instruments to realize common goals in the interaction between entities (and facilitating the data supply), and 3) User characteristics (Welle Donker & Van Loenen, 2017).

As described, these findings were evaluated in a validation session with a group of open data researchers from the Delft University of Technology. During this session, the attendants were in the first place asked about the completeness of the literature list (*Table 1*), and in the second place asked about the completeness and correctness of the set of data ecosystem governance criteria. The combination of the systematic literature review and this validation session led to the framework visualized in *Figure 1*. An overview of the articles in which each criterion appeared, can be found in *Table 3* in the Appendix.

[8] The Greater Horn of Africa comprises eight countries: Djibouti, Eritrea, Ethiopia, Kenya, Somalia, South Sudan, Sudan and Uganda.
[9] The Southern Africa Development Community (SADC) has 16 member states: Angola, Botswana, Comoros, Democratic Republic of the Congo, Lesotho, Madagascar, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia and Zimbabwe.

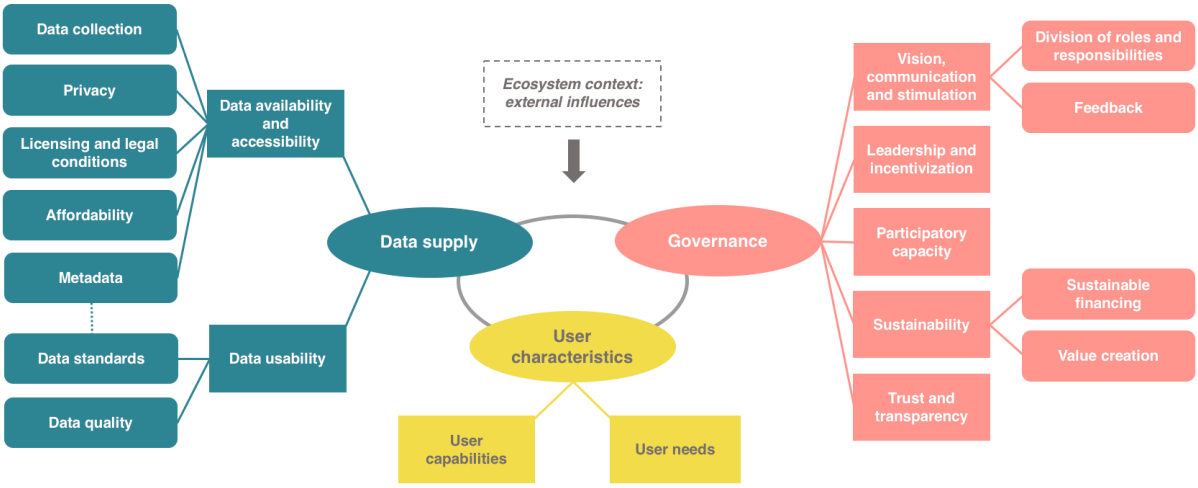


Figure 1: Final framework of criteria for a successful data ecosystem – findings from systematic literature review and expert validation combined

The Data Supply ('DS') criteria are described as follows:

DS1 **Data availability and accessibility:** It should be clear for actors in a data ecosystem how to find data, where to find it and how to access it.

DS1a **Data collection:** It should be clear to actors in the data ecosystem how data can be collected.

DS1b **Privacy:** It should be ensured that data protection laws are followed.

DS1c **Licensing and legal conditions:** Another important aspect to enhance the accessibility of data in a data ecosystem is the type of licensing associated with the data, which is necessary to ensure the legal foundation for the potential (re)use of data

DS1d **Affordability:** The data shared in the data ecosystem should be affordable.

DS1e **Metadata:** The existence of appropriate metadata can help to improve the availability and accessibility of data.

DS2a **Data standards:** An important aspect that determines the quality of data is the presence of standards to facilitate data interoperability.

DS2b **Data quality:** The data shared in the data ecosystem should be of high quality

The Governance ('G') criteria are described as follows:

G1 **Vision, communication and stimulation:** A collaborative, interactive environment should be established and cooperation between stakeholders should be stimulated. A fragmented approach should be avoided.

