MSc thesis

Determining relevant disparate disaster data and selecting an integration method to create actionable information;

For the professional and responding community in the disaster response and preparedness phase

Robert Monné
Rmonne@xs4all.nl
+31 6 13946811
Utrecht University
Master of Business Informatics
Sunday, 10 January 2016

SUPERVISORS **4**
LIST OF TABLES **5**
LIST OF FIGURES **6**
TABLE OF ABBREVIATIONS **7**
PREFACE **8**
EXECUTIVE SUMMARY **9**
1 INTRODUCTION **11**
1.1 CONTEXT 11
1.2 PROBLEM IDENTIFICATION & DIAGNOSIS 15
1.3 SCOPING 19
1.4 RESEARCH QUESTIONS 23
1.5 SUMMARY OF CHAPTER 23
2 RESEARCH DESIGN **25**
2.1 RESEARCH MODEL 25
2.2 OVERVIEW OF RESEARCH METHODS 27
2.3 RESEARCH METHOD: HOW DO WE DETERMINE AND INTEGRATE THE RELEVANT DISASTER DATA? 28
2.4 RESEARCH METHOD: 'INFORMATION NEED' IDENTIFICATION (RQ 1) 29
2.5 RESEARCH METHOD: IDENTIFYING DISASTER DATA SOURCES (RQ 2) 35
2.6 RESEARCH METHOD: IDENTIFYING DATA INTEGRATION METHODS (RQ 3) 36
2.7 SUMMARY 38
3 THEORETICAL FRAMEWORK **39**
3.1 EVALUATION FRAMEWORK 39
3.2 INTEGRATION METHODS 41
3.3 CURRENT STATE OF THE ART IN DATA INTEGRATION FOR DISASTER RESPONSE/MANAGEMENT 45
3.4 SUMMARY 46
4 RESULTS **47**
4.1 DATA INTEGRATION METHOD SELECTION APPROACH 47
4.2 THE 'INFORMATION NEED' OF DISASTER RESPONDERS (RQ 1) 51
4.3 IDENTIFIED DATA SOURCES ON FLOODS IN BANGLADESH (RQ 2) 59
4.4 DATA INTEGRATION METHOD EVALUATION (RQ3) 62
4.5 SUMMARY 62
5 RESEARCHER'S OBSERVATIONS FROM FIELD VISIT **62**
5.1 INFORMATION USAGE AT THE GRASS-ROOT LEVEL 63
5.2 DATA QUALITY AND AUTHENTICITY 64
5.3 GOVERNMENT VS NGOS 64
5.4 BENEFICIARY SELECTION 64
5.5 DUPLICATED DATA COLLECTION 64
5.6 DATA GRANULARITY LOSS IN GOVERNMENT 65
5.7 SUMMARY 66
6 DETERMINING 'TO-BE INTEGRATED' DISASTER DATA SOURCES AND SELECTING AN INTEGRATION METHOD (MAIN RESEARCH QUESTION) **67**
6.1 DETERMINING 'TO-BE INTEGRATED' DISASTER DATA SOURCES 67
6.2 SELECTING AN INTEGRATION METHOD 85
7 DISCUSSION **94**
7.1 OUR APPROACH WORKS! 94
7.2 INFORMATION NEEDS 94
7.3 DATA SOURCES 95

7.4 INTEGRATION METHODS 95
7.5 FIELD TRIP CHALLENGES 96
8 CONCLUSIONS 97
9 RECOMMENDATIONS 99
10 FUTURE RESEARCH 100
11 ACKNOWLEDGEMENTS 101
12 REFERENCES 102
13 APPENDIXES 107
13.1 LIST OF APPENDIXES 107

Supervisors

Name and title	Dr. Marc van den Homberg
Affiliation	Senior expertise consultant ICT4D - TNO
Address	Oude Waalsdorperweg 63 2597 AK Den Haag
Telephone number	+33 6 58840547
Email address	marc.vandenhomberg@tno.nl; marcjchr@gmail.com

Name and title	Dr. Marco Spruit
Faculty, department and Research group (chair)	Department of Information and Computing Sciences, Software Systems, Organizations and Information
Email address	M.R.Spruit@uu.nl

Name and title	Dr. Sergio España (2nd supervisor)
Faculty, department and Research group (chair)	Department of Information and Computing Sciences, Software Systems, Organizations and Information
Email address	s.espana@uu.nl

List of tables

Table 1 Interview target group 31
Table 2 Distribution of Interviews (Respondents vs Transcriptions) 32
Table 3 Search terms for literature review 37
Table 4 Evaluation framework related to data 39
Table 5 integration method evaluation framework related to developing world 40
Table 6 Identified Data Integration Methods 41
Table 7 Activity descriptions for proposed method 49
Table 8 High level clusters of Disaster Responder Activities 52
Table 9 Disaster responder activities + preparedness + recovery 52
Table 10 Disaster responder decisions 54
Table 11 High level Responder Information Needs 55
Table 12 Low level Disaster responder Information Needs 55
Table 13 Final list of information needs 56
Table 14 Quotes of respondents 58
Table 15 Overview of disaster data sources, types, sponsors and sources 60
Table 16 Overview of data integration methods 62
Table 17 Information Need vs Data Source Matrix 68
Table 18 Data source's coverage of information needs 71
Table 19 Information need fulfilment statistics 72
Table 20 Information needs not covered by data sources (given time constraints) 78
Table 21 Information needs not covered in the data sources (without timing constraints) 79
Table 22 Sorting coverage approach 80
Table 23 Deciding on "to-be-integrated" data (Most coverage sorting approach) with timing constraints 81
Table 24 Sorting approach without timing constraints 82
Table 25 Deciding on "to-be-integrated" data (Most coverage sorting approach) without timing constraints 83
Table 26 Summarization of results from Optimization approach 85
Table 27 to-be-integrated data combinations to analyse 86
Table 28 Analysis of dissimilarity between JNA excel and FFWC 86
Table 29 Analysis of dissimilarity between JNA PDF and FFWC 87
Table 30 Analysis of dissimilarity between JNA excel and District Disaster Management Plan 87
Table 31 Analysis of dissimilarity between JNA PDF and District Disaster Management Plan 88
Table 32 Analysis of dissimilarity between JNA excel and PDF 88
Table 33 Analysis of dissimilarity between District Disaster Management Plan and FFWC 89
Table 34 overview of analysis between data source dissimilarity 90

List of figures

Figure 1 Proposed extension to the CRISP-DM process 9
Figure 2 Geographical location of research site 12
Figure 3 Communities and their overlap in disaster response 13
Figure 4 phases of a disaster management process 14
Figure 5 Visualisation of challenges 15
Figure 6 Information requirements in disasters (Gralla, Goentzel, & Van de Walle, 2013) 17
Figure 7 Mismatch between information needs and data sources 18
Figure 8 From data to wisdom (Ackoff, 1989) 20
Figure 9 Overview of a computer based information system 20
Figure 10 NGO and Government information environment 21
Figure 11 CRISP DM Process (Wirth, 2000) 22
Figure 12 Research model 26
Figure 13 Relation between CRISP-DM and Research Question 1 27
Figure 14 Relation between CRISP-DM and Research Question 2 27
Figure 15 Relation between CRISP-DM and Research Question 3 27
Figure 16 Relation between Main Research Question, Research Model, and CRISP-DM 28
Figure 17 Relation between RQ1, Research Model and CRISP-DM 29
Figure 18 Pie chart with respondents 32
Figure 19 Labelling of Interviews 33
Figure 20 Relation between RQ2, Research Model and CRISP-DM 35
Figure 21 Relation between RQ3, Research Model and CRISP-DM 36
Figure 22 Categorisation framework for integration methods 41
Figure 23 Proposed method for selecting an integration method 48
Figure 24 Relation between proposed method, CRISP-DM and Research Questions 50
Figure 25 Relation between RQ1 and proposed method 51
Figure 26 van de Walle framework compared to 'information needs of disaster responders' 57
Figure 27 Timing of information needs 59
Figure 28 Relation between RQ2 and the proposed method 59
Figure 29 Timeline of data sources 61
Figure 30 Government Structure in Bangladesh 65
Figure 31 Data granularity loss in Government 66
Figure 32 relation between section, main research question and proposed method 67
Figure 33 Data source's total coverage of information needs 72
Figure 34 Coverage of context information needs 73
Figure 35 Coverage of coordination information needs 74
Figure 36 Coverage of flood news information needs 75
Figure 37 Coverage of location based information needs 76
Figure 38 Coverage of 'Needs' information needs 76
Figure 39 Coverage of situation overview information needs 77
Figure 40 Fulfilment of information needs plot (including timing constraints) 82
Figure 41 Fulfilment of information needs plot (excluding timing constraints) 84
Figure 42 Relation between section, main research question and proposed method 85

Table of abbreviations

Abbreviation	Meaning
ACAPS	The Assessment Capacities Project
BBS	Bangladesh Bureau of Statistics
CRISP-DM	Cross Industry Standard Process – Data Mining
DDM	Department of Disaster Management
DDMP	District Disaster Management Plan
DMIC	Disaster Management Information Centre
ETL	Extract Transform Load
FFWC	Flood forecasting and warning centre
GoB	Government of Bangladesh
Gov't	Government
HCTT	Humanitarian Coordination Task Team
HXL	Humanitarian exchange language
JNA	Joint Needs Assessment
MMS	Manab Mukti Sangstha
NGO	Non-Governmental Organisation
PIO	Project Implementation Officer
S&R	Search and Rescue
UNOCHA	United Nations Office Coordination Humanitarian Affairs
WFP	World food programme
XML	Extensible Markup Language

Preface

So, how did I get here?

About two years ago I started my master in Business Informatics, where I specialize in Data Mining and related subjects, because I want to become a data science consultant. I believe there are tremendous amounts of value encapsulated in the big amounts of data, which we as humanity store. The department at my university which specializes in the analysis of data (Business Intelligence etc.) receives multiple proposals from companies with a data mining related focus. However, I did not want to do one of these proposals which were just handed to me, I wanted to do something different.

I was very ambitious and wanted to do a very distinctive master thesis with social impact. My goal was to deliver value to the world in a social way. The opportunities this data mining field promises should not only be accessible for the large corporations, I did not want to make the rich even richer. Next to this I wanted some international experience. These motivations made me search for a data mining project in the developing world. I created a blog, changed my LinkedIn, started to go to conferences and talked to a lot of people in my network. Every week I spend multiple hours to find a suitable project, so I could help people on the bottom of the pyramid to get a better life. I had skype conversations (on very crappy internet connections) with people in among others: Kenia, Tanzania, Sierra Leone and Ghana and had talks around microfinance, farming and lots more. Unfortunately due to my very specialized requirements for a project, and the low availability of data in these developing countries I did hit a lot of dead ends in my quest.

Meanwhile, I only had one month left before I had to start with my thesis and I did not even have a project yet. I decided to also start looking for "common" data science project which did not involve the humanitarian aspect. But then I got very lucky, one of the first contacts I talked to about my thesis suddenly emailed me to introduce me to Marc van den Homberg, who was involved in the project, where I was already eight months looking for. We quickly arranged a meeting and we had a "deal" within a few days.

Marc is a consultant at TNO who specializes in information management around disasters. One of his projects involves river floods in Bangladesh, this is the project I got involved in, but you will read much more about this in the following sections.

Hopefully you will enjoy reading this Master thesis and maybe apply some of its thoughts in the real world. I really enjoyed the whole process and learned a lot around data integration, but maybe even more important, I learned a lot about myself.

Executive summary

Disasters have the potential to devastate everything in their path: people's livelihoods, information infrastructure, critical infrastructure, roads and ultimately people's lives. Floods in Bangladesh happen yearly, we can divide the floods in two basic categories (Ministry of Flood and Disaster Management, 2010). Where the first category includes the floods that happen irregularly and cause a mayor impact. The second category is the "regular" and gradual flooding that occur basically every year which is the focus of this research.

Lee et al (2011) describe the disaster response phase as: a complex process on intense time pressure, with high uncertainty, and multiple stakeholders which results in unpredictable information needs. There is a clear need for quality information in disaster response, however the disaster responders have incomplete, outdated or totally unavailable information.

All these problems have an underlying issue, there is a lot of potential in the data which is available, but the sources are not connected, so we do not have a clear picture of the situation and therefore we cannot make optimal decisions. Quick integration of the data sources is a mayor challenge in most data related projects. We decided to focus our research on data integration, whilst integration as a whole is naturally a much bigger concept (see section 1.3).

Which leads us to our main research question: *How can one determine and integrate required disparate datasets to fulfil the information needs of disaster responders in regularly recurring natural disasters?*

To answer this question we used a framework from the data mining field which is widely recognized (Kurgan & Musilek, 2006). We use and extend the CRISP-DM (Wirth, 2000) process. The result of this extension, which is our main artefact, can be found in Figure 1 and the related sections 1.3.3 and 4.1.

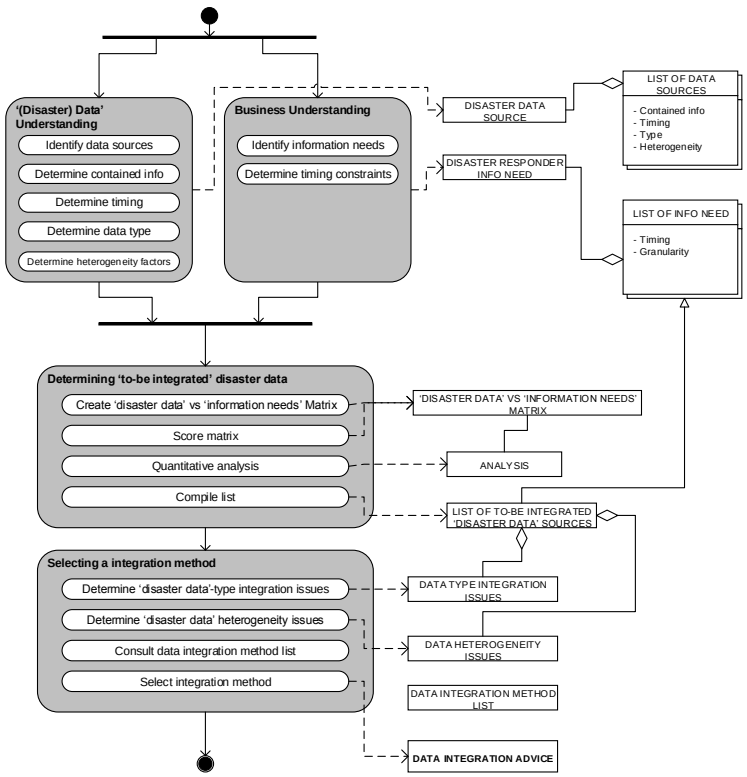


Figure 1 Proposed extension to the CRISP-DM process

The research focuses on 3 research objects, which are: the information need of disaster responders, relevant disaster data sources, and applicable integration methods. We identified 84 information needs, 15 disaster data sources, and 16 integration methods, which are elaborated upon in sections 4.2, 4.3 and 4.4, respectively. These artefacts can be reused for different contexts.

We conducted an analysis (Section 6.1.2) between the information needs and data sources. Which lead us to the conclusion that only 27% of the information needs are covered in time, whilst 63% is covered in the data sources if we do not take timing constraints into account.

By applying our method the user will get the most valuable an efficient set of disaster data sources to integrate. This will save a large integration effort and therefore a lot of time and money. Secondly, the user will get an applicable integration method for the situation.

This research suggests the most efficient set of disaster data sources and the applicable integration methods for our specific case. The advice can be found in sections 6.1.3 and 6.2.2.

We will conduct an experiment to validate the suggested methods in future research. We choose 2 data sources which cover a large amount of information needs, then we try to apply text mining to extract structured information.

1 Introduction

In this chapter we show the reader the context (1.1) in which this research takes place, we describe the problem domain (1.2), determine the scope of this research project (1.3) and present the research questions (1.4).

1.1 Context

Disasters have the potential to devastate everything in their path, people's livelihoods, information infrastructure, critical infrastructure, roads and ultimately people's lives. Disasters can be classified by multiple factors, one of them is temporal variability. First we have, slow onset disasters, which are disasters that are happening gradually (like sea level rise, drought etc.). Secondly we have rapid onset disasters, which require an immediate response (like floods or earthquakes) (Cutter et al., 2008). Disasters can also be classified whether they're "man-made" or not (e.g. war vs floods). This research addresses natural rapid onset disasters as it's focused on river flooding in Bangladesh.

1.1.1 Research partners

Several actors are involved in the described flooding context. Our partners in this research are Cordaid and TNO. Cordaid is a Dutch NGO (Non-governmental organisation) which is one of the largest development aid organizations in the Netherlands. They work with more than 600 local partners to provide humanitarian aid. Their expertise areas include healthcare and disaster response (Cordaid, n.d.). This organization is very relevant to our research due to their expertise and activities in disaster response.

TNO is a Dutch innovation institute, their mission is to: "connect people and knowledge to create innovations that boost the competitive strength of industry and the well-being of society in a sustainable way". TNO has 3000 professionals who focus on five different research themes: Healthy Living; Defence, Safety & Security; Urbanisation and Energy (TNO, n.d.). TNO is hired by Cordaid to provide technical consultancy in their TamTam project. The TamTam project tries to increase the information sharing and information usage by people on the most local and rural areas in the Sirajganj area of Bangladesh. The TamTam project focusses on sending a flood-early-warning-message via a voice based mobile message to people in the rural areas. These messages basically warn people for the coming flood. Normally, these people do not have access to this kind of information. Cordaid also trained volunteers to send text messages with water height messages to improve the predictive models from the FFWC (Flood Forecasting and Warning Centre in Bangladesh). Next to the ongoing project, Cordaid will expand their activities by gathering information during the floods via a custom build mobile app for a more effective response and need identification. Our research will be an important building block in their future strategy for the TamTam project.

1.1.2 Floods and chars

Floods in Bangladesh happen yearly, we can divide the floods in two basic categories (Ministry of Flood and Disaster Management, 2010). Where the first category includes the floods that happen irregularly and cause a mayor impact. An example of an irregular flood is the flood of 1998 which ravaged approximately 60% of the land and affected 30 million people. Other examples of large disasters are the 1954, 1955, 1987 and 1988 floods (Kunii, Nakamura, Abdur, & Wakai, 2002). The second category is the "regular" and gradual flooding that occur basically every year which is the focus of this research. There are several causes for floods including: Heavy rainfall in the Himalayas or Assam valley, landslides or blockage of natural drainage. The complete list can be found in Appendix 1.

The char islands in Bangladesh are situated in one of the three large rivers (the Brahmaputra, Ganges and Meghna) which can be described as a highly volatile living environment due to the movement of the river, these islands keep moving around and are very prone to flooding since they're in the middle of the river or on the riverbank.

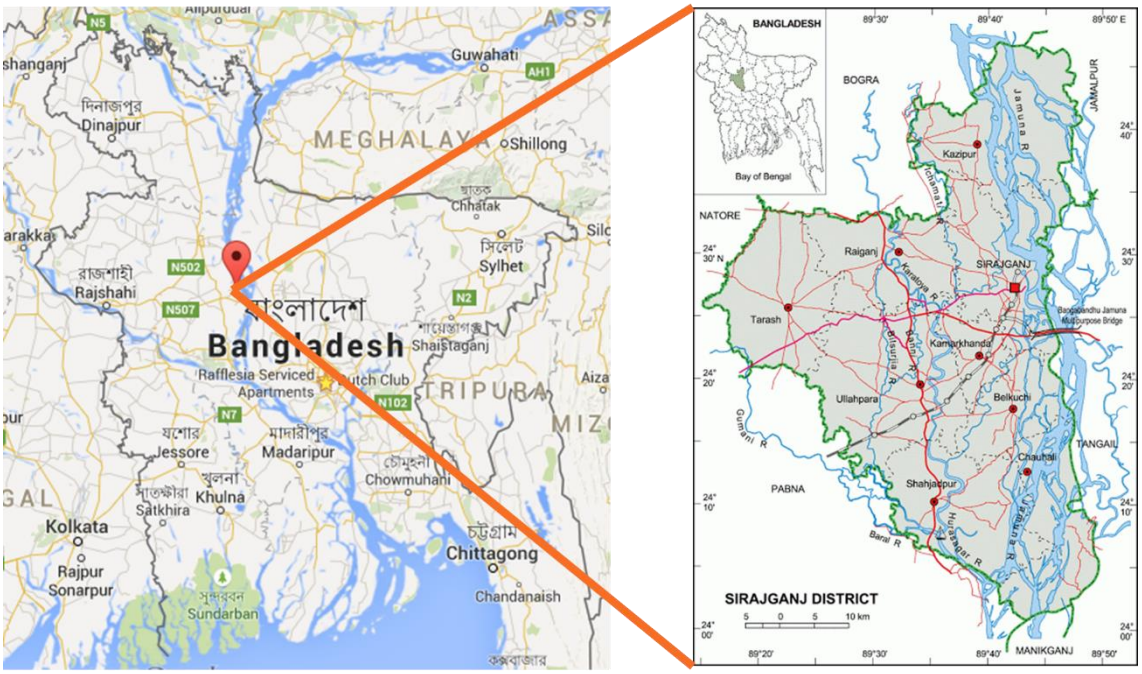


Figure 2 Geographical location of research site

Next to this, these char islands are a hard place to reach. The main livelihoods of the people on the islands consist of agriculture, fishing, and livestock. Efforts around education and health in relation to floods are minimal (Sarker, Huque, Alam, & Koudstaal, 2003). The char islands are one of the most severely and repeatedly hit places of the riverine floods and are populated by the most vulnerable people

in Bangladesh, therefore we choose the chars as a location for our research and we explain this further in section 2.3.