G1a **Division of roles and responsibilities:** All actors in the data ecosystem should be identified, and there should be a clear division of the roles and responsibilities of the actors in a data ecosystem.

G1b **Feedback:** The data ecosystem should include feedback mechanisms to enable data users to provide feedback to data providers.

G2 **Leadership and incentivization:** Activities in a data ecosystem should be stimulated, incentivized and coordinated by a problem owner, either top-down or bottom-up. Incentivization can take place by for example make data sharers feel intrinsically rewarded, or by lowering the barrier to entry for data sharing.

G3 **Participatory capacity:** Data supply should match data demand in a data ecosystem: public bodies require certain capacities to be able to participate in a data ecosystem, like:

- Technical knowledge on certain systems and technologies involved
- Data management knowledge, on how to ensure high data quality
- Operational knowledge, on how to incorporate data activities into current practices

G4 **Sustainability - financing and value creation:** For a data ecosystem to become sustainable,
G4a sustainable financing should be arranged and value should be generated for the ecosystem
G4b stakeholders.

G5 **Trust and transparency:** Data providers and data users should mutually trust and be transparent to each other.

The following criteria relate to the third category: User Characteristics (UC).

UC1 **User capabilities:** The capabilities of the data users in an ecosystem should be considered.

UC2 **User needs:** The needs of the data users in an ecosystem should be considered.

One last aspect related to data ecosystems coming forward in the literature is the importance of considering the external context of an ecosystem. This is not exactly a criterion for a successful data ecosystem; it is merely an overarching factor that should be considered when analyzing a data ecosystem. The functioning of a data ecosystem also depends on for example the local culture, the political system and historical influences. These aspects all together form the institutional conditions in which the ecosystem is or needs to be embedded, that influence how actors in the ecosystem function and how different ecosystem elements are arranged. Hence:

EC **External context:** The external context of the data ecosystem should be considered.

Practical findings: humanitarian data ecosystem success criteria

Subsequently, the application of the framework derived from literature to the humanitarian sector was assessed. By discussing the framework in a validation session with a group of seven people involved in the development of the Community Risk Assessment and Prioritization toolbox, it was found that the 'data supply' and 'user characteristics' criteria both relate to the design and development of the toolbox, whereas the 'governance' criteria relate to the creation of an optimal context for the development of the toolbox. Therefore, having practical experience with the development of the toolbox, the attendants of this session were all able to evaluate the relevance of the 'data supply' and 'user characteristics' criteria. All these criteria were considered relevant in the context of the humanitarian data ecosystem, though it might be possible that some of the elements will receive a stronger focus in a later stage of the data ecosystem: it was repeatedly stressed that the humanitarian data ecosystem is in a very early phase of its development.

In the current phase of the humanitarian data ecosystem there is a strong focus on the data supply criteria; mainly the ones directly related to data collection, as that is essential to be able to start developing the Community Risk Assessment and Prioritization toolbox in the first place, and thus where the data ecosystem starts to come into existence. Data standards for example, are indeed 'nice to have' and make the data collection and analysis easier, but it is also possible to get the data ecosystem running without data standards in place.

Also, in this first phase the user characteristics need to be considered. If the outcomes of the toolbox do not match the user capabilities or information needs of humanitarian decision-makers, the desired effect will not be achieved.

The governance part of the framework is very much underdeveloped. This is a large issue with the humanitarian data ecosystem around the Community Risk Assessment and Prioritization toolbox: There are no, or very few, governance structures in place yet that can stimulate the data ecosystem development. There are many fragmented organizations involved that only pursue their own objectives, while there is a lack of central governance to coordinate and align all these individual activities. This is in line what has previously in this paper been described as claims by Berens et al. (2016), Raymond et al. (2016) and Sabou (2016). As Berens et al. (2016) described, "in humanitarian work, the multitude of independent organizations, government actors, multinational initiatives, individuals, private sector recourse providers, and digital platforms constitute a dynamic ecosystem with no clear leader or dominant force" (p.1). They later refer to this as the existing 'governance disparity'.

To get to a structure to successfully govern the process of rolling out the Community Risk Assessment and Prioritization toolbox in a country, interviews with eight humanitarian data experts representing five different regions, and one having global experience, were carried out. It is assumed that a humanitarian data ecosystem spans one country, and that the governance is taken care of within this country.

The empirical findings resulting from these interviews are presented in *Table 1* below. Green indicates that the criterion was perceived relevant, orange indicates relevant, but only in a later, more developed stage of the humanitarian data ecosystem, and red means 'should be omitted'.

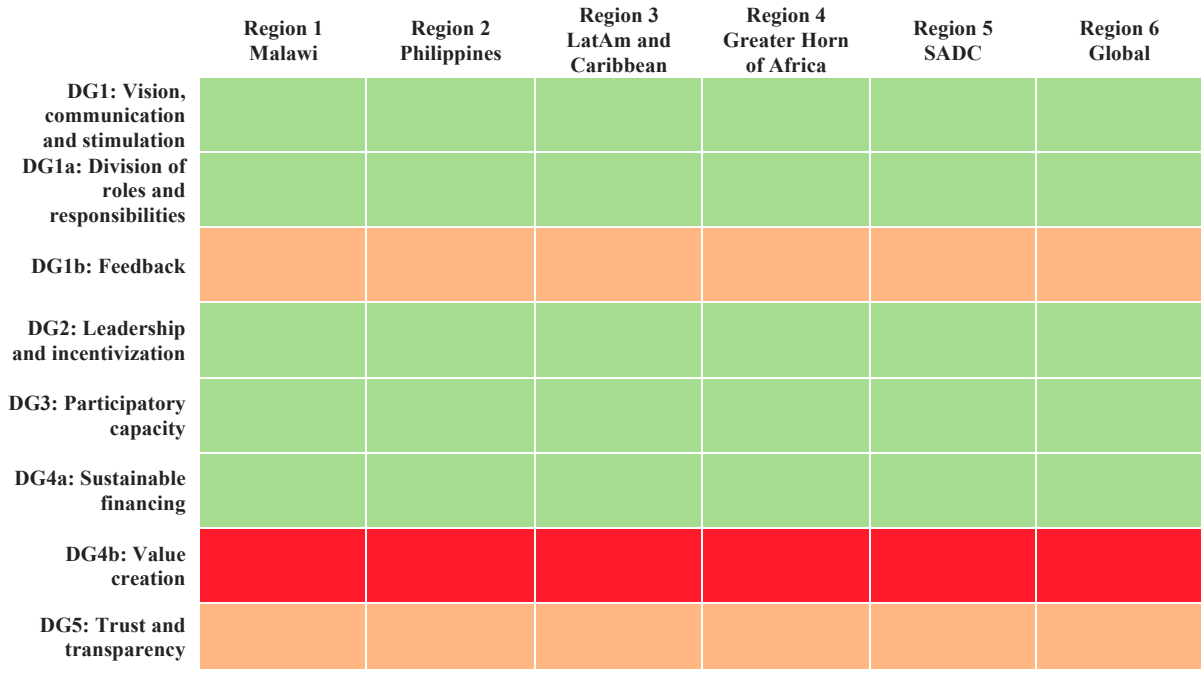


Table 3: Perceived relevance of governance criteria in the current humanitarian data ecosystem

As can be seen, the interviewees were very homogeneous in their answers. No additional criteria were suggested to consider.

The main practical implications of the theoretical governance criteria in the current humanitarian data ecosystem, as described by the interviewees, are summarized below.

1 **Vision, communication and stimulation:** To create a sense of responsibility and ownership, to have all ecosystem stakeholders understand the value and importance of data-sharing and to address the issue of fragmentation, it is important that the party in the leadership role actively involves all other relevant stakeholders in the establishment of the data ecosystem.

1a **Division of roles and responsibilities:** In the specific context of the Community Risk Assessment and Prioritization toolbox, there are four different roles to be identified and assigned, of which the first two have a leading role (and hence also relate to DG2 – Leadership and incentivization.