1.1.3 Information environment

The char islands inhabitants do not have the same amount of access to information as the people in the capital (Dhaka), but technology is also penetrating these rural areas of Bangladesh. There are 121,860,000 mobile telephone subscribers at the end of January 2015 (Bangladesh Telecommunication Regulatory Commision, 2015). The total population of Bangladesh is 158,209,034 (Bangladesh Bureau of Statistics, 2011) therefore approximately 77% of Bangladesh' population has access to a mobile phone connection. Unfortunately only 42,766,000 (about 27%) people have access to internet, either via mobile or ISP (internet service providers). Bangladesh is 123st on the world ranking the World Economic Forum's (WEF) Networked Readiness Index (NRI). A description of this index can be found below:

> *"Which measures the propensity for countries to exploit the opportunities offered by ICT. The NRI assesses the impact of ICT on the competitiveness of nations. The Index is a composite of three components: the environment for ICT offered by a given country or community (market, political and regulatory, infrastructure environment), the readiness of the community's key stakeholders (individuals, businesses, and governments) to use ICT, and finally the usage of ICT amongst these stakeholders."* (Access to Information Programme, 2013)

Next to these static figures, research shows that the trends for mobile phone acceptance and usage are promising (Tran et al., 2015). This research has to adhere to these technological constraints in order to deliver an effective solution and ensure adoption. In our fieldtrip we conducted minor observations to define the information environment of our specific research area.

1.1.4 Communities in disasters

As described in other disaster related research we divide the actors within the disaster context into three partly overlapping communities. On the one hand we have the *affected community*, which is directly and indirectly affected by a disaster and in need for direct humanitarian aid. On the other hand we have the *responding community*, which is populated by people within or outside the disaster area, who help in response and recovering from the disaster but who are not trained to do so. The third group is the *professional community*, which consists of professional and trained responders such as local and national government, NGOs and crisis response centres (Homberg & Neef, 2015). This research focuses on the responding and professional communities in a disaster situation. We don't focus on the affected community because they have the smallest role in the response to a disaster, and are not sufficiently educated and technologically equipped to use potential information from this research.

Figure 3 Communities and their overlap in disaster response

1.1.5 Phases of disaster

Disaster management can be divided into four phases, preparation, response, recovery and mitigation.

(1.) Preparedness is explained by the preparations a community takes to respond when a disaster is predicted or assumed to occur.

(2.) The Response phase consist of the activities during and directly after the disaster to preserve the livelihood of the affected but also the environment, social, economic and political structure.

(3.) The Recovery phase contains the actions to make the community back to its original state (or even better).

Figure 4 phases of a disaster management process

(4.) Mitigation entails the activities to prevent disasters or reduce the impacts. (Altay & Green, 2006).

A table included in Appendix 4 provides more detail on the general activities included in the four mentioned phases.

1.1.6 Disaster management in Bangladesh

In 2010, the Bangladesh government issued a report titled the "Standing orders on disasters" which states all relevant government bodies active in policy making, response or coordination of disasters with their respective functions and responsibilities (Ministry of Flood and Disaster Management, 2010). There are several national level coordination bodies who ensure the policy and overview activities, these are: The National Disaster Management Council (NDMC), Inter-Ministerial Disaster Management Coordination Committee (IMDMCC) and Cabinet Committee on Disaster Response (CCDR). Next to these national coordinators there are several layers in the structure of the Bangladesh government, from largest to smallest (in size and responsibility): District, Upazila, and Union. Every level has an accompanying disaster response coordination group. Next to these the LDRCG (Local Disaster Response Coordination Group) exists which combines four coordination committees into one body (Pourashiva, District, Upazila and Union) to ensure a cooperation when needed. Responsibilities include: establishing a local emergency operation centre, mobilize resources and team for disaster response, determine priorities related to relief goods etc. A full table of their responsibilities can be found in Appendix 2. Next to this, a "Local Level Multi-Agency Disaster Incident Management System" exists which has overall control on the local level. A local disaster incident manager is appointed with the following responsibilities: take control of disaster incident and establish a disaster incident management point, list resource availability etc. A full list of responsibilities can be found in Appendix 3. Since our research is aimed at the professional and the responding community, we will gather data from the actors mentioned above; however, this is further specified in our research design (chapter 2).

Next to the government multiple NGOs are currently involved in disaster management. All with a different scope, we can divide them in several groups: local, national and international NGOs. One of these international NGOs is Cordaid, which is one of the supporters of this research. They form an agglomerate with Concern Universal (national NGO) and MMS (local NGO), which are active in disaster reduction and response. Each of these three will be used for collection of data, as specified in our research design. Other NGOs active in this area are: Practical Action and Oxfam.

1.2 Problem Identification & Diagnosis

In section 1.1, we described the context in which our research is conducted. In this paragraph, we describe the problems related to information science. We focus on the different actors and their issues around information management, collection and usage. Basically, actors need information in response to a disaster, to make the most effective decisions. In the paragraphs below, the reader will find out their exact challenges. The problems can be divided into two main challenges: first the challenge for the responding and professional community to make sense of the information provided (1.2.1), and secondly challenges that arise related to data integration (1.2.2). The problems are visualised in Figure 5.

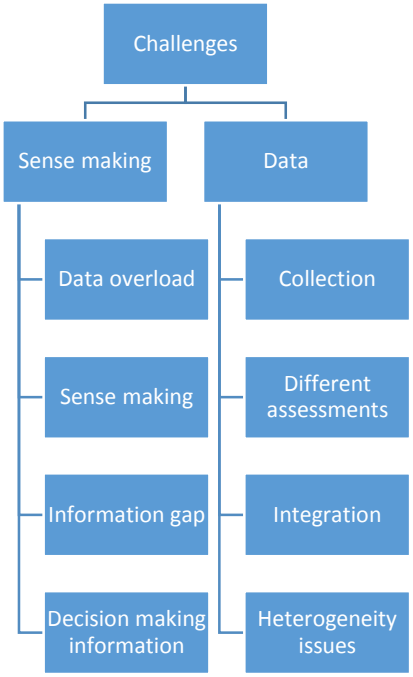


Figure 5 Visualisation of challenges

1.2.1 Sense making challenges

1.2.1.1 Data overload challenge

Two papers clearly describe the situation after a disaster strikes, regarding the information environment and its challenges (Lee, Bharosa, Yang, Janssen, & Rao, 2011; Preece, Shaw, & Hayashi, 2013). Lee et al (2011) describe the response phase as: a complex process on intense time pressure, with high uncertainty, and multiple stakeholders which results in unpredictable information needs. There is a clear need for quality information in disaster response, however the disaster responders have incomplete, outdated or totally unavailable information. Preece et al (2013) describe the following challenges:

> *"High levels of uncertainty. Extreme stress with significant consequences of actions. Compressed timelines. Significant lack of information available initially followed by extreme information overload. Difficulty assessing information quality. Multiple actors perhaps never having worked together."*

Other work emphasizes that managing information on a large scale is challenging due to differences in geography, the diversity, the extensive amount and the dynamic behaviour. Next to this, data failures, network outages and anomalous events are no exceptions (Li, Li, Liu, Ullah Khan, & Ghani, 2014).

In our disaster area multiple forms of information exist as earlier field research by Cordaid and TNO indicated, which might help the professional disaster responders. We can think of word of mouth from colleagues and affected population, static information from the government, reports from earlier

disasters, and news from television, maps with the base line situation, and maps from flooding area, raw data and reports based on the joint needs assessment (JNA) and D-form templates. The JNA is a damage and needs assessment performed by a consortium of NGOs whereas the D-form is an assessment performed by the Bangladesh Government. This is not an exhaustive list, but this short brainstorm clearly paints the picture of an (unconnected) information overload. This is a remarkable fact given the low resource environment in which Bangladesh is situated.

1.2.1.2 Sense making challenge

People interpret information in different ways, as is suggested in related research (Wolbers & Boersma, 2013). Examples are shown where people deem information irrelevant for other actors, whilst other actors are heavily dependent on this particular piece of information. Next to that, information is often incomplete in a disaster situation, but can be used for different perspectives by the actors. Shared understanding of the information is of utmost importance, we try to set the stage for this shared understanding by creating an information needs framework for disaster responders.

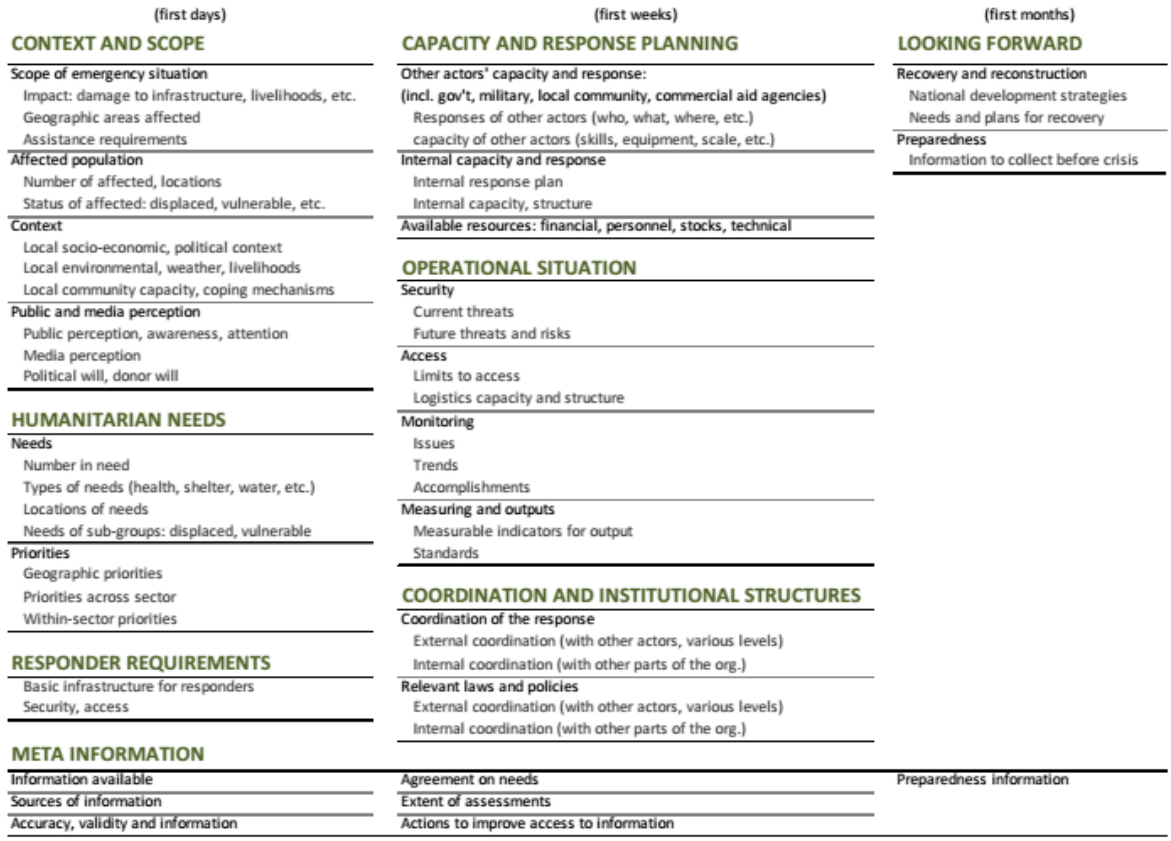
1.2.1.3 Information gap challenge

Next to the challenges created by the information overload for disaster responders another challenge exists in this specific problem domain. From the perspective of the affected and responding community an information gap exists, as opposed to an information overload. They do not have the means to express their humanitarian needs in an effective way or to acquire information about the situation of the disaster. A project in 2014 supplied the people on the char islands with voicemail messages where predictions were shared about the water levels (Cumiskey, 2014). This is very useful information before the disaster strikes, but does not really help the people when the disaster just hit, because they do not have any other relevant information like: location of food, health service for example. So, volunteer disaster responders from the affected community do not have any information to work with, their decisions are presumably just based on gut feeling or word of mouth. Possible types of useful information could be: shelter places, lead times of aid supply, evacuation routes, need sharing, assessment of damage etc. This list is not exhaustive and is merely a product of a quick brainstorm by the authors. We later examine and collect the information needs of these responders in a field trip. More on this can be found in the research design section 2.3.

1.2.1.4 Information needs for decision making challenge

As described above, disaster responders have a clear need for information. Several other organizations have tried to grasp the information needs of crisis responders. ACAPS interviewed eight respondents with regards to information in the direct aftermath of a disaster. They categorize the decisions after a crisis into two types: operational and strategic. These decisions need to be based on information which can be categorized into: pre-disaster information (how was the situation before the disaster, this information is collected before the disaster), and disaster specific information (related to the event and its impact). Examples of pre-disaster include: cultural information, background information, vulnerable groups, risks and disaster trends. Examples of disaster specific information include: needs and gap analysis, response capabilities, and operational constraints etc. (ACAPS, 2010). Another agglomerate of organizations performed a similar analysis of the information needs of disaster responders, but was focussed on the international responding community. They did not come to a definitive list but were able to develop a preliminary framework which can be found in Figure 6 (Gralla, Goentzel, & Van de Walle, 2013).

Figure 6 Information requirements in disasters (Gralla, Goentzel, & Van de Walle, 2013)



1.2.2 Data collection and integration challenges

1.2.2.1 Data collection challenge

Next to the information overload and challenges on the information receiving end, there are also challenges while collecting the data during a disaster. Morton and Levy (2011) conducted a survey on challenges in disaster data collection during the Indian Ocean Tsunami (2004, the Hurricane Katrina (2005), and the Haiti Earthquake (2010). They selected a few of the large set of key disaster indicators. These indicators were: morbidity, mortality, rapid health assessments, shelter needs assessments, nutrition and food aid assessments and infrastructure assessments. All challenges described in the article can be summarized to: environmental, political, economic, cooperation security and infrastructural challenges. A few examples from the article relevant for our problem domain are: low quality of birth registration available, low amount of pre disaster health data available, poor communication between actors (government, local government, NGOs and the affected), illiteracy, low response rates, secondary data collection with resulting misclassification bias and recall bias, lack of guidelines for data collectors, no standard data collection between organizations, relief agencies unwilling to share data (because they're competing for the same funding), and people are in hard to reach places (when not in large shelter facility). (Morton & Levy, 2011)

1.2.2.2 Different damage and needs assessments and disparate data sources

After a disaster of a "sufficient" size occurs in Bangladesh, multiple (international) NGOs perform a Joint Needs Assessment (JNA). This assessment is aimed to: Identify priority needs, provide approximate numbers of affected people, Identify severely affected unions and Upazilas, and provide initial recommendations for strategic decisions regarding resourcing and response planning. A part of this research is an analysis of the JNA since it is focussed on the needs of the affected population, while

this research is partly focussed on the information needs of the responders. These two "needs" should be a clear match, but some additional analysis is required. It might be the case that the joint needs assessment is not fully supporting the specific information needs of the responders.

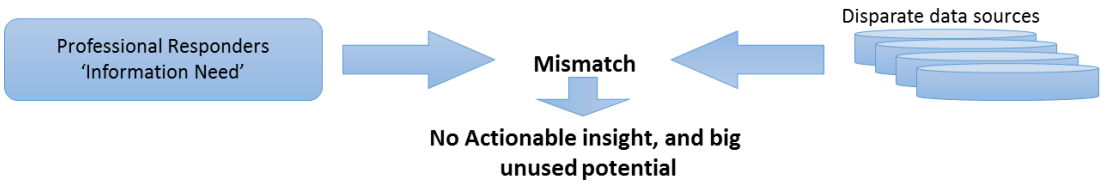
Next to the JNA the Bangladesh government also has a means to collect information after a disaster, they use their own formats called the SOS form and the D-Form. In the appendix you'll find the d-form template and the JNA template (Appendix 5 and Appendix 6). The difference and overlapping of the two will be subject to analysis.

There are a lot more data sources which could be used for effective disaster response, these could be combined with the two described assessments, this is a perfect example where the need of data integration originates, but we will get back to this in the remainder of this thesis.

1.2.2.3 Integration of the disaster related information

All these problems have an underlying issue, there is a lot of potential in the data which is available, but the sources are not connected, so we do not have a clear picture of the situation and therefore we cannot make optimal decisions. This translates to the unfulfilled information needs of disaster responders. Our problem is visualized in Figure 7. The main focus of this research will be to fill this gap. To reach this goal we need to understand the information needs of the disaster responders and we need to identify the disparate data sources. But that's not all, how do we connect these two? There are multiple perspectives to look into this problem, we describe our scoping in paragraph 1.3.

Figure 7 Mismatch between information needs and data sources



1.2.2.4 Data Integration Issues

Quick integration of the data sources is a mayor challenge in most data related projects. We decided to solely focus on data integration, whilst integration as a whole is naturally a much bigger concept (see 1.3). For example, we can think of integration on the highest level like: "Manual integration" where people are looking for data in different environments and integrate it themselves, or "integration via a common user interface", which can be compared to an online search engine. Next to this we can look at integration via middleware or on the application level, which use applications like workflow systems to integrate two applications with each other. Lastly we see integration on Data level (Uniform data access and Common data storage), where the first one is basically a logical integration of data and the data is kept separate, and the second is a physical integration of data (Ziegler & Dittrich, 2007). We are focussing on the lowest layer of integration options, as is explained in the scoping section.

Ziegler and Dittrich (2007) describe several data conflicts.

- The architectural view of the information system (which we described above)
- The functionality and content of the "to-be" integrated systems or data sources
- The type of information (multimedia, structured, unstructured, semi-structured etc.)
- The need and requirement of autonomy of the individual data or "to-be" integrated systems
- Type of usage of integrated system (read-only or also write access?)
- Performance requirements
- Availability of resources (skills, time and money)

Next to these problems, there are several types of heterogeneity between data sources as described by Ziegler and Dittrich (2007)

- Differences in OS (operating systems) and hardware
- Differences in Data Management software
- Data semantics, which is the way the data relates to a real world object (Bergamaschi, Castano, & Vincini, 1999)
- Data schemas, which is the physical structure of the data and exactly how it's stored. Types of data could be: "DateTime", Integers etc. and physical structure of the data could relate to the specific partitions in which the data is stored. (Abiteboul, 1997)
- Data models, which is an abstract representation of the data and its relations within, without the physical structure from the data schema. (Stair, Reynolds, & Chesney, 2008)
- Differences in middleware
- User interfaces
- Integrity constraints and business rules.

Another work focusses on the heterogeneity of data integration and the conflicts it creates (Boufares & Ben Salem, 2012)

- Inconsistency of the syntax, where for the same concept a different entry is used, like yyyy/mm/dd vs dd/mm/yy.
- Different measure units, where different measures are used, like pound vs kg
- Inconsistency of the representation, where for the same field different types of representation are used, like a scale for social class which differs between 1-5 vs 1-10.
- Redundancy of entities, where the same entity is present in both data sources
- Inconsistency of entities, where the same entity is represented in different ways in both data sources, for example a different street address.
- Violation of cardinality (one to many, or many to many relations for example) constraints, there might be a constraint of ten people with a specific role, while when you merge the databases this constraint might be broken
- Source dependency, we need to know the real value of an entity in data integration where multiple values exist. One might think to just grab the value that is mentioned most, but this could lead to serious problems when data sources (with wrong values) are copied and thus influence these statistics.
- Semantic inconsistencies, which is the inconsistency when the data represents two different real world objects.

In the results section 3.1 we present a framework to evaluate the disaster data sources and the integration methods.

1.3 Scoping

This part sets the stage for the rest of the thesis, the reader will find out what scope we use to solve the problems described in paragraph 1.2.

1.3.1 From data to wisdom

The first framework we use to position and scope our research, is one of the most widely used in the field of business intelligence: to depict the process which goes from raw data to wisdom (Ackoff, 1989). We do not cover this framework fully but our focus is in the process of going from data to information. Going from information to knowledge and from knowledge to wisdom is research with strong influences from psychology and organization theory, which is not the topic of this research. However, this is related to the sense making literature discussed in section 1.2.1.2.

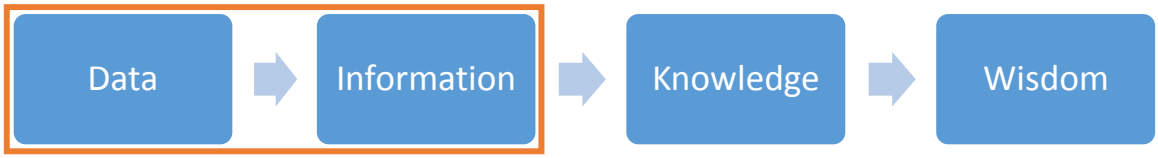


Figure 8 From data to wisdom (Ackoff, 1989)

1.3.2 Information systems architecture

Computer based information systems (CBIS) have six different components: hardware, software, people, processes, telecommunications and databases. They all cooperate to reach the common goal of the information system (Stair et al., 2008). In our case the goal would be to provide an actionable overview of the situation to disaster responders for effective decision making. A graphical overview is provided in Figure 9. The problem statement as described earlier requires a holistic approach with solutions for each of the six components. However, in our research, we focus only on the data layer, and we exclude the others. We see most potential in researching these elements because this is the base of all information systems, most systems support an export of the data, and this way we might be able to integrate all sources in one operation. Because if we would focus on the software layer, we probably would need to write connectors for every combination.

The people component of the CBIS framework is not one of our focus areas and therefore out of scope. We choose not to focus on this because we need all our focus on the data layer, otherwise our research would be too much scattered across multiple disciplines.

We could also focus on the processes of data collection, novel strategies of data collection can be found in the field, where the main players are: UNOCHA (United Nations Office Coordination Humanitarian Affairs), ACAPS and humanitarianresponse.info (subsidiary of UNOCHA). UNOCHA has developed several guideline documents which aim to streamline humanitarian data collection. An example is MIRA (Multi-cluster Initial Rapid Assessment) which is created to support assessments performed by multiple actors (Inter-Agency Standing Committee & Committee, 2012). Next to these guidelines they developed a humanitarian dashboard that helps to integrate all assessments. ACAPS is an organization mainly focused on improving humanitarian needs assessments, they offer different tools to support organizations in needs assessments. We choose however to assume the data collection as being of relevant quality, and use the data which comes from these assessments for our analysis. Nonetheless, if the data is integrated, the quality of the different sources could be assessed.

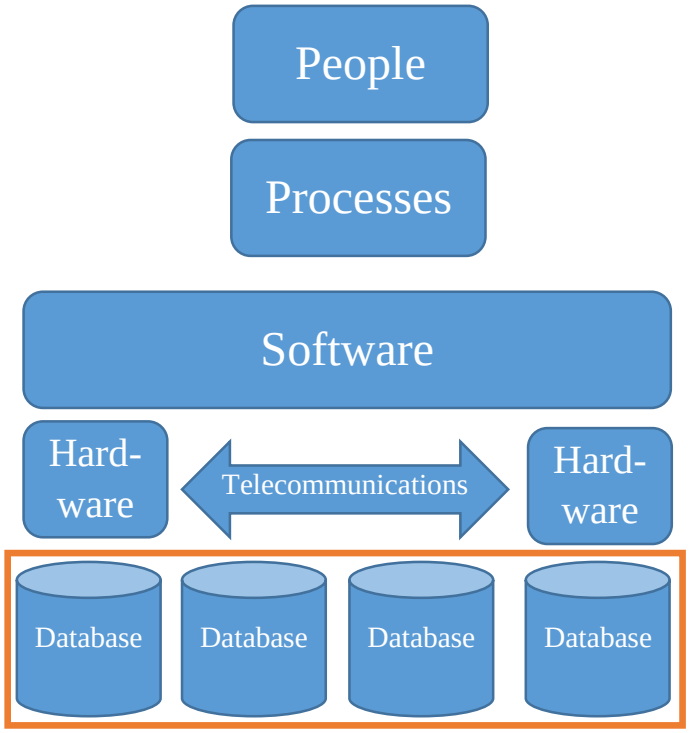


Figure 9 Overview of a computer based information system

In the humanitarian sector multiple software systems exists that deal with disaster response; basically every individual actor (NGO or Government) has their own (proprietary) software, hardware, processes, databases and telecommunications (Figure 10). It is a sheer impossible challenge to get all actors to work together and create a single information system to increase cooperation due to several reasons among which operational, political and competitive reasons. It will be very challenging to have a system that can work in several very different geographical contexts and cultures for different type of disasters. Basically only convincing every NGO to work with the same software will find much resistance since every actor will feel they would lose their 'competitive edge'.

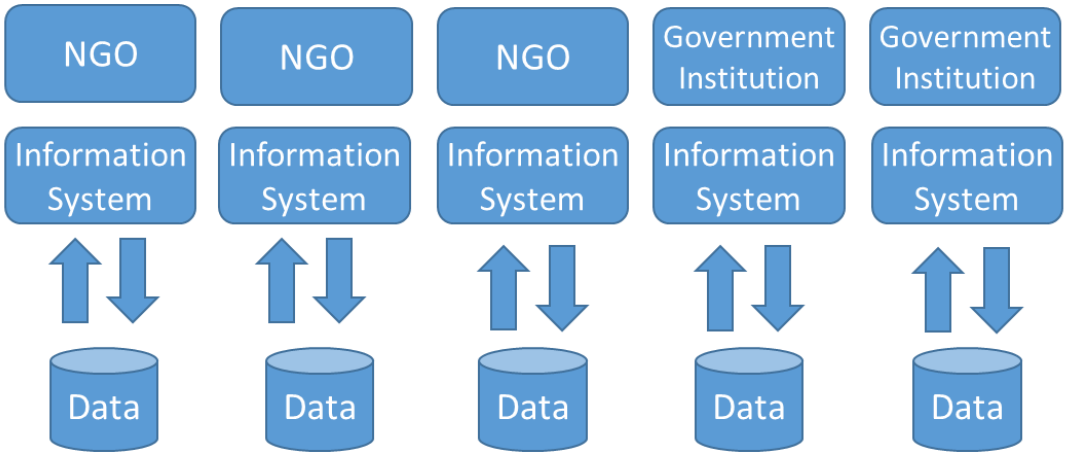


Figure 10 NGO and Government information environment

We see the data layer as an input and output component of the different actors in the domain. If we look at Figure 9 we see the data layer as the lowest layer. From our perspective, this layer is the best and required place to start with integration of the different information to get an actionable overview of the disaster situation. All actors are producing data which can be used for data integration. Since if we focus at the bottom layer we will avoid some of the political and competitive challenges which integration of information systems create. Furthermore, by validating this data integration approach we might get these actors more aligned. Next to this in the Bangladesh context data is easily accessible, a more thorough description of the data source collection can be found in section 2.5.

1.3.3 Cross industry standard process for data mining (CRISP-DM)

Within the data mining field a widely recognized framework (Kurgan & Musilek, 2006) is the CRISP-DM process which depicts an ideal process with sub steps to successfully complete a data (mining) oriented project. Our research falls in the first three parts of the framework (Business Understanding, Data Understanding and Data Preparation). The next phase (modelling) is out of scope since this is the process of building algorithms that extract information/knowledge from the data. All next phases are therefore also out of scope (since deployment and evaluation are the chronological successors of modelling). Below we further explain the phases we are covering and extending in our research.

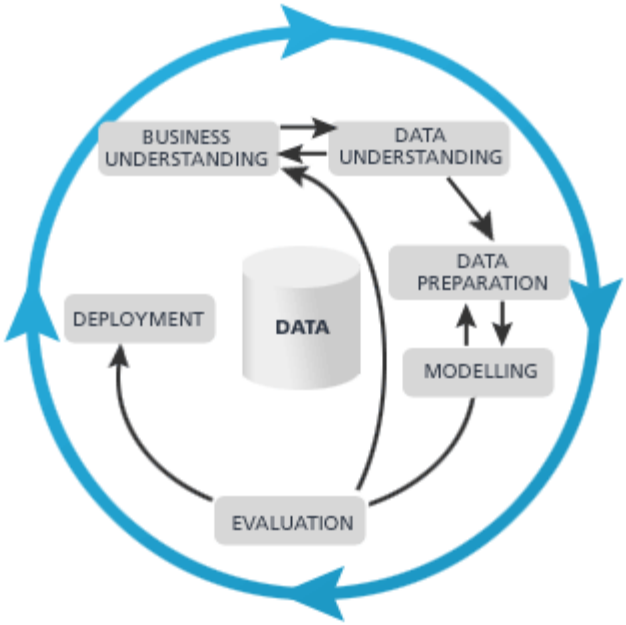


Figure 11 CRISP DM Process (Wirth, 2000)

All activities in the "business understanding" phase are aimed to understand the business requirements for the project. So, what kind of information does the decision maker require for effective decisions? For what type of decisions does he need the data mining activities? The deliverables from this phase can be found in the following sections of this research: 1.1 where we describe the disaster context, 1.2 where we describe the data related challenges, 4.2 where we present our results related to the information needs.

The activities in the "data understanding" phase are designed to understand the data from the problem context. Firstly all data is collected, described and familiarized by exploration which can be found in Section 4.3 and Appendix 11 of our research. Secondly the data quality is verified, our results can be found in 6.2. For our quality analysis we mainly focused on data integration related aspects.

Thirdly the data preparation phase is focussed to Select, Clean, Construct, Integrate and Format the data. We focus on the selection and partly on the integration sections of this phase. We feel that an increased focus on the selection activities can lead to a minimized integration effort, we show our approach in section 4.4. We base our selection of data sources on the results from the "business understanding" and "data understanding" phase.

Future research will entail an experiment to integrate the selected disparate data sources which followed by answering our main research question.

1.4 Research questions

We described multiple challenges in the information environment: information overload, information gap, data collection, information needs, disparate data sources and different damage and needs assessments. Simply put, we cannot go from data to information, and thus cannot sufficiently satisfy the information needs of the disaster responders. From our perspective, we hypothesize that we can fulfil these information needs by integrating the disparate data sets. This hypothesis leads us to the question on how we can effectively select relevant datasets and integrate them. Our research will propose a method and perspective to select the most effective data integration method for the specific case. On the practical side we will provide an advice to our sponsors on integration methods for their case.

The description of the problem and the scoping leads us to the research questions below. We're confident these research questions will direct our research towards useful, applicable and scientifically valid answers and conclusions.

Main question: How can one determine and integrate required disparate datasets to fulfil the information needs of disaster responders in regularly recurring natural disasters?

Sub-questions:

1. What are the information needs of the professional and responding community at national and local level?
    - What activities perform disaster responders in the preparedness and response phase?
    - What are important decisions for disaster responders in the preparedness and response phase?
    - What are the information needs of disaster responders in the preparedness and response phase?
    - What are the timing constraints associated with the information needs?

2. What are available and relevant disaster data sources?
    - Which data sources are available?
    - When do these data sources come available for usage?
    - What extra data is needed?
    - Which data source to use?

3. What data integration methods exist?
    - How can we evaluate these methods?
    - Which method to select?

1.5 Summary of chapter

The focus of this research will be on the responding community and the professional responders in the preparedness and response phase for floods in Bangladesh. Since they're the people saving lives when it matters the most, and we think that we could deliver the most value within these boundaries. They are the mayor players in the information environment, both in creation/collection of data and in usage of data. But they are also the main decision makers in a crisis situation, and they're still unsupported by quality information (Zhang, Zhou, & Nunamaker Jr, 2002). We will not focus on the affected community due to their small role in the information usage, but also because their role in the response phase (saving lives and providing relief aid) is smaller than the other two mentioned groups. We also described the scope of our research, we choose to focus on the data layer in the problem context. We will use the CRISP-DM to steer our research. Next to this we will create a method that extends the CRISP-DM for this case.

Earlier reports (ACAPS, 2010; Gralla et al., 2013) focus on the information need of international response, while our research focusses on a specific local case. This is an important differentiator of this research.

We introduced our research by describing the context and the scope and presented the research questions. The next chapter of this thesis entails the research design we created to answer our research questions.

2 Research Design

Here we present the design we created to conduct and structure our research. First we present the overview with a research model 2.1, which leads us to the overview of research methods and the relation to the CRISP-DM framework (2.2). The method for our main research question can be found in section 2.3. For every sub research question we created individual methods that can be found in sections 2.4, 2.5 and 2.6.

2.1 Research Model

We conduct a practice-oriented research (Verschuren & Doorewaard, 2010), which follows from the solution we try to develop for a real world problem which emerged from our partners (TNO and Cordaid). The main goal of our research is to determine which disaster data to integrate, and select integration methods for the specific case described in the introduction. To accomplish this goal we developed a generic method that can be used (hopefully across sectors) to help select disaster data sources and select an integration method. We empirically validate our approach in our results section. Our method is based on the CRISP-DM process, and extends it for (flood related disaster) data selection and integration method selection.

To conquer this goal we need to study three research objects, firstly we need to determine the information needs (research question 1) applicable for our target audience (disaster responders), secondly we need to evaluate the disaster data sources available in the context (research question 2), thirdly we need to evaluate the available data integration methods (research question 3).

An overview of the research model can be found in Figure 12. Based on the problem and described scoping of our research we can say that we take a data-driven research perspective on the problem due to the centrality of the disaster data in our research model and questions.

There are five different phases of problem solving in practice-oriented research: Problem identification, Diagnosis, Design, Intervention and Evaluation. Since a generic problem is already defined by our partners, we will focus on the first three phases starting from Problem Identification. Since our partners only broadly defined a problem, this does not lead to a fixed problem statement suitable for research. Therefore we need to thoroughly examine the context and problem domain, which is done in section 1.2. The diagnosis phase is focussed on the causes and the background of the problem, we examined that in section 1.2. Finally we perform the Design phase of the research, which is focussed on the design of a solution, this is found in our results section where we propose a method for selecting an integration approach/method (Section 4.1). The limited length of our research prohibits to also complete the last two phases.

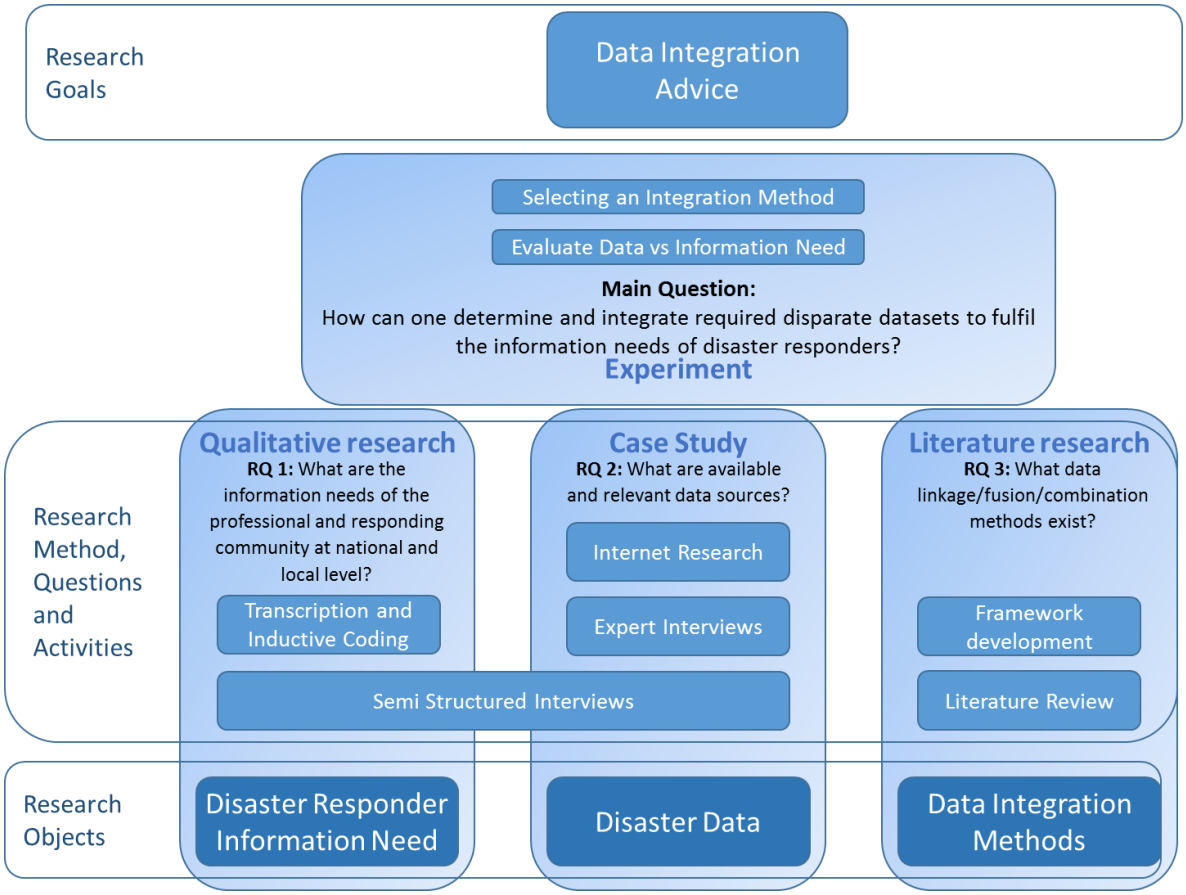


Figure 12 Research model

2.2 Overview of Research Methods

Our research method is a combination of semi-structured interviews, inductive coding, literature research, internet research, experimentation, and expert interviews. An overview of this research projects overall design can be found in Figure 12.

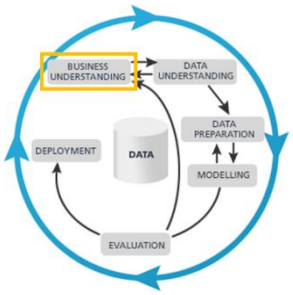


Figure 13 Relation between CRISP-DM and Research Question 1

First we determine the information needs of the professional community (research question 1) based on literature, exploratory interviews and semi structured interviews in Bangladesh. We can compare this with the "business understanding" phase in CRISP-DM (Wirth, 2000). The approach is further elaborated in paragraph 2.3.

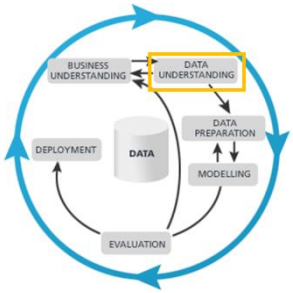


Figure 14 Relation between CRISP-DM and Research Question 2

Then we determine the available disaster data sources based on the semi-structured interviews, separate expert interviews and Internet research (research question 2). The research design of this part can be found in section 2.4. Which relates to the data understanding phase of the CRISP-DM process.

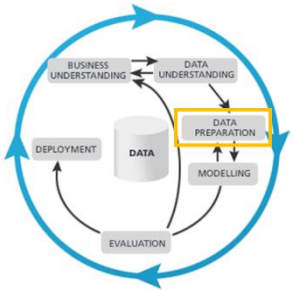


Figure 15 Relation between CRISP-DM and Research Question 3

As described in the problem description these disaster data sources are not sufficiently connected, therefore we need to determine ways to integrate these disaster data sources (research question 3). The options for integration will be gathered based on a literature review, the design can be found in 2.6.

These three research questions help us to answer our main research question for which we develop a generic method which extends the CRISP-DM to get to a conclusion. We empirically validate our proposed method by an experiment.

In future research we aim to actually integrate the selected disaster data sources, which resulted from applying our proposed method.

2.3 Research Method: How do we determine and integrate the relevant disaster data?

Main question: *How can one determine and integrate required disparate datasets to fulfil the information needs of disaster responders in regularly recurring natural disasters?*

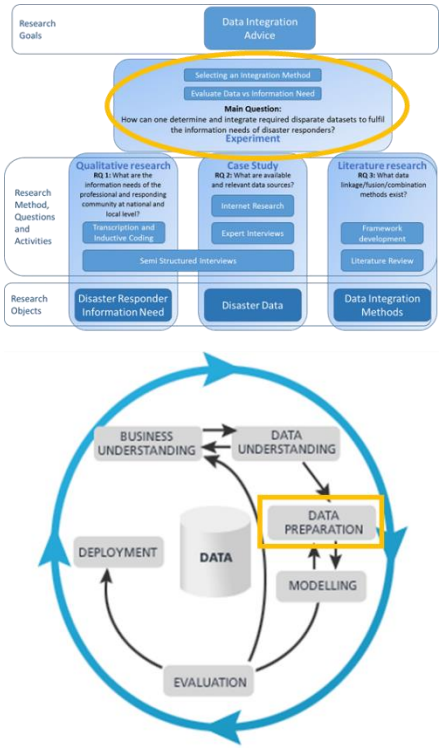


Figure 16 Relation between Main Research Question, Research Model, and CRISP-DM

To answer our main research question, we develop a generic method, which is based on the CRISP-DM process and can be seen as a specification or extension to the method (see section 1.3.3). We explain the connection between the process and our method in the result section. Afterwards, our method is empirically validated by applying it in a case study around the Bangladesh Floods in 2014. In this experiment we use the information needs from research question 1 (section 2.4), the disaster data sources from research question 2 (section 2.5), and the data integration methods from research question 3 (section 2.6)

The method is visualised in section 4.1. We use the modelling technique associated with method engineering (Van De Weerd, Brinkkemper, Souer, & Versendaal, 2006).

2.4 Research Method: 'Information Need' Identification (RQ 1)

What are the information needs of the professional and responding community at national and local level?

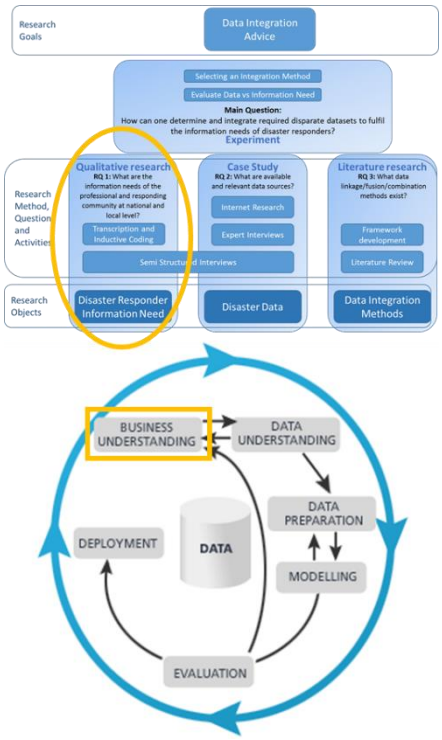


Figure 17 Relation between RQ1, Research Model and CRISP-DM

To collect research data for answering these research questions we chose a qualitative approach. We want people to think creatively, so we do not want them bounded by a fixed and structured questionnaire. Next to this reason we only found one partly applicable framework in our literature study which means we could not create a reliable structured questionnaire. These factors justify a semi structured interview approach. The questionnaire, which can be found in Appendix 10, is designed to gather the information needs of disaster responders. Because we want to increase the reliability of our research we also asked the respondents about their actions and their decisions. This way we approach the "information need" from a different angle, since these can be cross-referenced to find blind spots in their expressed needs. This is also done to ensure the bias of non-recollection for our respondents, we assume people can better express "what they have done" than to express "what they need", because the latter requires a higher capability of vision and creativity. The exact term for our approach is an 'oral history semi structured interview', which is focused on the oral narration of a person's history, beliefs, values, behaviour, stories etc. (Bryman & Bell, 2007) However, semi structured oral history interviews bring at least two problems, the first one is bias introduced by bad memory (Grele, 1991) and the second is memories could be perceived unimportant (Samuel, 1976). But since our research has an explorative focus (we do not know the information needs and do not want to presume anything) we let the advantages of this oral history interview approach outweigh the problems.

Next to gathering the information needs of the responders, the researchers observed the local context to identify other information related challenges and opportunities. Observations can be found in the results section at paragraph 5.

Qualitative research is conducted based on the following phases (Bryman & Bell, 2007):

Phases 1 to 4 are further elaborated below. Phase 5 and 6 are presented in the result section.

2.4.1 General Research Question

What are the information needs of the professional and responding community at national and local level?