- *The initiator/coordinator,* which is the leading agency, most likely a United Nations agency, in setting up the humanitarian data ecosystem. In the toolbox context, this party should initiate, coordinate and support the development of the toolbox in a country, bring all relevant actors together and incentivize them to facilitate data-sharing and to use the toolbox.
- *The local lead*, which is a national body that becomes the local manager after the initiator/coordinator has set out the initial data ecosystem, and is responsible for locally promoting participation. This role is important to ensure the sustainability of the ecosystem.
- *Data providers* should be identified and are responsible for sharing their data that is considered relevant.
- *Toolbox users*, which are humanitarian or government agencies who should use the toolbox (i.e. the data processed into understandable information) and incorporate the outputs in their operations and decision-making.

1b **Feedback:** The possibility to provide feedback to data providers is considered useful, but irrelevant in the current phase of the humanitarian data ecosystem: collecting relevant data is already sufficiently challenging. The interviewees rather look for alternatives than putting effort in the development of feedback mechanisms, hence the color orange: relevant, but in

a later, more developed stage of the humanitarian data ecosystem. The same holds for transparency from data providers on how they got their data – 'transparency' is considered to be overlapping with 'feedback'.

2 **Leadership and incentivization:** It is suggested to split up this element, as both aspects can be assessed individually. 'Leadership' was already addressed with the 'division of roles and responsibilities' (indicating an overlap) and 'incentivization is considered very important, and something that relates to all identified roles in the data ecosystem. One of the most suggested approaches to incentivize parties to participate in the data ecosystem was by meeting the 'vision, communication and stimulation' criterion: to involve stakeholders from the start.

3 **Participatory capacity:** For 'participatory capacity' holds the same as for 'incentivization': very important, and something that needs to be addressed and built for every role in the ecosystem. Essential in the humanitarian data ecosystem, as stakeholders generally lack technical and information management capacities, especially at lower administrative levels.

4 **Sustainability: financing and value creation:** Sustainable financing should be the responsibility of the initiator/coordinator of the data ecosystem. 'Value creation' is considered to be overlapping with incentivization: parties can be incentivized to participate by creating value for them – hence red: should be omitted.

5 **Trust and transparency:** As described, transparency should be merged with 'feedback' as they relate to each other. Trust between actors in the data ecosystem is indeed considered important. However, it is not a criterion for a successful humanitarian data ecosystem, but rather something that follows from the creation of the 'vision, communication and stimulation' criterion.

To summarize, the following suggestions for changes in the original framework – when applying it to the humanitarian data ecosystem – were proposed:

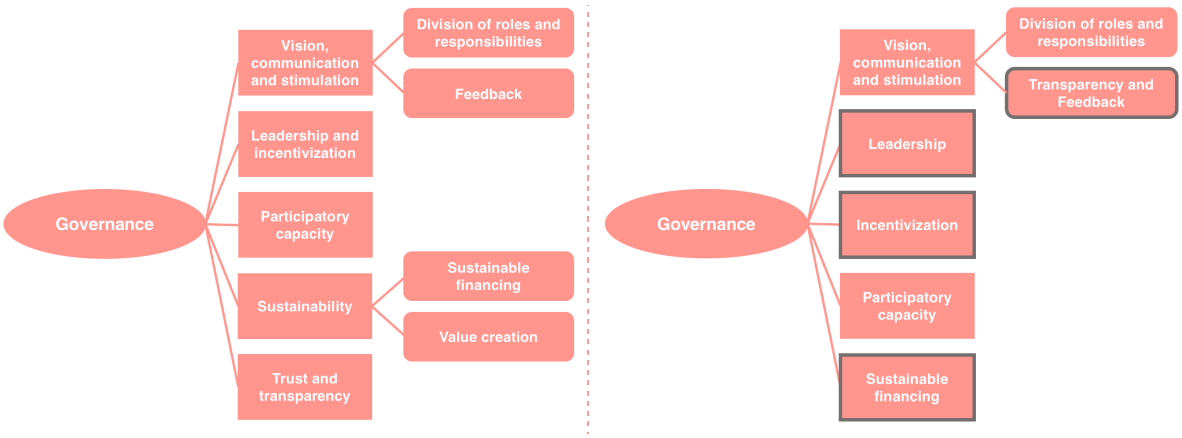


Figure 2: Left: governance criteria before empirical validation – Right: governance criteria after empirical validation