2.4.2 Selecting Relevant Sites and Subjects

Our target group consists of disaster responders. Different actors participate in the response to a disaster: first we have response from the affected community and (local) private sector, secondly from the NGOs and/or the Government. We can also divide the actors by high or low level positions (National or Local Response). We target all of these subgroups to get a comprehensive overview and to see which information needs each actor has (if there is overlap, to discover everybody's preferences). Therefore we chose the following five actor groups in our problem domain: National Government, Local Government, National NGO, local NGO and Responding community.

Our respondents were selected based on the following three key criteria.

- Has participated in one or more disaster responses or preparations and most notably also in the 2014 flooding.
- Is part of the professional community (government, NGO) or the responding community (local disaster management volunteer, local private sector)
- Has either a local, district or a national focus on disaster response (as for example a ground-level responder or as a more high level coordinator)

Table 1 Interview target group

Actor	Organization	Focus	Roles
NGO	Concern Universal	National	• Disaster Response Coordinator • Disaster Preparation Coordinator • Ground level responder
	JNA consortium	National	
	MMS	Local	
Responding community	Local Disaster Responders	Local	• Imams • Teachers • Entrepreneurs • Farmers • UISC/Digital centre entrepreneurs • Volunteer Disaster management committees
	Private Companies	Local	
Government	Department of Disaster management	National	• Disaster Response Coordinator • Disaster Preparation Coordinator • Ground level responder
	Upazila Disaster Management Committee	Local	
	Union Disaster Management Committee	Local	
	Union Parishad	Local	
	Ansar and Village Development Police	Local	

The district of Sirajganj is one of the most flood prone areas of Bangladesh, this leads to a continuous presence of experienced NGOs and Government officials active in disaster response. We have a high chance to find respondents which adhere to our constraints in this region so we chose this as our target region. Based on some informal discussions we determined what NGOs are active in this region and approached them via email. All people who were approached responded to the request with a positive reply. To get more respondents we used a snowballing technique. We talked to six High level NGO employees and to ten low level NGO employees.

The local government officials were approached by one of our contacts from the local NGO and Cordaid, we organized three focus group sessions. The most important respondents are: one high ranking disaster manager on National level, two Upazila Administrative Officers, two Union Chairman, ten low level government staff, six locals from the responding community. The full list can be found in Appendix 9. And we graphically visualised the distribution in a pie chart (Figure 18)

Table 2 Distribution of Interviews (Respondents vs Transcriptions)

Group	Nr. of People	Nr. of Interview & focus group transcriptions
NGO low level	10	2
NGO high level	6	6
Government low level	14	4
Government high level	1	1
Responding community	6	2
Total	37	16

*some interview were conducted in a focus group session, therefore the lower amount of transcriptions as compared to interviewees.

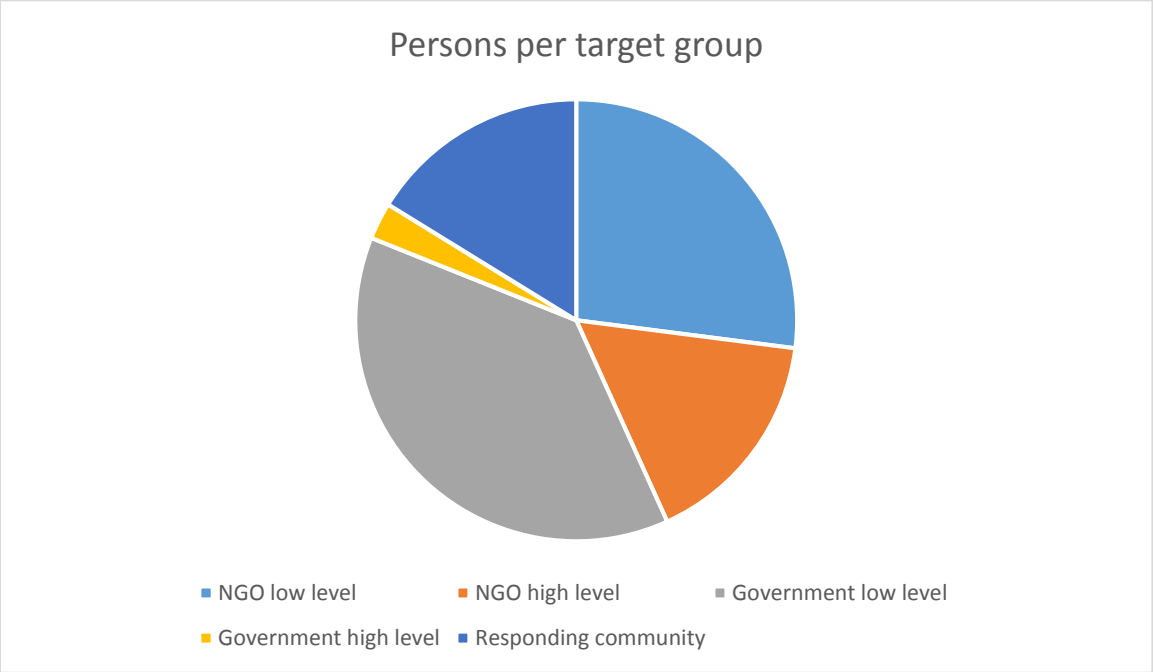


Figure 18 Pie chart with respondents

2.4.3 Collection of Relevant Research Data

To get the most valuable results, interviews were recorded on a recording device, and conducted in a private setting with a dedicated time agreement (Bryman & Bell, 2007). There are issues associated with the recording of interviews: high amount of work because all interviews need to be transcribed and possible hesitance from the interviewee because of recording. However this does not outweigh the advantages of recording because it corrects the natural limitations of the researcher's memory, allowing a more thorough analysis. If the research data is kept it helps in raising the validity of the research, since other researchers can evaluate the research data and the conclusions. Interviews on a high level (NGO and Government) were all recorded. Transcriptions can be found in Appendix 14.

The interviews on the lower level response (NGO, Government and responding community) were not recorded due to difficulties with the environment and group setting. Most interviews were done outside with a high amount of noise and therefore a low quality of recording. Since most of the respondents were non-English speaking we used an interpreter, this would result in recordings with a lot of unusable parts, and therefore recording was not ideal. The researchers chose to take notes at these locations to

better be able to capture the response of our interviewees. Directly after each interview the notes were written to a more elaborate interview report.

Bryman and Bell (2007) mention several useful criteria for a successful interviewer, each of which is kept in mind when preparing and conducting the interviews. These principles can be found in the appendix. Bryman and Bell also describe several types of questions, these were used for structuring our interview guide: Introducing questions, Follow up questions, Probing questions, Specifying questions, direct questions, Indirect questions, Structuring questions, Silence, and Interpreting questions.

2.4.4 Interpretation of Research Data

The output of these interviews are approximately ten hours of voice data and hand written notes (from the not-recorded interviews). Due to time constraints we transcribed only the most vital parts of the interviews. Hence we will focus on information relevant to our research questions. To validate our results every respondent is asked to review and validate their interview transcription. Only some minor changes were implemented, since the response was limited.

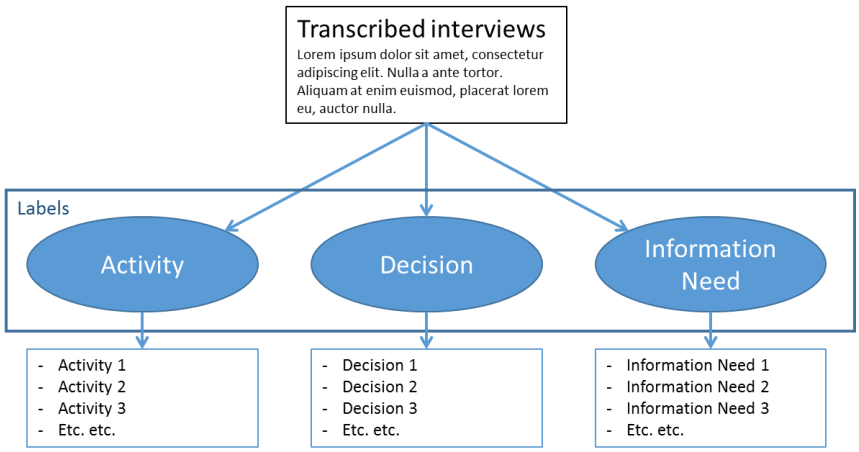


Figure 19 Labelling of Interviews

After the transcription of the interviews the researchers read every transcription and interview report to get familiarized with the research data. After the familiarization phase, we read every individual interview and tagged sentences and paragraphs by the following themes: Activity, Decision and Information Need (see Figure 19). For this process we used NVIVO 10 for Windows. Every interview was read once for every theme, which means every interview was read and tagged three times. This resulted in three lists with subthemes occurring within each theme. For example, a subtheme within the Activity theme is: 'Sharing relief goods'. These subthemes emerged by reading through the transcriptions. We made cross sections per actor to see differences in the different themes. We have a long list of very specialized information needs, activities and decisions. For each of these themes a clustering exercise is performed to make more sense of the research data. This clustering is based on experience emerged from the familiarization phase.

Our list of (low level) information needs (subthemes) is plotted on the framework proposed by Gralla et al (2013) to see where the differences are, results can be found in the results section. Next to the comparison with the academic literature we want to determine how we can fulfil these information needs.

After the finalization of these three lists (activities, decisions and information needs), we asked two domain experts to review them for validation and add additional information needs which emerged from either the activities/decisions or from their review of the information needs list. This resulted in a sharper and more extensive list of information needs.

2.4.5 Analysis of threats to the validity of the results

External Validity: relates to the degree in which findings can be generalized. Normally qualitative research is marked with a low external validity due to the high influence of the sample size, however, we feel the results of these interview can be easily used for other floods, which occur in several other Asian countries. Therefore we would like to share the results with a "medium" external validity.

Internal Validity: relates to the strength of the relation between theory and observation. This is a strong point of qualitative research according to Bryman and Bell, therefore the internal validity will be high.

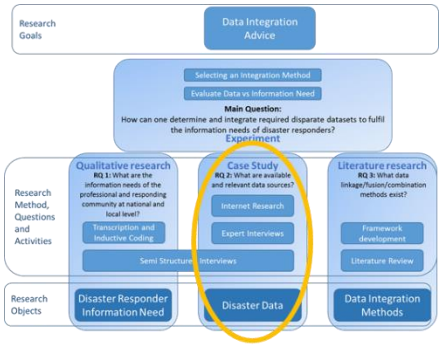
External Reliability: relates to the degree a study can be replicated. We clearly described our research approach and documented everything carefully, therefore the external reliability of this research will be high.

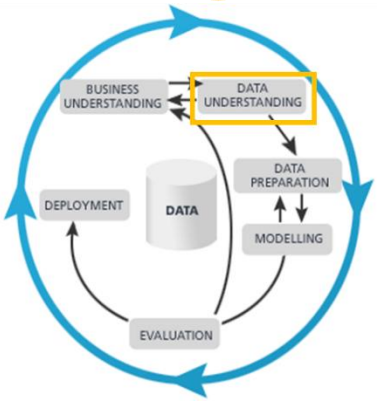
Internal Reliability: relates to the agreement between observers on what is perceived. Unfortunately this is a weak point of the research since only one observer will be used due to financial and time constraints.

Triangulation: By interviewing multiple stakeholders we will triangulate our results. But the results are also triangulated by incorporating earlier research from different authors (Gralla et al., 2013).

2.5 Research Method: Identifying Disaster Data Sources (RQ 2)

What are available and relevant disaster data sources?





In this paragraph we show the reader how we identified disaster data sources to develop one of our research objects.

2.5.1 Case: Flood 2014

To find disaster data sources, we choose to focus on a specific case in the past. This way we are able to validate our approach and we can focus in our future research for an implementation in an ongoing or future disaster. The most accessible, recent and relevant case is the flood of 2014. This one is the most recent and the most severe of the last years, because of the recent occurrence it's quite easy to extract the relevant disaster data.

A disadvantage is that we do not have the pressure of the real disaster in the response phase which decreases our external validity, since, how do we validate our approach works with the obvious time constraints a disaster enforces? However, it also raises the possibility for us to come up with the most optimal solution, and in the future we could apply it in a real world case.

2.5.2 Semi-structured interviews

Some questions in the questionnaire (Appendix 10) are about the disaster data and information usage in the flood situation of 2014. These disaster data sources were obtained during the field trip to Bangladesh.

2.5.3 Internet research

The approach we choose is internet research to find disaster data sources around the flooding in Bangladesh. Because we assume most disaster data is publically available, and otherwise it is not part of the "information overload" described in the problem definition section. We applied more or less a snowballing method, starting at some websites of known government and NGO institutions who are active in this domain and browsed on from there.

2.5.4 Domain Experts

Next to this we interviewed two domain experts who are currently implementing a disaster data collection application in the same geographical location. We interviewed them and asked for the disaster data sources they use for their application, next to this we asked them to compile a list of relevant government and NGO institutions who are collecting and possibly sharing disaster data.

2.6 Research Method: Identifying Data Integration Methods (RQ 3)

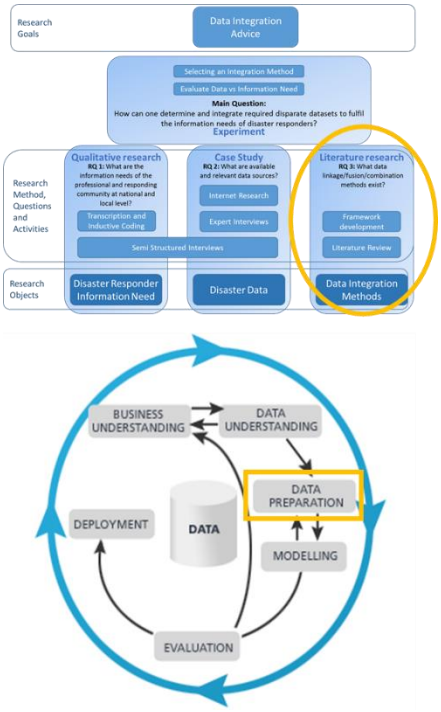


Figure 21 Relation between RQ3, Research Model and CRISP-DM

We decided to conduct a literature study to determine the available data integration methods. A literature study starts with the planning, then the actual conducting and finally reporting. We will use this framework for our literature review (Wohlin et al., 2000). We are interested the main features, strengths and limitations of the methods, and not in an in-depth analysis or to what extent they have been validated. This, along with the fact that the review we are performing is very broad (all different kinds of integration methods), made us choose a mapping study instead of a systematic literature review. This decision saved time for more valuable research endeavours.

2.6.1 Planning

Relevant research questions:

- What data integration methods exist?
    - How can we evaluate these methods?
    - Which method to select?

Based on these research questions, there are three sub deliverables we want to deliver from this literature study:

1. The first goal of this literature review is to find out applicable data integration methods. The result will be a list of non-domain and non-application specific data integration methods. The results can be found in section 3.2
2. In the second track we determine what data integration methods are currently and previously used in the disaster information management domain. Results can be found in section 3.3.
3. The third goal of this literature review is to determine an evaluation framework for the found integration methods. Results can be found in section 3.1

To get some domain understanding and to focus our approach we first did an orientation search around the keywords: linkage, integration and fusion and unstructured, text, relational, structured data. Based on the review of the articles we determined separate sub themes for further investigation, these themes include: semantic integration, text mining, NoSQL, data spaces, data warehousing. Further specification of the search terms can be found in Table 3.

We use google scholar to search for the primary academic studies. From this preliminary search we determine the most relevant articles, from these articles we go on with a snowballing approach to reduce the amount of false positives.

The selection criteria for the general track are:

1. The study integrates data types which are used in our problem domain (for example: the social media penetration in rural Bangladesh is low, therefore literature on social media integration is irrelevant). Relevant data types are: Excel sheets, relational databases, pdf, text, geographic information.
2. The method is portable in a developing world context
3. The study describes in which situation its applicable
4. The study describes the minimum required criteria for implementation

The selection criteria for the domain specific track are:

1. The study integrates data types which are used in our problem domain
2. The study is applied in a disaster context

First we determined a framework for evaluation, which can be found in section 3.1, based on that we extracted relevant information.

2.6.2 Conducting

We conducted the review based on the following search terms:

Table 3 Search terms for literature review

Search terms	Search terms
Evaluation framework	• Data integration evaluation framework • Data integration problems • Data integration conflicts • Heterogeneous data conflicts
General methods	• allintitle: (heterogeneous OR disparate OR dissimilar) data integration/fusion/linking • allintitle: (heterogeneous OR disparate OR dissimilar) information fusion/integration/linking • allintitle: survey information fusion/integration/linkage/linking • allintitle: survey data fusion/integration/linkage/linking • linking AND "unstructured text" AND relational data • integrating AND "unstructured text" AND relational data
Domain specific integration methods	• allintitle: "data linkage" (Disaster OR Emergency OR Crisis) • allintitle: "data integration" (Disaster OR Emergency OR Crisis) • allintitle: "data fusion" (Disaster OR Emergency OR Crisis)

2.6.3 Reporting

Goal	Articles found	Articles Selected	Used in final paper
Evaluation Framework	100 approx.	4	2
General Data Integration Methods	10.000 approx.	51	16
Domain Specific Integration Methods	22.000 approx.	37	11

2.7 Summary

We shared the research design we created, to answer our research questions. Every sub question has its own (combination of) specific research methods, which are elaborated in this chapter. The next chapter creates a theoretical framework based on the literature review from research question 3.

3 Theoretical Framework

What data integration methods exist?

This section defines the theoretical framework based on the literature review conducted for research question 3. First we determine evaluation frameworks that are eventually used to select the optimal integration method for our case (3.1). Then we present the relevant integration methods from the literature (3.2), followed by the approaches from the current disaster information management practices (3.3). The methods found in this literature review are used in our data integration advice.

3.1 Evaluation Framework

To effectively evaluate the found linkage and integration techniques/methods we need to select a framework. We can use this framework to informed decisions when choosing an integration method. We found 2 articles in the academic literature, among others, which list issues regarding data integration. These issues can be used to evaluate the methods. A broad description of these issues can be found in the introduction of this thesis, in section (1.2.2.4). We combine the two to create a generalizable framework to be applied across cases.

Table 4 Evaluation framework related to data

Location of data		Is the data located on the same server, or in disparate systems?
Structure of data		Which resembles the format of the files (text, pdf, excel etc.)
Data model		Which is an abstract representation of the data and its relations within, without the physical structure from the data schema. (Stair et al., 2008)
Data schema		Which is the physical structure of the data and exactly how it's stored. Types of data could be: "DateTime", Integers etc. and physical structure of the data could relate to the specific partitions in which the data is stored. (Abiteboul, 1997)
Heterogeneity	Inconsistency of syntax	Where for the same concept a different entry is used, like yyyy/mm/dd vs dd/mm/yy.
	Different measurement units	Where different measures are used, like pound vs kg
	Inconsistency of representation	where for the same field different types of representation are used, like a scale for social class which differs between 1-5 vs 1-10
	Redundancy of entities	Where the same entity is present in both data sources
	Violation of Cardinality	One to many, or many to many relations between entities
	Semantic	Which is the inconsistency when the data represents two different real world objects.

Secondly we created a framework which is focussed on the constraints we get from the developing world context.

Table 5 integration method evaluation framework related to developing world

Functionality of to-be system		Required to be able to combine data sources which are ready for analysis and visualisation
Availability of resources	Skills	brings both opportunities and constraints, on the one hand we can get very cheap manual labour to integrate the data, on the other hand, we don't have an extreme high abundance of skilled IT personnel to our exposure. Time is obviously also a constraint, due to the pressure of the disaster we cannot wait for everything to be 100% perfect, we might need to choose the "next best answer" instead. Money is obviously also a tough constraint in the developing world context.
	Time	
	Money	
Low resource implementation	Internet connectivity variability	The "developing world implementation constraints" require us to limit ourselves to personal computers for the integration of the datasets, we don't have the processing power to use grids of multiple computers for example, next to this, for the integration we must not rely on a stabile internet connection.
	Processing limits	

3.2 Integration methods

Based on the research design of research question 3 (section 2.6) we reviewed the current status of data integration in the academic literature. Below you'll find a summary of most interesting and applicable methods.

Table 6 Identified Data Integration Methods

1. Generating mapping schemas	5. Peer-to-Peer data management	9. Collaborative integration	13. Extract structured information from unstructured text
2. Adaptive Query Processing	6. Data Warehousing	10. Dataspace systems	14. Integrating unstructured data into relational databases
3. XML	7. Extract, Transform and Load processes (ETL)	11. Humanitarian exchange language (HXL)	15. Ontology guided information extraction from unstructured text
4. Model management	8. Personal data integration	12. Usage of text mining algorithms	16. Unstructured information integration through data-driven similarity discovery

To structure our review, we divide the methods into different categories. We defined two axis' on which we can categorise the methods. The first axis is the focus the methods has, it can either have an high level focus, where it is more oriented towards the human factor of the data integration, whereas there can also be a more low level focus, which has a more technical focus. The second axis is the data type which the method is suitable for, where it can either have a focus towards unstructured data (text, audio, video etc.), or it can have a focus towards structured data (excel, relational databases etc.). A further analysis, based on the framework from section 3.1, can be found in the result section 4.4.

	Low level focus	High level focus
Unstructured data	12, 13, 14, 15, 16 *(Section 3.2.3)*	10 *(Section 3.2.4)*
Structured data	1, 2, 3, 4, 6, 7, 11 *(Section 3.2.1)*	5, 8, 9 *(Section 3.2.2)*

Figure 22 Categorisation framework for integration methods

There is a plethora of integration methods, we used some 'literature reviews' from academic sources as a starting point for our review. We found an excellent paper from 2006 which provides an overview of research directions in data integration (Halevy, Rajaraman, & Ordille, 2006). We identified 16 separate data integration methods from the academic literature. Which are depicted in Table 6.

3.2.1 Low level and Structured Data Integration

The first category we describe is the low level focus with structured data integration. There are 7 methods associated with this category. The following methods are described below: Generating Mapping schemas, adaptive query processing, XML, model management, data warehousing, ETL and HXL.

The first one is generating mapping schemas (1), which is the process to automatically generate semantic mappings between sources and a mediated schema. This was followed by the automatic creation of a schema based on other input than the schemas themselves. Finally, Machine Learning techniques were applied to predict mappings between unseen schemas (Halevy et al., 2006).

The second method is Adaptive Query Processing (2), which tries to optimize the queries fired at disparate datasets. This leads to challenges because unlike a regular database management system (DBMS), which has query optimization and query processing divided, a data integration environment has much less information on optimization because of the dynamic nature. A unified architecture for adaptive query processing was found (Halevy et al., 2006).

The third research direction is XML (3), which is a unified and machine readable data exchange mark-up language. This solved a lot of the syntactic issues in data integration, but failed to address the semantic integration issues. This lead to XML files which were still very much tied to the applications which used them, which still lead to challenges in data integration. However, this movement actually increased the urgency and need for data integration research (Halevy et al., 2006).

The fourth research direction is Model management (4), which goal is to create an algebra for manipulating schemas and mappings, in order to not repeatedly create the same operations for every new model or context. Sub areas of research are schema merging and composing mappings (Halevy et al., 2006).

Data Warehousing (6) is a technique where a common data storage is used for the integration of the data. These are loaded by Extract, Transform and Load processes (ETL) (7) in which the integration logic is encapsulated. There are multiple architectures for implementation of Data Warehouses. The first one is an independent data marts architecture, where every data mart operates individually for a specific goal, the data among every mart isn't integrated however. The second is a data mart bus architecture, which incorporates middleware to integrate the individual data marts. Thirdly a hub-and-spoke architecture exists, which is developed iteratively for one specific goal at a time, this could lead to redundancy and latency however. Fourth is the centralized data warehouse, this is one big data warehouse that serves the needs of all actors, this could give a holistic and perfect view of the problem domain, it is very expensive and time intensive to implement however. The last one is a federated data warehouse, which is a compromise technique, it leaves all data in their respective physical location and accesses the sources as needed, most integration is done by middleware (Turban, Sharda, & Aronson, 2008; Ziegler & Dittrich, 2007). Another approach within data warehousing is operational data stores, which are basically loaded with fresh data after the sources are updated, the sources are not cleansed or aggregated and data histories are not supported (Ziegler & Dittrich, 2007).

The last method in this category is the humanitarian exchange language (11) (HXL) which is a project from the UN OCHA (United Nations Office for Coordination of Humanitarian Affairs), aiming to improve data management and sharing for effective disaster response. This approach mainly consists of a common terminology agreed by all actors. This terminology (and its corresponding tags) are used to put a layer on top of the data, this way the data is made interoperable (Hendrix & Keßler, 2009). The work involves a tool to assist in tagging the data. The usage of this method was validated in an operational dashboard which only used HXL data. HXL is related to XML since it is also a language to mark specific parts of the data with tags, however, XML is more focussed towards text files, and whilst HXL is more focused to structured files (excel for example).

3.2.2 High Level and Structured Data Integration

This second section describes the identified methods related to structured data but with a higher level focus. Methods associated with it are: peer-to-peer data management, personal data integration and collaborative integration.

The first direction is the field of Peer-to-Peer data management (5), which is related to widely known peer to peer (p2p) file sharing concepts like (Bit torrents and Popcorn Time etc.). This direction has two main advantages, first many organizations want to share data but do not want to take responsibility for creating mediating schemas and mappings. Mediating schemas are used to map certain entities in the data to entities in another dataset, in order to make them coherent. This is taken away by the distributed nature of this approach because the actor only needs to create mappings for its neighbours. Secondly, it might not be possible to create one single mediated schema for every data integration situation (Halevy et al., 2006).

A second approach is personal data integration (8), which is a special form of manual integration. In this approach a declarative integration language is used to define tailored integrated views, each view exactly matches the information needs by showing all relevant entities with applicable semantics from the real world. This leads to a very personalized view on the data (Ziegler & Dittrich, 2007).

The third approach is collaborative integration (9) (also a manual integration form), which is built on the idea of a user's contribution before usage of the data integration system. Users need to answer questions about mappings, these answers are used to enhance the integration (Ziegler & Dittrich, 2007). This is also strongly related to coordinated data scrambles (Campbell, 2015), which is a method currently applied in the humanitarian response world.

3.2.3 Low Level and Unstructured Data Integration

Our third category focusses on unstructured data integration with a low level focus. Methods associated with this category are: text mining, extracting structured info from text, linking text to relational databases, and information integration by data driven similarity.

The first approach we see valuable for our case is the usage of text mining algorithms (12) to integrate the available data sources, since in our context a lot of information is shared in unstructured pdfs. We reviewed some literature review papers (Agrawal & Batra, 2013; Gharehchopogh & Khalifehlou, 2012; Gupta & Lehal, 2009; Tan, 1999) in this field and found some applicable methods. The main goal for text mining is to extract interesting patterns from unstructured text (Gupta & Lehal, 2009). The main problem of text mining is to extract the explicit and implicit relations found in natural language. There are multiple sub areas of text mining, we only discuss the ones applicable for our problem domain. The first is Information Extraction, which addresses the problem to get structured information by transforming the text of the unstructured sources. It extracts key phrases and relationships from the text based on the text itself. After this process the structured data can be analysed for further knowledge gain (Gupta & Lehal, 2009). A second subfield of text mining that could be relevant to our field is the topic of summarization, where large portions of text are automatically summarized to some key phrases. This can be extremely usable/useful in the disaster context, where a disaster responder does not have the time to read through every document due to the time pressure he faces in the response and preparation phases. Summarization focusses on the user level of integration, not on the data level, as is our scope (Gupta & Lehal, 2009). A third subfield of text mining possibly applicable in our case is the topic of 'question answering'. This is a tool in which a human can pose a natural language question and the machine will provide an adequate answer, not by giving a link to a website (like google), but by extracting that answer and presenting it to the user. This technique could be very effective in our case since the responders can get focused answers. A fourth subfield of text mining is association rule mining. This is a topic widely applied in data mining, but also found its way to text mining. Algorithms of this type try to determine terms within the text that are frequently mentioned together. This can be

used to get increased insight into the large amount of documents which are shared in the disaster response.

A clear example of unstructured integration is a method which links text documents to structured information (13) (Chakaravarthy, Gupta, Roy, & Mohania, 2006). This system basically identifies the entities from the database in the unstructured text (even when the entities do not exactly match because it exploits the available context). They give an explanatory example: a retail store receives emails with complaints or feedback from its customers, the store doesn't exactly know when the situation of the complaint has arisen, but by applying this system, they could tie the complaint to a specific transaction in their ERP system to increase the effectivity of their response to this complaint, or to better solve the issue (Chakaravarthy et al., 2006).

A third approach is focused on integrating unstructured data into relational databases (14) (Mansuri & Sarawagi, 2006). The authors propose an approach to use statistical models for extracting and matching structure from the unstructured sources. The system is capable of loading unstructured records into columns across different tables in the database. They first train a classifier to detect the entities from the database in the text, then the algorithm can automatically fill the database based on a set of text documents. The example the authors use is to fill a database with authors, journals and article titles (Mansuri & Sarawagi, 2006).

A fourth approach is ontology guided information extraction from unstructured text (15) (Anantharangachar, Ramani, & S, 2013). They propose a method that fills an existing ontology from unstructured text, the type of data they extract are semantic RDF triples, and these can afterwards be used for analysis. A RDF triple is a very basic data type, it only consists of subject-predicate-object, and an example is: "Sirajganj-IsSituatedIn-Bangladesh". An ontology is a predefined agreement along domain experts about the meanings of concepts and the relations within a specific domain (Anantharangachar et al., 2013).

A fifth approach is unstructured information integration through data-driven similarity discovery (16) (Ananthanarayanan, Reinwald, Balakrishnan, & Yee, 2009). This method is focussed on the actual data, without focussing on the meta-data. An experiment was conducted and promising results were obtained. They used multiple existing tools from NLP (natural language processing) and data mining. These techniques are applied to collect the unstructured data, identify the related sets, and relate these sets to reference data sets.

3.2.4 High level and Unstructured Data Integration

On the high level and unstructured data integration category, we have only one approach, which is dataspace systems (10). Which was proposed in an article by Franklin et al (2005) which proposes a database management system that is not focussed on rigid relational database structure and heavy integrated data warehouses. They propose a DataSpace Support Platform, which is a higher layer of abstraction across all databases which offers a specific set of services so developers can focus on developing applications instead of on data management issues. One of the key factors is the "pay as you go" principle. Which means that the users do not integrate *all* data before they start analysing the data, but only integrate when required. This saves time and money, because users can experiment with the data before making the big investment of integration. (Franklin, Halevy, & Maier, 2005)

A dataspace has four distinguishing properties that set it apart from other database management solutions.

1. It deals with data in a wide variety of formats, and can be accessed from different kinds of systems. Hence it's designed to support 'all' data instead of a subset (like DBMS (Database Management Systems))
2. A DSSP (DataSpace Support Platform) is not fully in control of the data, most responsibility stays with the owner of the data.

3. The service level of a DSSP is varying, it may give 'best approximation' answers, while relational databases claim to have one version of the truth
4. A DSSP will deliver tools to create a better integration of the data when needed.

3.3 Current state of the art in data integration for Disaster response/management

There are several integration approaches and methods applied in the current Disaster Management research. We provide the reader with an overview in this section, which is based on our literature review. We found some very useful articles, but also some who are still at the spring of their development, we choose to not review them in depth (Ashish, Lickfett, Mehrotra, & Venkatasubramanian, 2009; Mescherin, Kirillov, & Klimenko, 2013; Naumann & Raschid, 2006; Xiang-hong, Ji-ping, Sheng-hua, & Yong, 2011; Ying, Daoping, Guangli, & Di, 2010).

Research from Hristidis (2010) conducts a survey of 'data management and analysis in disaster situations'. They define several roles, challenges and solutions for data management and analysis in disaster situations. We summarize the relevant part of their research for our context, which is the part on data integration. The authors describe data integration as: integration of multiple heterogeneous sources, data fusion or ingestion. Challenges are: data is both structurally and semantically heterogeneous. Examples of data include: 1. Situation reports or incident reports. 2. Damage analysis reports. 3. Geo-data and road status. 4. Logistic data, for example delivery times. 5. Messaging and communication. 6. Financial data. 7. Blogs. Next to the heterogeneous aspects, the data is also uncertain. As a solution the authors propose three main components which interact with each other. First is 'data spaces' which is a loosely integrated set of data sources where integration happens when needed. Second is organization coordination & communication middleware, which can be described as the means in which organizations communicate with each other and the rest of the stakeholders. Third is Ontology and Semantic web, which are used to create a common vocabulary to identify semantically common objects that can be used to link data together. These three components work together to create a clear and comprehensive overview of the data (Hristidis, Chen, Li, Luis, & Deng, 2010).

Secondly we found 2 frameworks (Kou et al., 2010; J. Xu, 2011) which try to develop a way to integrate all heterogeneous data sources available in a disaster context. Where Kou et al (2010) focus on a XML middleware approach, Xu (2011) proposes an approach with a central integrated database and separate converters per data source.

There is another research direction focussed at integrating GIS or spatial data (Bakillah, Mostafavi, Brodeur, & Bédard, 2007; Rishe & Yesha, 2011; Stancalie, Craciunescu, & Irimescu, 2009; Vatseva, Solakov, Tcherkezova, Simeonova, & Trifonova, 2013). Stancalie et al (2009) focus on earth observation data to monitor floods, this approach does not incorporate other data sources like operational data or needs assessments. Rishe and Yesha (2011) describe the TerraFly software which is used by governments and news agencies to visualize GIS data, it includes transformations to enhance datasets with geographical references. However, this is a very computationally expensive solution and therefore not applicable in our context. Bakillah et al (2007) propose a mapping language to support the mapping of different ontologies in the spatial data domain, but due to the low amount of GIS data we choose not to further investigate this approach. Vatseva et al (2013), propose a data integration by using a GIS system. They visualized several geo-datasets but this was mostly focussed on structured data, which is not the only data structure in our domain.

Next to the frameworks, the GIS integration and the survey we described, we found some interesting approaches we wish to share (Ashish & Mehrotra, 2010; Fahland & Quilitz, 2007; Huang, Chen, & Xiao, 2014; Kovavisaruch, Kamolvej, Prommoon, & Iamrahong, 2013; Meisner et al., 2009; Yang et al., 2012). Ashish and Mehrotra describe a community driven approach, where they develop a service for new application builders to use integration capabilities, this paper has similar thoughts as the

community and P2P approach described in section 3.2.2. Kovavisaruch et al (2013) describe a way to integrate all data from different government bodies in Thailand, this approach only focusses on structured (CSV, excel and relational) data unfortunately, and is therefore only partly applicable for our case. Huang et al (2014) describe an approach to integrate heterogeneous data by using XML, they focus only on structured and semi structured data. Next to this they also try to incorporate application integration into their solution, where our solution focussed on data integration, therefore this method is out of scope. Fahland et al (2007) propose a flexible solution that incorporates semantics to counter the heterogeneity issues. They transform the data first to schema-less RDF for maximum flexibility, and tag it with time and space for analysis. Unfortunately this system was not ready at the time of publishing and no further articles are written about it. Yang et al (2012) describe MADIS, which is a multimedia aided disaster information system for emergency management, this system analyses pictures and the descriptions to provide the disaster responder with situational awareness. There is unfortunately no integration of unstructured pdf files incorporated. At last, Meisner et al (2009) provide an overview of data integration and visualisation technologies. They conclude that high resolution imagery and digital elevation models are of utmost importance to this set of technologies, but these are unfortunately not yet available for our context.

There are also other initiatives to improve information interoperability in the humanitarian sector. The earlier mentioned Symposium on Best Practices in Humanitarian Information Exchange, mentions some useful best practices. The best practices for data products can be found in Appendix 7 (Tsui, 2002). These are mainly focussed on: Defining user needs of decision makers in the field, focus on operationally and strategically relevant themes, consult data providers and affected communities, and finally create templates for assessments. The best practices for preparedness can be found in Appendix 8. These can be summarized as follows: prepare toolboxes for distributing information, create capacity for rapid deployment, maintain quality of data even when disasters are not current, and define an exit strategy when project is closed (Tsui, 2002).

3.4 Summary

In this chapter we identified and described the evaluation frameworks we will use for the selection of data sources and integration methods. Next we described the identified integration methods from the literature, and categorised them in a 2x2 matrix to provide a clear overview. At last we described the current state of the art around data integration in the disaster management sector. In the next chapter we will describe our main artefact, which is our proposed method. Followed by the results related to research question 1 and 2, which are the disaster responder information needs and the identified disaster data sources.

4 Results

In this section the reader finds the results of our research, where we empirically validate our proposed method to select a data integration method. In section 4.1 we first share the overview of our method. The following results are divided by the three separate research questions from section 1.4. In section 4.2 the reader finds the answers to research question 1 (the Information need of disaster responders). In section 4.3 the reader finds the identified data sources, which relate to research question 2. In section 3 we share our literature research to data integration methods. In section 4.4 we analyse the identified data integration methods from section 3.

4.1 Data integration method selection approach

In Figure 23 we present the approach we developed based on the problem context as an extension to the CRISP-DM process. Parts of the approach are linked to the research questions we developed in 1.4. The 'Disaster Data Understanding' activity relates to research question 2, and to the CRISP-DM 'Data Understanding Phase'. The 'Business Understanding' phase revolves around the disaster responder information needs, which relates to research question 1. In these first two steps we determine the information needs of disaster responders, and the available disaster data in the problem environment. Then we combine the two components to finally get a list with data sources which should be integrated to cover most of the information needs. We use the results of the literature review from this thesis to decide which data integration method is best suited for this specific problem. All activities in the method are extensively described in Table 7.

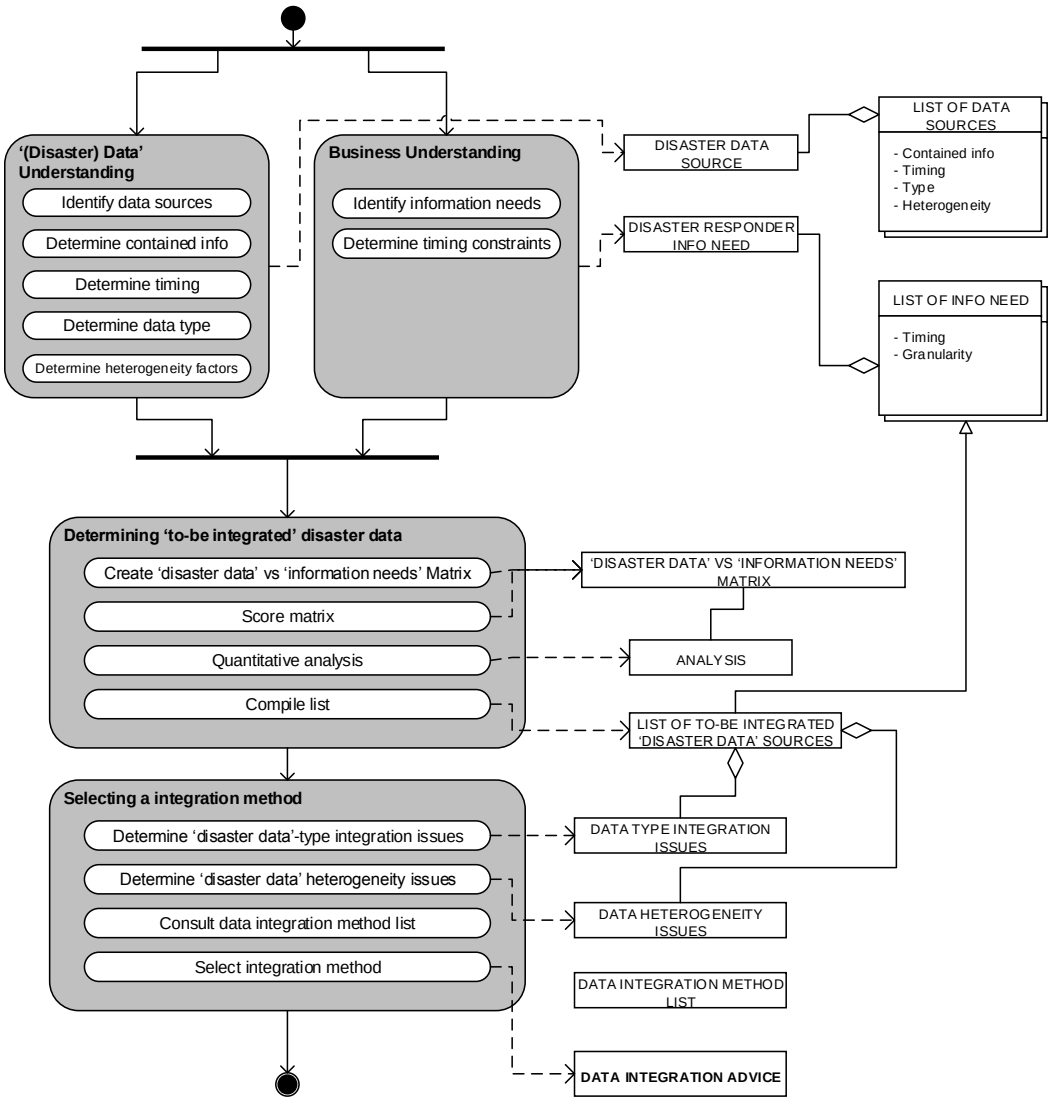


Figure 23 Proposed method for selecting an integration method

Table 7 Activity descriptions for proposed method

High Level Activity	**Activity**	**Description**
Disaster Data Understanding	Identify data sources	The team interviews stakeholders to compile a first list of disaster data sources associated with the disaster. Respondents could be: NGO employees, government disaster management authorities etc. Next to the interviews, the team performs an internet search for open data sources
	Determine contained info	The team familiarizes with the data by exploring it and writing down what information it contains.
	Determine timing	The team determines at what time the data sources come available, or are available.
	Determine data type	For every data source the data type is determined, which is required for successful integration.
	Determine heterogeneity factors	For every data source, the team determines how the heterogeneity factors could be filled in. In the activity 'Determine Disaster Data Heterogeneity Issues' the knowledge from this phase is used to determine, per pair of disaster data sources, what heterogeneity issues arise.
Business Understanding	Identify information needs	The team interviews multiple experienced disaster responders, the results lead to a list of disaster responder 'information needs'. Note: Multiple data collection approaches are possible, we used unstructured interviews combined with inductive coding. Our results can be used as starting point for future teams who want to identify the information needs in a different situation.
	Determine Timing Constraints	Every information need is marked with a timing constraint: when is the information needed in the disaster phase?
Determining 'to-be integrated' disaster data	Create 'Disaster data' vs 'information needs' Matrix	The team creates a matrix where on one axis they put the 'disaster data' and on the other axis: the 'information needs'. Also the timing constraints are put in the matrix, these are implemented so the team can see which data sources will be available 'in time' and which not.
	Score matrix	For every cell in the matrix (e.g. a combination of an individual information need and a disaster data source), the team determines if the information need is covered by the data source.
	Quantitative Analysis	Using count functions (in for example excel) the coverage of the information needs by disaster data sources can be calculated. Now we now for each data source, how much 'information needs' they cover. By using an optimization approach we can select the most efficient set of disaster data sources to cover the most information needs.
	Compile list	The results of this optimization approach lead to a list of data sources that should be integrated.
Selecting an integration method	Determine 'disaster data'-type integration issues	For every combination of disaster data sources, we determine the issues that arise from different types. For example, excel files don't easily integrate with PDF files.
	Determine 'disaster data'	There are multiple heterogeneity factors which can be found in Table 4. These are used to score every combination of data

High Level Activity	Activity	Description
	heterogeneity issues	sources with their "Heterogeneity issues". These issues need to be solved by the integration method.
	Consult data integration method list	Based on the type-issues, and heterogeneity-issues from the previous steps, we consult data integration methods to determine which one(s) could solve these issues.
	Select data integration method	Based on the consultation of the former activity, we select a data integration method applicable for our situation.