DISCUSSION AND CONCLUSION

This study has taken a 'data ecosystem' approach to create understanding about how to facilitate, stimulate and coordinate data-sharing between humanitarian actors to increase the level of data preparedness in a country. By doing so, this study is the first scientific study to address the 'humanitarian data ecosystem'. A framework of criteria for the successful development of a data ecosystem (in general) has been developed based on a systematic literature review, and was validated with a panel of open data experts. Subsequently, this framework has been tested on its application and relevance in the humanitarian sector by means of a case study: should these same criteria be considered when establishing a data ecosystem in the complex, fragmented humanitarian context around the Community Risk Assessment and Prioritization toolbox developed by the Netherlands Red Cross? The point of attention in the humanitarian data ecosystem appeared to be the governance part of the framework, as this is still very much underdeveloped. Therefore, eight respondents representing five different regions with varying data landscapes were asked about how they perceived the relevance and application of the different governance criteria in their region. It was hereby assumed that a humanitarian data ecosystem spans one country, and that the

governance is taken care of within this country. The interviewees also specified the practical implications of the criteria for the humanitarian data ecosystem specifically.

'Value creation' and 'trust' were found to be overlapping and hence omitted, but all the other criteria were considered very relevant. A remark must be made for 'transparency and 'feedback' though; according to the interviewees these only become relevant in a later data ecosystem stadium, due to the current focus on enhancing, facilitating and speeding-up data-sharing practices.

By means of the interviews, it was also found that many of the governance criteria closely interrelate with 'data supply' and 'user characteristics' criteria: they are all interdependent. The division of roles and responsibilities for example, should also include the identification of data providers – relating to the data supply part of the framework, and toolbox users – relating to the user characteristics part of the framework. Moreover, the 'leadership' is also a part of the 'division of roles and responsibilities', and likewise 'incentivization' and 'participatory capacity' hold for every identified role.

Thus, it is argued that it is relevant to approach data-sharing in the humanitarian sector with the data ecosystem perspective, as it provides insight in the elements to be addressed and helps to provide structure. However, the interrelations between the criteria are a lot more complex than visualized in *Figure 1*, indicating the humanitarian data ecosystem might be too complex to capture in a static framework. Additionally, some of the criteria are considered to be of secondary importance: they only become relevant in a late stage of the data ecosystem. The humanitarian data ecosystem is emerging and currently in its initial development phase. There is a strong focus on 'data supply' and on – though less but still present – 'user characteristics', but the governance part requires attention to create an optimal context to facilitate meeting the data supply and user characteristics criteria.

Future research should extend this study by also testing the theoretical framework of data ecosystem success criteria to a set of other case studies, as an extra validation. This would both provide insight in the generic applicability of the framework, and in the specificity of the findings for the humanitarian data ecosystem. Other case studies can also provide insight in the evolvement of the maturity of other data ecosystems, which is something the immature humanitarian data ecosystem can learn from.

REFERENCES

510 Global. (2016). *Personalised Humanitarian Aid Through Big Data: an initiative of the Dutch Red Cross.* Retrieved from http://510.global (accessed March 21, 2017).

510 Global. (2017). *Community Risk Assessment and Prioritization toolbox.* Retrieved from https://profiles.510.global/#!/ (accessed March 15, 2017).

Benini, A. (2015). *The use of Data Envelopment Analysis to calculate priority scores in needs assessments.* A note for ACAPS.

Berens, J., Raymond, N., Shimshon, G., Verhulst, S., and Bernholz, L. (2016). The Humanitarian Data Ecosystem: the Case for Collective Responsibility. Retrieved from https://pacscenter.stanford.edu/wp-content/uploads/2017/08/humanitarian_data_ecosystem.pdf (accessed September 19, 2017).

Comes, T. (2016). *Cognitive biases in humanitarian sensemaking and decision-making lessons from field research.* Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), IEEE International Multi-Disciplinary Conference on. IEEE, 2016.

Grant, M.J. and Booth, A. (2009). A typology of reviews: an analysis of 14 reviews types and associated methodologies. *Health Information and Libraries Journal, 26*, pp. 91-108.

Harrison, T.M., Pardo, T.A. and Cook, M. (2012). Creating Open Government Ecosystems: A Research and Development Agenda. *Future Internet 2012, 4*, pp. 900-928.