4.1.1 Relation between proposed method, CRISP-DM and Research Questions

In Figure 24 we present the relationship between our proposed method, the CRISP-DM process and the associated research questions.

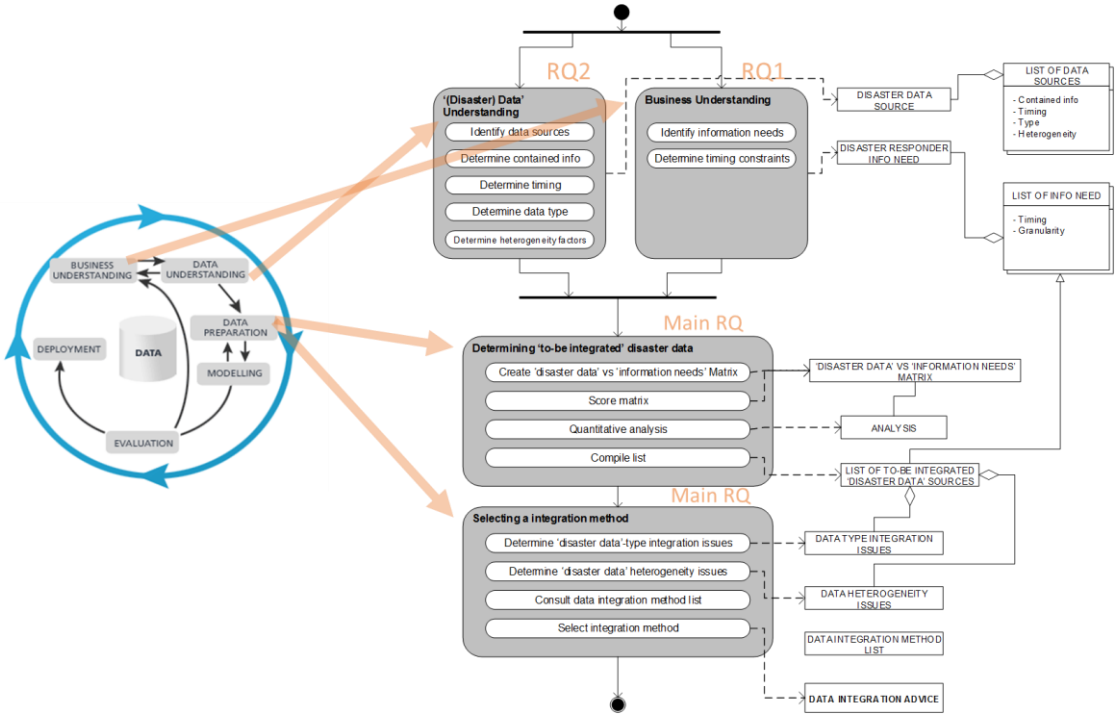


Figure 24 Relation between proposed method, CRISP-DM and Research Questions

A contribution of our study is a further specification of a phase in the CRISP-DM process. One task is titled: 'select data', we propose to use a quantitative approach to select the disaster data sources with the most added value. This approach can be found in section 6.1.

4.2 The 'Information Need' of Disaster Responders (RQ 1)

What are the information needs of the professional and responding community at national and local level?

This section presents the results gathered from our fieldtrip to Bangladesh and answers the first research question. The results are divided by into three parts: The activities of disaster responders (4.2.1), their decisions (4.2.2), and most importantly: their information needs (4.2.3).

We should interpret the numbers by keeping in mind that the interviews of actors are not evenly distributed over the actor groups. See

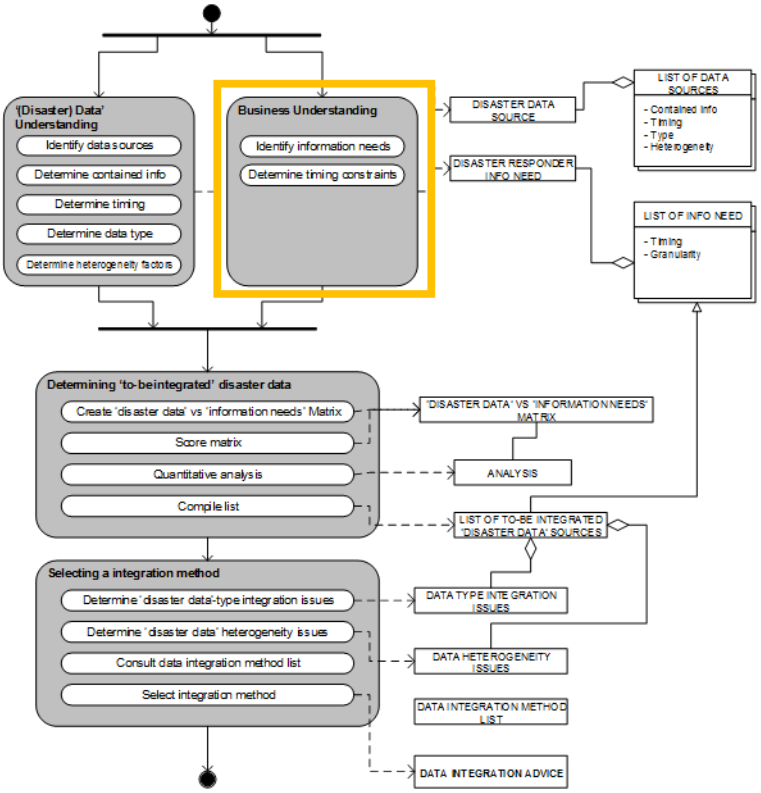


Figure 25 Relation between RQ1 and proposed method

Table 2 with the distribution of people vs transcriptions for more information. Also, since the focus groups are not transcribed literally these could yield a lower amount of themes. The researchers only had one chance to listen to the respondent and note down the themes that occurred, as compared to the 1-on-1 interviews which are recorded and transcribed literally. Also, since about 15 people participated in a focus group it automatically leaves less time to speak for an individual, therefore there is less depth in their answers. However, since this is a qualitative research we could not really analyse the results quantitatively from the start, therefore we should only keep it in mind while looking at the results, and it does not affect the quality of the output. A bit further in this thesis we analyse the results based on a cross section of the subgroups, to see whether the activities, decisions and information needs differ per group.

The "Sources" statistic relates to the amount of sources in which the (sub) theme is found. Every cluster is an aggregation of all sub-themes. In the most right column we see the term "References" These depict the total sum of tags of the subtheme, hence if the subtheme is mentioned three times in an interview, a score of 3 is added to the references column.

4.2.1 Disaster Responder Activities

Table 9 shows the list with activities emerged from the tagging of the interview transcriptions by inductive coding. We asked the respondents to their activities to increase our domain knowledge and so we can deduce their information needs from it. This exercise where the information needs emerges from the activities will be shared in paragraph 4.2.1.1.

There are 16 separate interview transcriptions used for these results. 12 of them are 1-on-1 interviews, and four are interview reports from focus groups. In the left column we see a list of activities (sub themes) mentioned in the interviews, summarized by clusters. The clustering exercise, which we described in the research methods section, resulted in eight clusters for 37 activities: Information Collection & Sharing, Recovering, Protecting livelihoods, Relief goods, Response, Search and Rescue, Surviving and Training. See Table 8 for the high level view. Next to this we have activities like: Collecting info, Gauge reading etc. These can be found in the lower level view of Table 9.

Table 8 High level clusters of Disaster Responder Activities

Activity	Sources	References
Information Collection & Sharing	11	39
Protecting livelihoods	10	32
Relief goods	10	21
Response	9	16
S&R	8	12
Training	7	11
Recovering	2	3
Surviving	2	3

We specifically asked about the preparedness and the response phase of the disaster, therefore we see a low amount of sources in which activities around the "Recovering" theme are mentioned. The amount mentioned of the 'surviving' subtheme can be explained by the low amount of interviews with locals from the responding community, specifically only 2 focus group discussions were conducted with this group. Intuitively the responding community is the only group who is actively surviving since they are the only ones affected by the disaster within our target group.

Next to these two low scores we see a pretty dense top of the table. Most of the clusters are fairly close to each other and are mentioned in at least seven of the interviews (43.75 %) to 11 of the interviews (68.75 %). When we sort the clusters, we get 'Information Collection & Sharing' as the most mentioned

cluster. We assume this is mentioned most because the researchers mention their interest in information usage around disasters, therefore the respondents have that on top of their mind when responding to the questions. The next two high scorers are 'Relief goods' and 'Protecting livelihoods'. 'Relief goods'-activities are focussed at the stocking, distribution, sharing and receiving of relief goods. 'Protecting livelihoods'-activities have subthemes like: moving cattle, raising plinth and preserving firewood and food for example. The 'Protecting livelihoods'-activities are mostly performed in the preparation to a disaster and are more or less started when the civilians know the flood is coming or when the flood season is near. The next most mentioned cluster is 'Response'-activities, this is comprised of things like: taking shelter and coordination activities. Lastly we get the clusters: 'Training' and 'Search and Rescue'. Where the first one is aimed at the training of the responders for more effective response, the latter is the process of finding displaced or people in need of rescue.

Table 9 Disaster responder activities + preparedness + recovery

Activities		Sources	References
		16	137
Information Collection & Sharing		11	39
Collecting info		10	17
Sharing info		7	13
Needs assessment		4	4
Identify need		2	2
Gauge reading		1	1
River erosion checking		1	1
Understanding need		1	1
Recovering		2	3
Provide employment		1	1
Rebuild community		1	1
Rebuilding infrastructure		1	1
Protecting livelihoods		10	32
Awareness building		5	9
Early warning		5	6
Preserving food and firewood		4	4
Creating contingency plans		3	4
Raising plinth		3	3
Harvest crops		2	2
Moving cattle		2	2
Community risk assessments		1	1
Create portable oven or stove		1	1
Relief goods		10	21
Food support		4	4
Create relief goods stock		3	4
Sharing relief goods		3	5
Provide WASH goods		2	2
Receive relief		2	3
Providing medical support		1	1
Providing non-food items		1	1
Providing relief goods		1	1
Response		9	16
Taking shelter		9	13
Coordinating		2	2

Activities		Sources	References
Security		1	1
S&R		8	12
Surviving		2	3
Managing cattle		1	1
Praying		1	1
Waiting for flood to leave		1	1
Training		7	11
Training responders		4	4
Advice on preparedness activities		3	3
Capacity building		3	3
Swim lessons		1	1

4.2.1.1 Information needs deduced from activities

We want to derive the information needs of disaster responders not only by directly asking our respondents, but also by deducing them from their activities. This increases the internal validity of our research since we will find blind spots our respondents have, and we can counter issues like no-recollection bias. We conducted a brainstorming exercise with 2 domain experts to derive the information needs from the activities. We first matched the already available information needs from our results to determine the gaps, afterwards we tried to deduce the blind spots by asking domain experts for their input.

4.2.2 Disaster Responder Decisions

What are important decisions for disaster responders in the preparedness and response phase?

Below is the list of decisions made by the disaster responders. There is a pretty sharp distinction between the three most mentioned decisions (Location of help, Type of support and Beneficiary selection) and the rest. These three are followed by two type of decisions around 'Coordinating' and 'Response' where the latter is basically the decision whether to respond or not.

Table 10 Disaster responder decisions

Decisions	Sources	References
Location of help	8	10
Type of support	8	14
Beneficiary selection	7	10
Coordinating	4	4
Response	3	3
Assessment	2	2
Location for S&R	2	2
Selecting experts	2	2
Time of response	2	2
Determining relief package	1	1
Harvesting	1	1
Shelter location	1	1
Where to buy and sell food	1	1

The results of this question are a bit disappointing as compared to the large list of information needs and activities which emerged from the interviews. The researchers assume that people find it hard to think back of the exact type of decisions they made, or they deem the decisions they make irrelevant

for the research. Or maybe the respondents had superiors making the decisions. These biases are described in the research design section. However we can assume that these are the most relevant decisions responders are making since these are on top of mind of our respondents.

4.2.3 Disaster Responder Information Needs

What are the information needs of disaster responders in the preparedness and response phase?

The last theme around which we asked questions in the interviews or focus groups was aimed at the information needs of the disaster responders. For this theme we also did a clustering exercise. We created cards with all individual information needs, without knowing how much and which clusters we would create, we grouped all information needs together based on our domain knowledge and the experience from the interviews.

The clustering lead to the following result: six clusters for 51 information needs. The high level clusters can be found in Table 11. We see two top scorers which are called 'situation overview' and 'needs'. 'Situation overview' contains sub themes like: Nr of affected population, or impacted area. The 'needs' theme is mainly focussed on the needs people in the disaster area have, this can include health, water etc. Thirdly we see an 'information need' around coordination, Subthemes in this cluster are mainly focussed at an enhanced coordination between separate NGOs and the government. This is also an observation we did when talking informally to some people. Next item on the list is the context/livelihood/baseline cluster, this is focussed on the "normal" situation of the people in the affected area. Basically, NGOs and the government want to know what people were doing before the disaster to be able to supply them with more focussed and personalized relief. The fifth cluster is about 'flood news', this is basically the information about the time of arrival, the time of inundation etc. Lastly a cluster is formed of information about the locations of specific services like: water supply, doctors etc.

Table 11 High level Responder Information Needs

	Sources	References
Situation Overview	12	39
Needs	11	23
Coordination	9	27
Context/Livelihood/Baseline	8	15
Flood news	6	9
Locations	3	10

Table 12 Low level Disaster responder Information Needs

Information Need	Sou-rces	Ref-erences	Information Need	Sou-rces	Ref-erences
Context/Livelihood/Baseline	8	15	**Needs**	11	23
Context of people	6	8	Needs of affected	9	19
Vulnerabilities	3	4	Health	2	2
Demographics	2	2	Targeting	2	2
Security plans	1	1	**Situation Overview**	12	39
Coordination	9	27	Nr of affected	5	7
Coordination	5	7	Market situation	4	5
Activity of other NGO or Gov't	3	5	Inundated area	3	4

Information Need	Sou-rces	Ref-erences	Information Need	Sou-rces	Ref-erences
Capacity	2	3	News	3	3
Community leaders	2	2	Damage to livestock	2	2
Gap	2	3	Nr of damaged houses	2	2
Presence of NGO workers	2	2	Accessibility	1	1
Burying strategies	1	1	Damage	1	1
Capacity (boats)	1	1	Destroyed houses	1	1
Relief goods stock	1	1	Hazard	1	1
Staff skills	1	1	Impacted area	1	1
Telephone numbers	1	1	Losses	1	1
Flood news	6	9	Nr of affected houses	1	1
Flood news	4	5	Nr of affected population	1	1
Flood duration	2	2	Nr of people dead	1	1
Earlier predictions	1	1	Nr of people injured	1	1
Time of inundation	1	1	People in need of rescue	1	1
Locations	3	10	River bank erosion	1	1
Shelter location	2	3	Scope	1	1
Doctor location	1	1	Situation	1	1
Food location	1	2	Sub-groups	1	1
Labour location	1	1	Submerged people	1	1
Location of drinking water	1	1			
Medicine location	1	1			
Shelter location for cattle	1	1			

The two domain expert validated and enhanced the results which gives us the following (definitive) list of information needs (Table 13). This list is used for the analysis in further parts of the research. Every italic information need is added by the 2 domain experts

Table 13 Final list of information needs

Context/Livelihood/Baseline	Flood news (continued)	Situation Overview
Context of people	Earlier predictions	Nr of affected
Vulnerabilities	Time of inundation	Market situation
Demographics	*Drainage systems*	Inundated area
Security plans	*Dry area, elevation*	News
Local governance and religion	*Temperature, humidity*	Damage to livestock
Preparedness of people	*Trend analysis*	Nr of damaged houses
Coordination	*Water quality*	Accessibility
Coordination	*Yearly disaster risk periods*	Damage
Activity of other NGO or Gov't	**Locations**	Destroyed houses
Capacity	Shelter location	Hazard
Community leaders	Doctor location	Impacted area
Gap	Food location	Losses

Presence of NGO workers	Labour location	Nr of affected houses
Burying strategies	Location of drinking water	Nr of affected population
Capacity (boats)	Medicine location	Nr of people dead
Relief goods stock	Shelter location for cattle	Nr of people injured
Staff skills	*Location of people*	People in need of rescue
Telephone numbers	*Location of displaced*	River bank erosion
Communication channels	*Location of relief goods*	Scope
Helicopter capacity	*Locations at risk for disaster*	Situation
Incidents registration	*Meeting points*	Sub-groups
Response activities of entrepreneurs or companies	*Pickup points*	Submerged people
Staff training	**Needs**	*Affected medical personnel*
Stock of emergency items	Needs of affected	*Damage to infrastructure, health facilities, public buildings*
Storage location of relief goods	Health	*Nr of people saved*
Evacuation routes	Targeting	*Displaced people*
Flood news	*Capacity of Affected*	*Road access*
Flood news	*Access to Health, Sanitation, Water and Education*	*Security*
Flood duration	*Availability of materials, like shelter and fuel*	*Telephone accessibility*
		Type of diseases

4.2.3.1 Comparison with Literature

CONTEXT AND SCOPE	0	(first weeks)		(first months)	
Scope of emergency situation	3	CAPACITY AND RESPONSE PLANNING	0	LOOKING FORWARD	0
Impact: damage to infrastructure, livelihoods, etc.	6	**Other actors capacity and response:**	0	**Recovery and reconstruction**	0
Geographic areas affected	2	**(incl. gov t, military, local community, commercial aid agencies)**	1	National development strategies	0
Assistance requirements	0	Responses of other actors (who, what, where, etc.)	2	Needs and plans for recovery	0
Affected population	0	capacity of other actors (skills, equipment, scale, etc.)	1	**Preparedness**	0
Number of affected, locations	5	**Internal capacity and response**	0	Information to collect before crisis	0
Status of affected: displaced, vulnerable, etc.	1	Internal response plan	0		
Context	1	Internal capacity, structur	2		
Local socio-economic, political context	2	**Available resources: financial, personnel, stocks, technical**	3		
Local environmental, weather, livelihoods	1	OPERATIONAL SITUATION	0		
Local community capacity, coping mechanisms	1	**Security**	0		
Public and media perception	0	Current threats	0		
Public perception, awareness, attention	0	Future threats and risks	0		
Media perception	0	**Access**	0		
Political will, donor will	0	Limits to access	1		
HUMANITARIAN NEEDS	0	Logistics capacity and structur	0		
Needs	1	**Monitoring**	0		
Number in need	3	Issues	0		
Types of needs (health, shelter, water, etc.)	7	Trends	0		
Locations of needs	3	Accomplishments	0		
Needs of sub groups: displaced, vulnerable	1	**Measuring and outputs**	0		
Priorities	0	Measurable indicators for output	0		
Geographic priorities	0	Standards	0		
Priorities across sector	0	COORDINATION AND INSTITUTIONAL STRUCTURES	0		
Within sector priorities	0	**Coordination of the respons**	0		
RESPONDER REQUIREMENTS	0	External coordination (with other actors, various levels)	1		
Basic infrastructure for responders	0	Internal coordination (with other parts of the org.)	1		
Security, access	1	**Relevant laws and policies**	0		
		External coordination (with other actors, various levels)	0		
		Internal coordination (with other parts of the org.)	0		
(First days)		(first weeks)		(first months)	
META INFORMATION	0	META INFORMATION	0	META INFORMATION	0
Information available	0	Agreement on needs	0	Preparedness information	0
Sources of information	0	Extent of assessments	0		
Accuracy, validity and information	0	Actions to improve access to information	0		

Figure 26 van de Walle framework compared to 'information needs of disaster responders'

In Figure 26 we present the comparison between a framework found in the literature (Gralla et al., 2013) and our list that emerged from our fieldtrip and analysis. We used the first list of information needs gathered in Bangladesh, so without the additional information needs from the 2 domain experts. Interesting to see is that the two are not overlapping very much. Most important overlapping concepts are: context and scope of the disaster, coordination and the humanitarian needs which are the most important factors in the earlier response. Most other information requirements are not mentioned in our interviews. The difference between our list and this framework is mainly explained by the international focus of the framework, whereas our research has a local focus. First, since the responders are mostly local and have experience with the disaster and the local environment, they have a large amount of knowledge which leads to a lower need of information (hence a lower amount of our concept 'information need'). If we compare this to the Gralla et al (2013) framework which is focussed at large disasters where mostly international disaster responders were used as respondents (who do not have any knowledge of the local environment), we see a clear difference. Secondly, the flood repeats itself almost yearly, this leads to an increase in the experience of the disaster responders and practised coping mechanisms, and therefore a lower information need.

4.2.3.2 Quotes of respondents

To enhance the strength of our results we share some quotes from our respondents. Which can be used to validate our list of information needs.

Table 14 Quotes of respondents

Cluster	Quote
Situation Overview	*"Next to that we don't know the accessibility of the areas, and where to go."* *"So you need to know whether the market is functioning and where to get the relief goods."* (NGO disaster responder)
Needs	*"Ehhh, the health information, we're really not getting the correct health information what are the […] health needs in these areas."* *"That "we need water", "we need evacuation", when we get the information, we make the decision. "* *"And most importantly what are the primary needs of the people"* (Multiple NGO disaster responders)
Coordination	*"We check what are the gaps there, is government able to cope with that. Is other agencies there, is the need already met? Of is there still need for care? And if care needs to deploy there."* (NGO disaster response coordinator)
Context/Livelihood/Baseline	*"The other important thing is: what were they doing before the disaster stroke. You need to know what their livelihood was in order to help them rebuild in your response."* (Government official)
Flood news	*"Situation monitoring (water levels) by the water development board and the FFWC"* *"We would like earlier predictions."* (NGO disaster responders)
Locations	*"Where are people hiding"* *"where is the food"* *"where is the shelter"* (Government official)

4.2.3.3 Timing of Information Needs

What are the timing constraints associated with the information needs?