Heimstädt, M., Saunderson, F. and Heath, T. (2014). *Conceptualizing Open Data ecosystems: A timeline analysis of Open Data development in the UK.* Discussion paper, Freie Universität Berlin, School of Business & Economics: Management, No. 2014/12.

Hilhorst, D. (2013). Disaster, conflict and society in crises: everyday politics of crisis response: Routledge.

Macharis, C. and Crompvoets, J. (2014). A stakeholder-based assessment framework applied to evaluate development scenarios for the spatial data infrastructure for Flanders. *Computers, Environment and Urban Systems, 46*, pp. 45-56.

Monaghan, A and Lycett, M. (2013). *Big Data and Humanitarian Supply Networks: Can Big Data Give Voice to the Voiceless?* Paper presented at the 2013 Global Humanitarian Technology Conference (GHTC), IEEE.

O'Brien, S. (2017). *This is how we build a stronger, data-driven humanitarian sector.* Part of the World Economic Forum Annual Meeting 2017, retrieved from https://www.weforum.org/agenda/2017/01/this-is-how-we-build-a-stronger-data-driver-humanitarian%20sector (accessed July 3, 2017).

Ortuño, M.T. et al. (2013). Decision Aid Models and Systems for Humanitarian Logistics. A Survey. *Decision Aid Models for Disaster Management and Emergencies,* Volume 7 of the series *Atlantis Computational Intelligence Systems,* pp. 17-44.

Parsons, M.A., Godøy, Ø., LeDrew, E., De Bruin, T.F., Danis, B., Tomlinson, S. & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science, 37*(6), pp. 555-569.

Raymond, N. and Al Achkar, Z. (2016). *Data preparedness: connecting data, decision-making and humanitarian response.* Harvard Humanitarian Initiative: signal program on human security and technology - standards and ethics series 01. Retrieved from http://hhi.harvard.edu/sites/default/files/publications/data_preparedness_update.pdf (accessed July 4, 2017).

Raymond, N., Al Achkar, Z., Verhulst, S. And Berens, J. (2016). *Building data responsibility into humanitarian action.* OCHA Policy and Studies Series 018. Retrieved from http://www.unocha.org/sites/unocha/files/Building%20data%20responsibility%20into%20humanitarian%20action.pdf (accessed July 4, 2017).

Sabou, J. (2016). *The Humanitarian "Data" Ecosystem: Intersection of Innovation & Disaster Relief.* Retrieved from http://www.leidensafetyandsecurityblog.nl/articles/the-humanitarian-data-ecosystem-intersection-of-innovation-disaster-relief (accessed July 3, 2017).

Van den Homberg, M., Monné, R. and Spruit, M. (2017 – to be published). Bridging the information gap by optimizing data selection using cost and quality to cover information needs of disaster responders. CAGEO.

Van den Homberg, M., Visser, J. and Van der Veen, M. (2017). *Unpacking Data Preparedness from a humanitarian decision-making perspective: toward an assessment framework at subnational level.* Proceedings of the 14th ISCRAM conference, Albi, France.

Van de Walle, B. and Comes, T. (2014). Risk Accelerators in Disasters: Insights from the Typhoon Haiyan Response on Humanitarian Information Management and Decision Support. In M. Jarke et al. (Eds), *Advanced Information Systems Engineering, Vol. 8484* (pp. 12-23). Springer International Publishing.

Welle Donker, F. and Van Loenen, B. (2017). How to assess the success of the open data ecosystem? *International Journal of Digital Earth, 10*(3), pp. 284-306.

Whipkey, K. and Verity, A. (2015). *Guidance for incorporating Big Data into Humanitarian Operations.* Document created as part of the Digital Humanitarian Network's community of interest on Decision Makers Needs. Retrieved from http://digitalhumanitarians.com/sites/default/files/resource-field_media/IncorporatingBigDataintoHumanitarianOps-2015.pdf (accessed April 5, 2017).