After the collection of the primary data we decided that we wanted to add one more variable, specifically the timing of the information needs. We assume some information is needed at the start of a disaster, while other information might be far more useful in the remainder of the disaster. During the validation of the results with respondents we asked them if they could map the information needs based on the

timing. They came up with three phases, which are related to the phases we described earlier. The phases are: first 48 hours, after 1-2 weeks and 2 months. If we visualise the mapping performed by our respondents we get the picture of Figure 27.

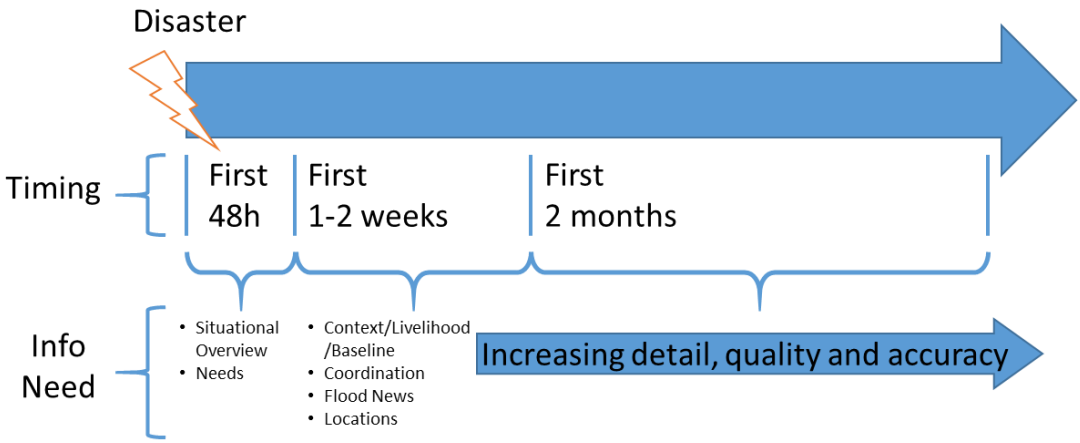


Figure 27 Timing of information needs

This mapping clearly provides us with the priorities for the first 48 hours and the phases after that. Also interesting to see is that the respondents want a lot of information at the early phases, but accept the information to not be perfect. However, they need the information to become more detailed and accurate, and they want a higher quality in the remainder of the disaster.

We have shared the results related to research question 1, next section will share the results for research question 2.

4.3 Identified Data Sources on Floods in Bangladesh (RQ 2)

What are available and relevant disaster data sources?

The following section shares the results related to our disaster data source related research question. We researched the internet for available data sources which could be used to satisfy the information needs of disaster responders. As described in the research method section we chose to focus on the 2014 flood case in (among others) the Sirajganj district. For every dataset, we determined the indicators it includes, the format in which the data is available, and the timing of publication. The flood of 2014 started around 13th of august, with heavy rainfall in India and lower lying regions of Bangladesh (HCTT, 2014a).

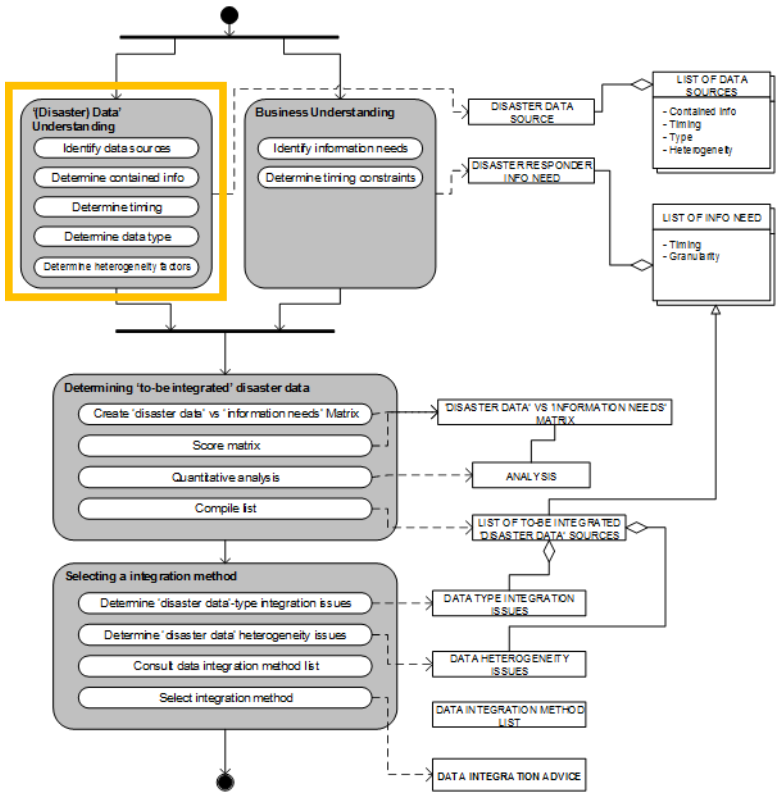


Figure 28 Relation between RQ2 and the proposed method

4.3.1 Identified disaster data sources

Which data sources are available?

Based on the interviews and the internet research (see section 2.5 for details on the research design), we compiled a list of all data sources applicable to our problem domain.

We interpreted all data sources as individual data sources, however, some have some overlap or are built from several sub-datasets. For example, the Geodash is presented as a data source, while it is in fact a geographical information sharing platform used by several departments from the Government of Bangladesh. This tool is used to make information publically available. We reviewed all information in this platform with one set of indicators as a result, for convenience reasons, otherwise we would get a very large amount of individual data sources. For the analysis of the information needs versus the data sources (section 6.1), this does not make a difference.

Table 15 Overview of disaster data sources, types, sponsors and sources

	Type of Data	Sponsor/Guardian	Source
JNA	Excel and PDF report	HCTT (Humanitarian Coordination Task Team)	NGO Assessment Team in affected area
D-Form	Excel	DMIC (Disaster Management information centre)	Government Official in affected area
Geonode WFP	Geographical layers	WFP	Multiple Government Agencies
DMIC portal - 4W DB	Relational DB	DMIC	Input from NGOs and department of disaster management
DMIC portal - Situation Reports (Inundation)	PDF	DMIC	DMIC, Department of Disaster Management and FFWC (Flood forecasting and warning centre)
District Disaster Management Plan	PDF	Government of Bangladesh (Local District Office)	Local government employees and local citizens
Secondary data assessment (ACAPS/HCTT)	PDF	HCTT	Multiple government and NGO sources.
DMIC disaster incident database	Relational DB	DMIC	Department of Disaster Management
DMIC hazard map	JPEG	DMIC	Department of Disaster Management
DMIC union fact sheets	PDF	DMIC	Department of Disaster Management
FFWC (Flood Forecasting and Warning Centre)	Website	FFWC	Gauges and field observations by FFWC employees
BBS (Bangladesh Bureau of Statistics)	Website	BBS	Multiple government sources and census data
Flood shelter list	Excel	DMIC	DMIC
National Water Resources Data	Website	Water resources planning organisation	Multiple Government Sources
News	Website	News agencies	Local reporters

Appendix 11 gives for every data source a short introduction, the contained information (indicators), the timing and the structure (pdf, web, GIS etc.). The disaster data sources are mostly accessible online, and contain by approximation 40-60 indicators per source. These indicators are used for the selection of the data sources. The three most occurring data types are: excel sheets, pdf documents and websites (html and pictures).

4.3.2 Timing of data sources

When do these data sources come available for usage?

We gathered for every data source its public availability which resulted in Figure 29. It is interesting to see that a lot of data sources are available before the disaster, these can be integrated beforehand, since the floods in Bangladesh are a regular occurrence. But we don't want to get too much ahead of things, since these results will be further used in the section around selecting a data integration method (6.2.2).

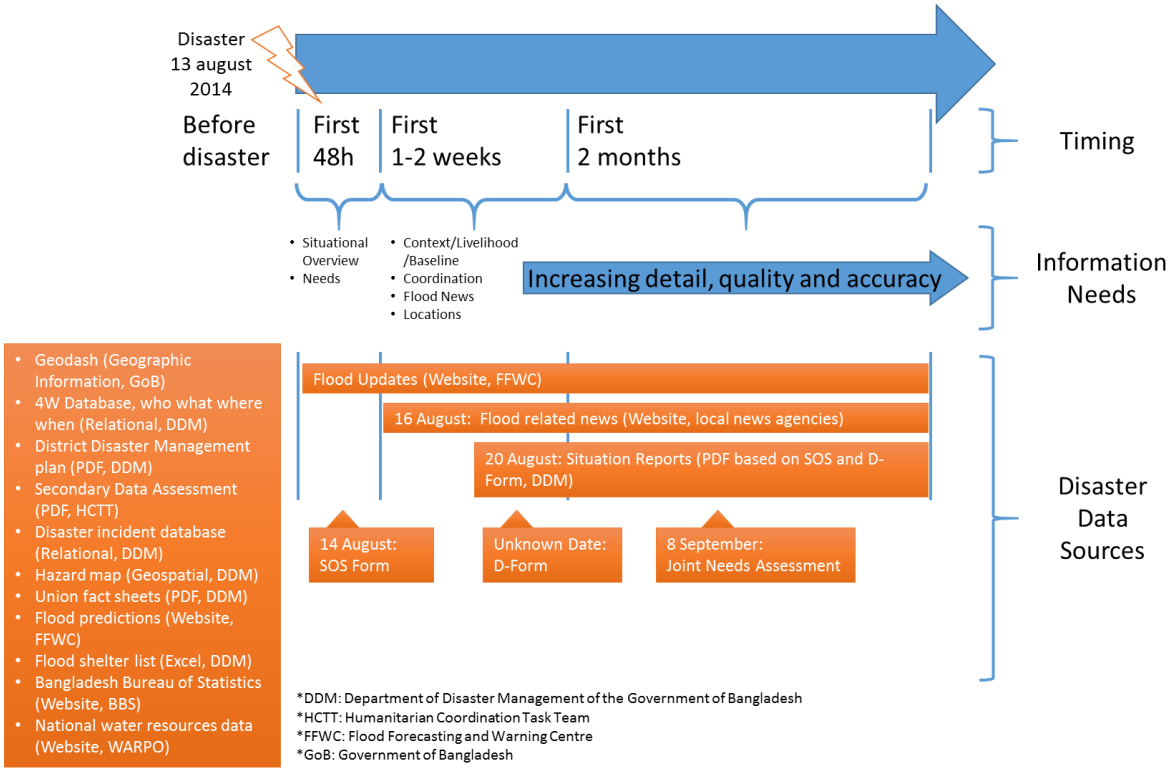


Figure 29 Timeline of data sources

We described the disaster data sources we identified, the next section is related to the data integration methods (research question 3)

4.4 Data integration method evaluation (RQ3)

In paragraph 3.2 we have described the following integration methods (for convenience we duplicated the table below):

Table 16 Overview of data integration methods

1. Generating mapping schemas	5. Peer-to-Peer data management	9. Collaborative integration	13. Extract structured information from unstructured text
2. Adaptive Query Processing	6. Data Warehousing	10. Dataspace systems	14. Integrating unstructured data into relational databases
3. XML	7. Extract, Transform and Load processes (ETL)	11. Humanitarian exchange language (HXL)	15. Ontology guided information extraction from unstructured text
4. Model management	8. Personal data integration	12. Usage of text mining algorithms	16. Unstructured information integration through data-driven similarity discovery

Every integration method is analysed by the framework we identified (section 3.1 Table 4). Since the analysis covers multiple pages, we present the results in Appendix 12 and Appendix 13. Based on this evaluation we will choose the (combination of) methods best suitable for our case study.

4.5 Summary

The first section of this chapter was our proposed method to select data and an integration method. In the following three sections we shared the results about our sub research questions (1, 2 and 3), which relate to the information need of disaster responders, the disaster data sources and the data integration methods. In the next chapter we present some not anticipated results from our research. In chapter 6 we continue with the anticipated results, where we validate our proposed method in an experiment.

5 Researcher's observations from field visit

There are some observations and challenges which were not directly targeted as a research goal in the methodology, however these observations came up during the interviews or were directly observed by the researchers. Some of these things we cannot directly reference due to privacy issues or since they were mentioned in an informal discussion with the researchers. However we still like to share them because they give the reader a thorough view of the problem domain.

5.1 Information usage at the grass-root level

When we asked the grass-root affected about their information needs, the researchers mostly did not get a satisfactory response. We assume they are not familiar with the concept of information. Asking provocative questions or giving some examples did not work, people hardly understood what we were looking for. We tried to get to the answer from a different angle, for example: "how are you deciding where you steer your search and rescue boat? And how do you know where people are in need of help?" The answer we got was pretty interesting: "we just go to the areas where we think people are in need, because these areas are lower than us". The critical thing to observe here is that people base their decisions merely on direct observation within their line of sight, and on previous experiences, they do not use actual operational information for their response activities. This can be caused by their information usage capabilities, or by the fact that not much information is available to them. This was earlier described at the introduction as an information gap. There is information available, but we need to figure out an effective way to share these back to the grass-root affected.

A challenge when designing a tool for a diverse set of stakeholders is obviously their differences in capabilities, interest and scope. However, in this specific domain also the literacy of people is an issue. A large portion of the Bangladesh population is illiterate. Next to the illiteracy most people on the grass-root affected areas are only capable of speaking Bangla (the country's main language) and cannot understand English.

Measurements are also an issue, where the Europeans have the metric system, and the United States of America use the Imperial system, people on the grass root level in Bangladesh have a different kind of measurement. Therefore we cannot share information based on meters or centimetres for example. One of our respondents told us that they use a "bigot", which is the length of a human hand in a specific position.

5.2 Data quality and Authenticity

Multiple actors shared some informal and 'off the record' comments about the data quality and authenticity issues occurring in this domain. Some actors from the NGO sector think the Government is lowering the damage-numbers and the amount of affected for political reasons. On the other hand, the Government thinks the NGOs are exaggerating the numbers to be able to provide their help. The truth is probably in the middle, but we can draw the conclusion that there is not full trust between these two actors. We should take this into account when developing a solution in this problem domain.

Next to this dispute, the quality of the data is questioned. For example, if you ask somebody on the ground "how many schools are situated in your area?". You could get multiple different answers like: the amount of government schools, the amount of private schools, the amount of pre-schools etc. This is just a minor and simplified example but the quality of the questions asked to the affected is of vital importance for the resulting quality of the data.

5.3 Government vs NGOs

In Bangladesh a disaster is only a recognised disaster if the government announces that a disaster has occurred. Only then most government departments come into action. It is prohibited for individual NGOs to provide ad hoc disaster relief without permission of the government (NGO Affairs Buro). This creates an additional barrier before support can be given, and decreases the agility of the response. We assume that the government has political reasons for this mechanism, they do not want to be seen as the country which is continuously under the pressure of disasters.

Next to this we heard some informal comments that the system is corruptible. Every government Union or Upazila Chairman is in charge of the distribution of relief goods (like rice etc.). They could decide to serve their loved ones first instead of distributing it evenly among the civilians. We heard once that a lot of the goods do not reach the designated destination. This is however a subject which is completely out of scope for the current research.

5.4 Beneficiary selection

Another theme that frequently appeared in our discussion was the concept of beneficiary selection. This is the process within NGOs that selects which people should be incorporated in their programme. The normal process is first getting an understanding of the context of the possible beneficiaries, by doing assessments and field visits. However this process gets under a lot more pressure in a disaster context. On the one hand the NGOs cannot easily assess the beneficiaries anymore due to the disaster and the resulting challenges like accessibility etc. On the other hand that the earlier assessments might not be true anymore. The situation of their beneficiaries could be worsened, but also other people's situation could be worsened.

5.5 Duplicated data collection

From an information perspective the floods in Bangladesh can be seen in different time boxes. Directly after the flood a low amount of available information, then a medium amount of disparate information and finally a large amount of disparate information. We can approximately divide that in the following way:

- First 72 hours after the flood – the government and every individual NGO are collecting information via their private channels. Government calls from the top down to their informants for information about the scope and the needs in the disaster area. This is done in an informal way via ad hoc situation reports and phone calls. Every individual NGO is collecting information from their informants (field staff and volunteers).
- First week after the flood – the damage and needs assessments are started and people on the ground are collecting data by filling in the respective forms from either the Government, the individual NGO or a consortium of NGOs.

- First month after the flood – NGOs, the consortiums and the Government are compiling their reports and sharing them.

As can easily be concluded from the description above, which is based on the interviews and field observations, we see that every actor in the disaster situation is collecting a lot of the same information, therefore a lot of duplication is occurring. This duplication is a waste of resources and adds to the issues most NGOs and Government responders shared about a lack of coordination between the responders. There is a large opportunity here to streamline this process and integrate all data in order to stop resource waste and improve coordination.

5.6 Data granularity loss in Government

After a disaster, the government has their own damage and needs assessment processes to collect the data they require. First we need to understand the structure of their government. On the lowest level we have the villages, above that we have wards (which consist of multiple villages). Unions are a level higher above wards, after which the level of Upazila emerges. Districts are the highest level under the seven Divisions of Bangladesh. A graphical explanation can be found in 'Figure 30 Government Structure in Bangladesh'. The data collection for the government lies within the responsibilities of the PIO (Project Implementation Officer), which operates at the Upazila level. They collect the data from every Union in a specific format (D-Form, see Appendix 5 for more information), aggregate it, and send that to the district office.

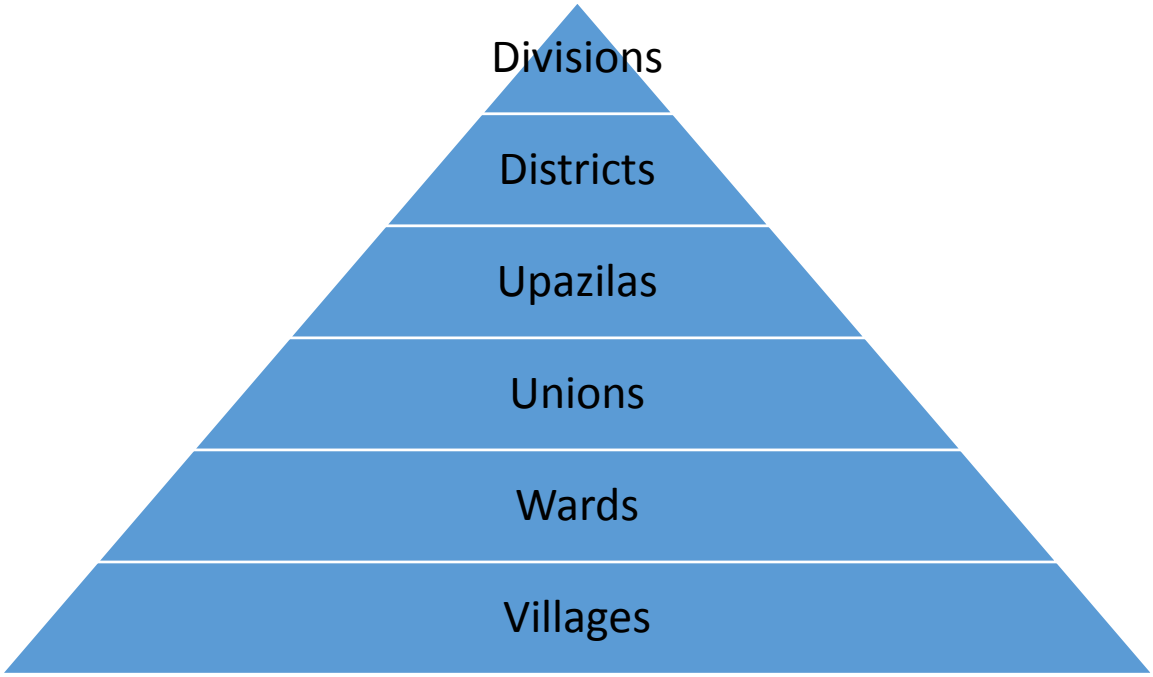


Figure 30 Government Structure in Bangladesh

The district office collects all data from all underlying Upazilas aggregates them and sends them to the division level. The PIO on the Upazila level collects the information by phone from his contact (like the Union Chairman), who their selves call around in their villages to gather the requested information. As can be perceived from this short process description, a lot of data, and therefore granularity is lost during this process. The researchers think the data can be more effective if all data is kept, instead of aggregating it on every different level of the government. Next to this, the data collection is fairly closed, only at the end of the process the government might release a report around the flooding. The researchers feel that if the government would make all data available online when collected, the other actors in disaster response could use it earlier. This way less duplication in information collection is achieved.

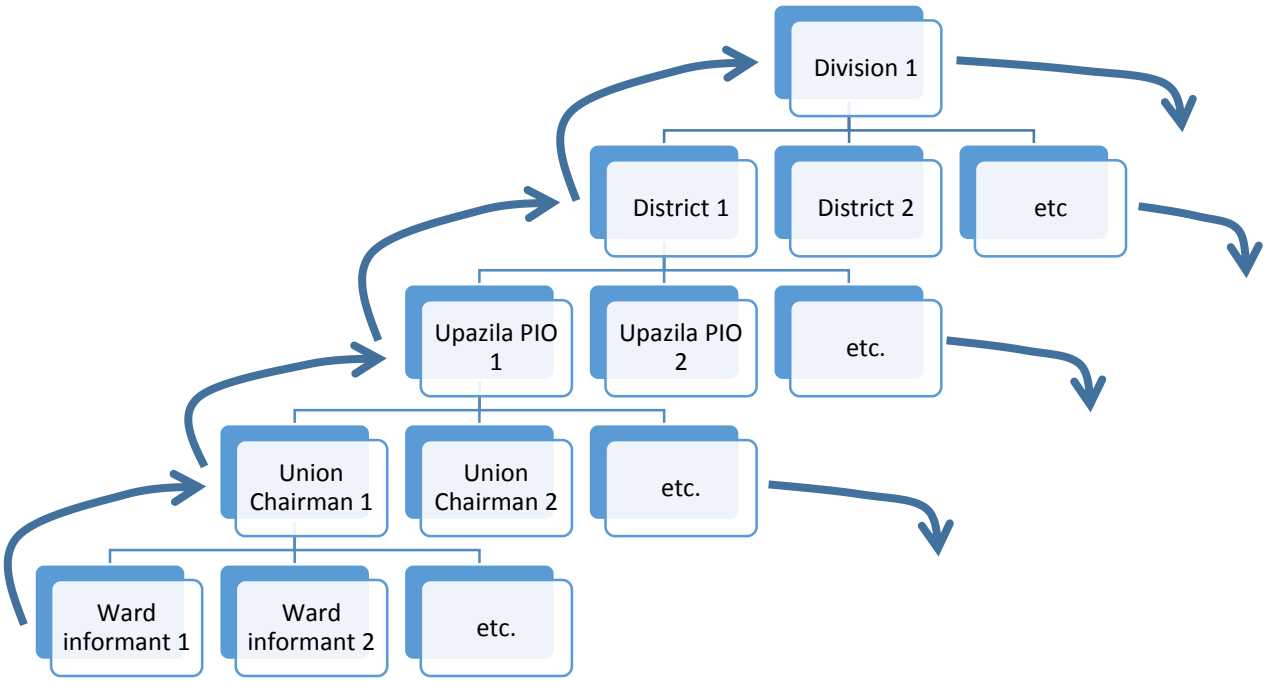


Figure 31 Data granularity loss in Government

5.7 Summary

We shared 6 interesting observations from our field visit, which were not anticipated upon and don't directly relate to a research question. These observations are interesting nonetheless, so we chose to give them a separate chapter to share them. Next chapter will provide the answer to our main research question: which data sources to integrate, and how to do thing?