APPENDIX

CATEGORY	CRITERION	DESCRIPTION	SOURCE
1. DATA SUPPLY	Data availability and accessibility	It should be clear for actors in a data ecosystem how to find data, where to find it and how to access it	• Attard, Orlandi & Auer (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016) • Zuiderwijk, Janssen & Davis (2014)
	Data collection	It should be clear to actors in the data ecosystem how data can be collected.	*Added after expert validation*
	Privacy	It should be ensured that data protection laws are followed	• Dawes, Vidiasova & Parkhimovich (2016) • Immonen, Palviainen & Ovaska (2014), • Lee (2014) • Wiener et al. (2016)
	Licensing and legal conditions	Another important aspect to enhance the accessibility of data in a data ecosystem is the type of licensing associated with the data, which is necessary to ensure the legal grounding for the potential (re)use of data	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Zuiderwijk, Janssen & Davis (2014)
	Affordability	The data shared in the data ecosystem should be affordable.	*Added after expert validation.*
	Metadata	The existence of appropriate metadata can help to improve the availability and accessibility of the data	• Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016) • Zuiderwijk, Janssen & Davis (2014)
	Data usability: *Data quality*	High quality of the data shared in a data ecosystem should be ensured to enhance its usability	• Attard, Orlandi & Auer (2016) • Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Zuiderwijk, Janssen & Davis (2014)
	Data standards	An important aspect that determines the quality of data is the presence of standards to facilitate data interoperability	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Macharis & Crompvoets (2012) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)

CATEGORY	CRITERION	DESCRIPTION	SOURCE
2. DATA GOVERNANCE	Vision, communication and stimulation	A collaborative, interactive environment should be established and cooperation between stakeholders should be stimulated	• Attard, Orlandi & Auer (2016) • Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Immonen, Palviainen & Ovaska (2014) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016) *(And other scholars implicitly: a 'collaborative environment' is a characteristic inherent to ecosystems)*
	Division of roles and responsibilities	There should be a clear division of the roles and responsibilities of the actors in an ecosystem	• Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Van Schalkwyk, Willmers & McNaughton (2016) • Zuiderwijk, Janssen & Davis (2014)
	Feedback	The data ecosystem should include feedback mechanisms to enable data users to provide feedback to data providers	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Zuiderwijk, Janssen & Davis (2014)
	Leadership and incentivization	Activities in an ecosystem should be stimulated, incentivized and coordinated (either top-down or bottom-up, depending on the context)	• Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
	Participatory capacity	Data supply should match data demand in an ecosystem; public bodies require certain capacities to be able to participate in an ecosystem	• Attard, Orlandi & Auer (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Lee (2014) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
	Sustainability: financing and value creation	For a data ecosystem to become sustainable, sustainable financing should be arranged and the value should be generated for the ecosystem stakeholders	• Davies (2012) • Harrison, Pardo & Cook (2012) • Heimstädt, Saunderson & Heath (2014) • Jetzek (2017) • Lee (2014) • Parsons et al. (2011) • Ponte (2015) • Van Schalkwyk, Willmers & McNaughton (2016) • Welle Donker & Van Loenen (2017) • Wiener et al. (2016)
	Trust and transparency	Data providers and data users should mutually trust and be transparent to each other.	*Added after expert validation.*
3. USER CHARACTERISTICS	User capabilities	The capabilities of the data users in an ecosystem should be considered	• Barthélemy (2016) • Davies (2012) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Lee (2014) • Welle Donker & Van Loenen (2017) • Zuiderwijk, Janssen & Davis (2014)
	User needs	The needs of the data users in an ecosystem should be considered	• Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012) • Immonen, Palviainen & Ovaska (2014) • Lee (2014) • Welle Donker & Van Loenen (2017)
A. EXTERNAL CONTEXT	-	The external context of the data ecosystem should be considered	• Barthélemy (2016) • Dawes, Vidiasova & Parkhimovich (2016) • Harrison, Pardo & Cook (2012)

CATEGORY	CRITERION	DESCRIPTION	SOURCE
			• Macharis & Crompvoets (2012) • Parsons et al. (2011) • Van Schalkwyk, Willmers & McNaughton (2016) • Zuiderwijk, Janssen & Davis (2014)

Table 3: Criteria for a successful data ecosystem found in the systematic literature review, complemented with insights from expert validation