6 Determining 'to-be integrated' disaster data sources and selecting an Integration Method (Main research question)

To provide Cordaid and TNO with a clear approach to integrate the data in the disaster context we follow two high-level steps. First we determine which information needs are covered by which data sources (section 6.1), then we determine which data sources should be integrated to get the best results (section 6.2). These steps are related to our proposed method.

6.1 Determining 'to-be integrated' disaster data sources

6.1.1 Create and Score the Information Needs vs Disaster Data Matrix

To analyse which data sources should be integrated we first need to determine which data sources cover which information needs. Table 17 is used for the analysis, on the Y axis of the table we placed all information needs, and on the X axis of the table we placed every data source. For every information needs we determined whether it was mentioned in one of the data sources. Table 17 represents a trimmed down version of the matrix. If required we can share the total table (including analysis and formulae).

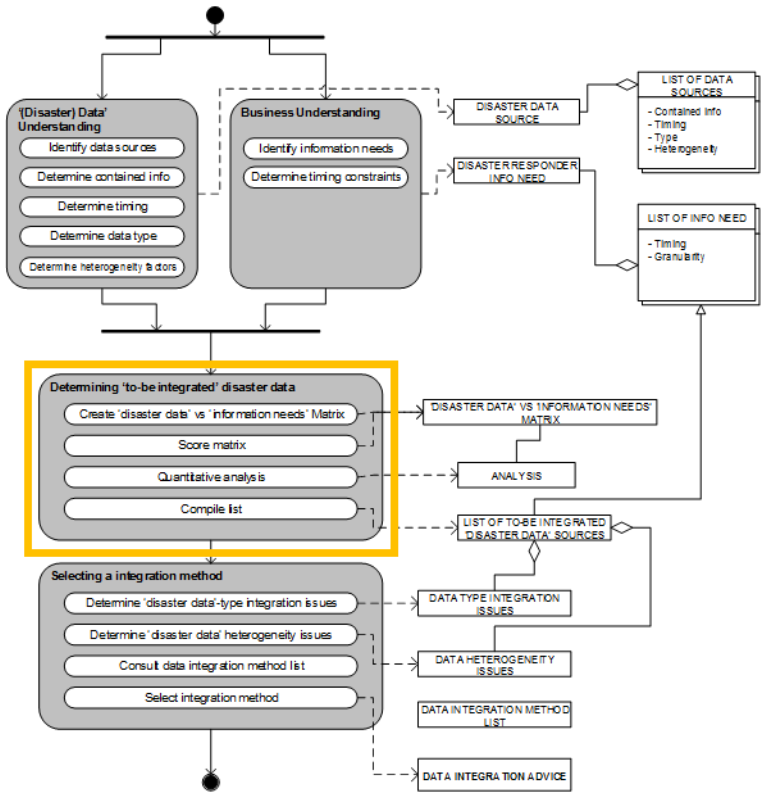


Figure 32 relation between section, main research question and proposed method

Every cell in this matrix is scored with: Yes, No or partly, which depicts the coverage of the information needs by the data source. The division of total and partial coverage by the data sources is a tricky one, since it could be tempered by subjectivity. However, we rated the matrix to the best of our knowledge. It should be noted that a 100% coverage of the information needs is almost never accomplished, we should take this into consideration when analysing these numbers.

The *coverage of information needs by the disaster data sources* is calculated by counting the amount of "yes" or "partly" statements, and dividing that with the total amount of information needs.

Table 17 Information Need vs Data Source Matrix

		Data Sources														
		JNA	**D-Form**	**Geonode WFP**	**DMIC portal - 4W DB**	**DMIC portal - Situation Reports (Inundation)**	**District Disaster Management Plan**	**Secondary data assessment (ACAPS/HCTT)**	**DMIC disaster incident database**	**DMIC hazard map**	**DMIC union fact sheets**	**FFWC (Flood Forecasting and Warning Centre)**	**BBS (Bangladesh Bureau of Statistics)**	**Flood shelter list**	**National Water Resources Data**	**News**
Information Needs	**Context/Livelihood/Baseline**															
	Context of people	No	Yes	Partly	No	No	Yes	Yes	No	No	Yes	No	Partly	No	Yes	No
	Vulnerabilities	No	No	Partly	No	No	Yes	Yes	No	No	No	No	No	No	No	No
	Demographics	Partly	Partly	Partly	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No
	Security plans	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No
	Local governance and religion	No	No	No	No	No	Yes	Yes	No	No	Yes	No	No	No	No	No
	Preparedness of people	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
	Coordination															
	Coordination	Partly	No	No	Partly	Partly	Yes	No	No	No	No	No	No	No	No	No
	Activity of other NGO or Gov't	No	No	No	Partly	Partly	Yes	No	No	No	No	No	No	No	No	No
	Capacity	Yes	No	No	Partly	No	No	No	No	No	No	No	No	No	No	No
	Community leaders	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No
	Gap	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Presence of NGO workers	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
	Burying strategies	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Capacity (boats)	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Relief goods stock	Partly	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Staff skills	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Telephone numbers	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No
	Communication channels	No	No	No	No	No	Partly	No	No	No	No	No	No	No	No	No
	Helicopter capacity	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Incidents registration	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Response activities of entrepreneurs or companies	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Staff training	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Stock of emergency items	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Storage location of relief goods	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

	Evacuation routes	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
	Flood news															
	Flood news	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes
	Flood duration	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes
	Earlier predictions	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No
	Time of inundation	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes
	Drainage systems	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Dry area, elevation	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Yes	No
	Temperature, humidity	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Trend analysis	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No
	Water quality	No	No	No	No	No	No	No	No	No	No	No	No	No	Partly	No
	Yearly disaster risk periods	No	No	No	No	No	Yes	Yes	No	Yes	No	Yes	No	No	No	No
	Locations															
	Shelter location	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No
	Doctor location	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Food location	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Labour location	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Location of drinking water	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Medicine location	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Shelter location for cattle	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Location of people	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Location of displaced	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Location of relief goods	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Locations at risk for disaster	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
	Meeting points	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Pickup points	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Needs															
	Needs of affected	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Partly
	Health	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Targeting	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Capacity of Affected	Yes	No	No	Partly	No	Yes	No	No	No	No	No	No	Partly	No	No
	Access to Health, Sanitation, Water and Education	Yes	Yes	No	No	No	Partly	No	No	No	No	No	No	No	No	No
	Availability of materials, like shelter and fuel	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Situation Overview															
	Nr of affected	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
	Market situation	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Inundated area	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
	News	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
	Damage to livestock	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Nr of damaged houses	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Accessibility	Yes	Yes	Partly	No	No	No	No	No	No	No	No	No	No	Yes	No
	Damage	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No

	Destroyed houses	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Hazard	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Impacted area	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
	Losses	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes
	Nr of affected houses	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Nr of affected population	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
	Nr of people dead	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes
	Nr of people injured	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	People in need of rescue	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	River bank erosion	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Scope	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No
	Situation	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No
	Sub-groups	Partly	Partly	No	No	No	No	Partly	No	No	No	No	No	No	No	No
	Submerged people	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes
	Affected medical personnel	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Damage to infrastructure, health facilities, public buildings	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Nr of people saved	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Displaced people	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Road access	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Security	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Telephone accessibility	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
	Type of diseases	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No

6.1.2 Quantitative Analysis of Information Needs vs Disaster Data Matrix

We first analyse the matrix on a high level (section 6.1.2.1), then we determine which information needs are not covered by performing a gap analysis (section 6.1.2.2). At last we select the most valuable and efficient set of disaster data sources to cover the information needs. We try several approaches to get to the result (section 6.1.2.3)

6.1.2.1 Overview Analysis of Data vs Info need matrix

To analyse the Disaster Data vs Information Needs matrix we counted for every data sources the amount of information needs it satisfies, to determine a percentage for the 'total coverage of information needs by the data source'. There is a total of 84 information needs, and 15 data sources. We scored every data source and every information need with a time constraint, obtained from: "Figure 27 Timing of information needs" and "Figure 29 Timeline of data sources". So all data sources have one of the 4 phases associated with them, and the information needs have one of the 4 phases associated with them. We use these phases as a constraint, if these don't match, the information need is not covered. This way we could analyse whether the timing constraints of our stakeholders are satisfied. The results of this analysis can be found in Table 18 is also visualized in Figure 33.

As we expected no data source alone answers to all information needs. Other interesting observations are the high amount of 0% scores, one data sources even manages to fulfil no information needs at all (DMIC disaster incident database). We expect this has to do with the information contained in this source, since it is very historically oriented and on a very high level. If we compare the rows with and

without time constraints we see clear differences. So the timing of the information needs is a very clear bottleneck according to our data.

High performers in total coverage without time constraints are the JNA and D-form, unfortunately they lose their high scores when we incorporate the timing constraints. In the JNA case this is because much of the covered information needs are required at the start of the disaster, while the JNA is only published almost a month after the start of the disaster. In the case of the D-form, the government does not share the dataset it gets from the assessments based on the d-form, the results are only shared via the situation reports (which is a very trimmed down version), and therefore we see 0% scores in the column with timing constraints.

Other high performers are the situation reports (14%), District Disaster Management Plan (18%) and the (online) News (13%). Especially the District Disaster Management Plan which also performs very well with time constraints (18%).

But as we can clearly see none of the data sources fulfils a critical mass of information needs to justify a 100% focus on one source. We need to integrate these sources to get a coherent picture of the total disaster situation.

Table 18 Data source's coverage of information needs

	Time constraints		**No time constraints**	
	Total coverage	Total coverage	Partly coverage	Partly Coverage
JNA	0%	35%	5%	0%
D-Form	0%	31%	2%	0%
Geonode WFP	0%	0%	5%	5%
DMIC portal - 4W DB	0%	0%	5%	5%
DMIC portal - Situation Reports (Inundation)	6%	14%	2%	2%
District Disaster Management Plan	18%	18%	2%	2%
Secondary data assessment (ACAPS/HCTT)	7%	7%	1%	1%
DMIC disaster incident database	0%	0%	0%	0%
DMIC hazard map	2%	2%	0%	0%
DMIC union fact sheets	4%	4%	0%	0%
FFWC (Flood Forecasting and Warning Centre)	7%	7%	0%	0%
BBS (Bangladesh Bureau of Statistics)	1%	1%	1%	1%
Flood shelter list	1%	1%	1%	1%
National Water Resources Data	5%	5%	1%	1%
News	4%	13%	1%	0%

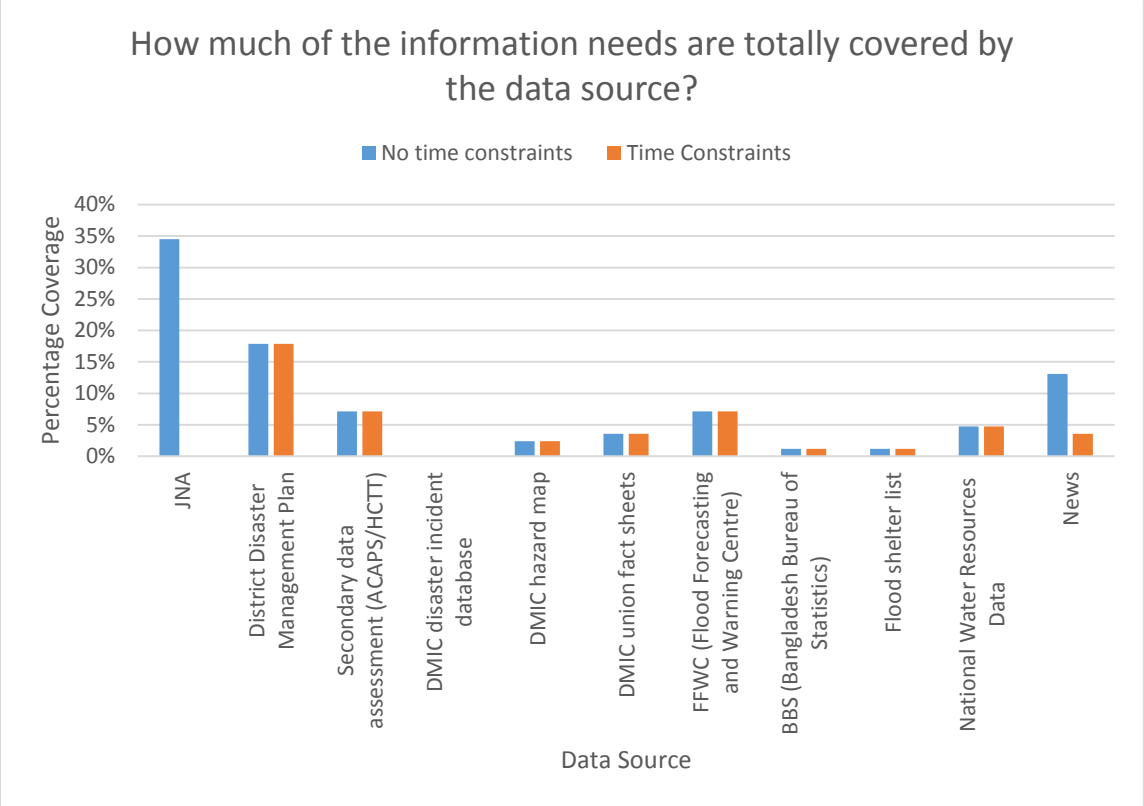


Figure 33 Data source's total coverage of information needs

To get a broad picture of the fulfilment of information needs in our case we did a basic statistical analysis, these are shared in Table 19. The reader must remember there are 84 information needs in total. First we look only at the information needs which are *totally* covered. We counted the amount of information needs which are *not* covered *in* time (61) and the amount which are *not* covered *regardless* of time (31). A simple calculations tells us that only 27% of all information needs is covered in time by the data sources, and that 63% is actually covered when time would not play a part.

Table 19 Information need fulfilment statistics

Timing Constraints	**Amount**	**%**	***No*** **Timing Constraints**	**Amount**	**%**
Total information needs	84		Total information needs	84	
Not covered	61	73 %	Not covered	31	37 %
Covered	23	27 %	Covered	53	63 %

6.1.2.2 Gap analysis

What extra data is needed?

We determine which information needs are not covered by the data sources, this way we can determine the direction in which additional data should be collected. For every cluster of information needs we

created a plot which shows for every 'information need' in how many data sources it is found. We used four statistics:

1. Total coverage in data sources (which depicts how many data sources incorporate an answer to the information need). These are depicted with light orange bars.
2. Partly coverage in data sources (which depicts how many data sources have a partial answer to the information need). These are depicted with dark orange bars.
3. Total coverage in data sources given time constraints (which depicts how many data sources incorporate an answer to the information need, where we set constraints for the timing to fulfil these information needs). These are depicted with dark green bars.
4. Partly coverage in data sources given time constraints (which depicts how many data sources have a partial answer to the information need, where we set constraints for the timing to fulfil these information needs). These are depicted with light green bars.

These four statistics give us a clear overview which information needs are fulfilled and which should be fulfilled by linking or by additional data collection. For quick referencing, the most wanted colours in the graph are (from most to least wanted): Dark green, light green, dark orange, light orange.

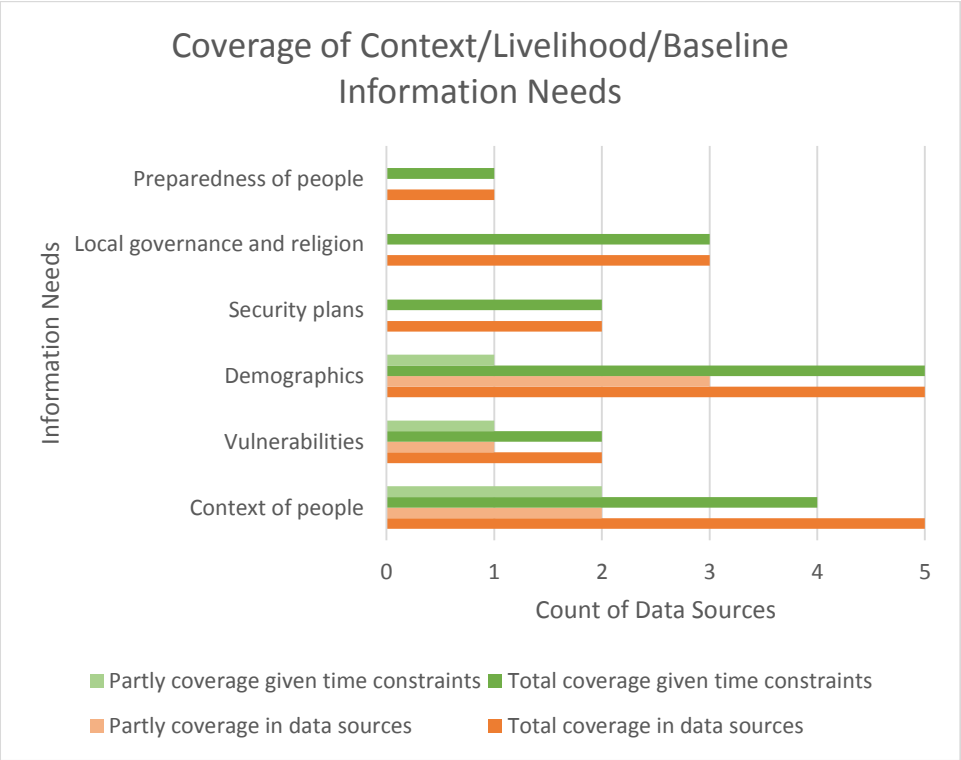


Figure 34 Coverage of context information needs

In Figure 34 we present the coverage of context related information needs. As we can see every information needs is totally covered by at least one data source, even if we incorporate the timing

constraints as mentioned before, in most cases the scores even stay the same.

Figure 35 Coverage of coordination information needs

Figure 35 depicts the coordination information needs. Interesting to observe is that not every information need is fulfilled by at least one data source. This leads us to the conclusion that we need to

collect additional data to fulfil these. There isn't much variance to the numbers when incorporating the timing constraints.

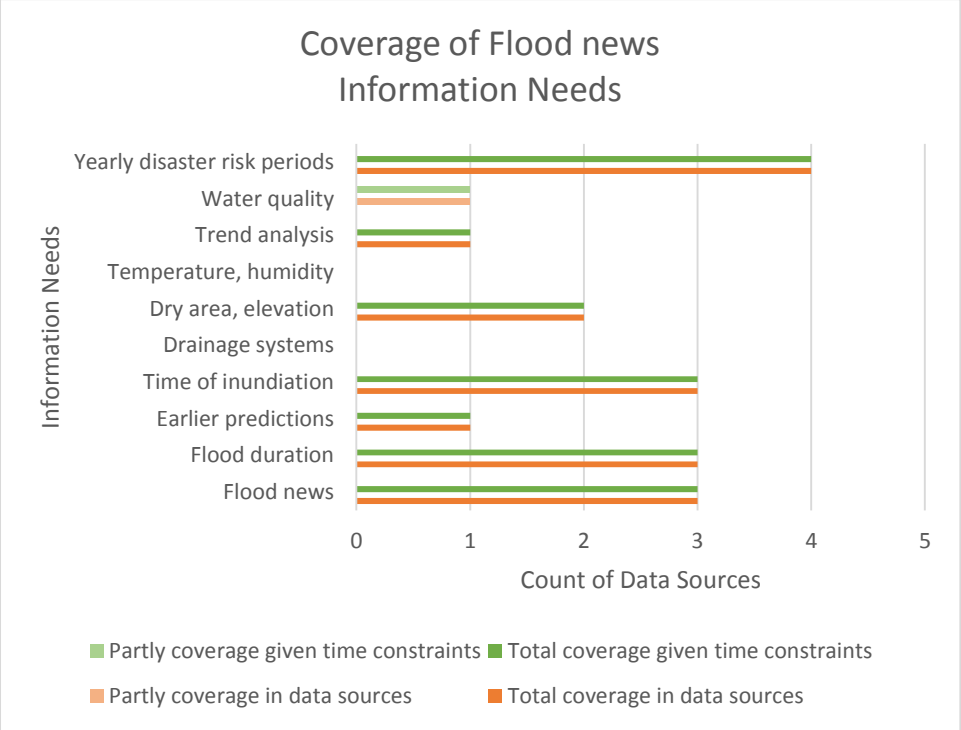


Figure 36 Coverage of flood news information needs

Figure 36 shows the coverage of flood news information needs. Interesting to see is the total absence of partial coverage statistics in this figure, which is a good indicator of a low need for additional data collection. Also the total absence of variation when incorporating time constraints shows us the data sources covering these information needs are well established. However, temperature and humidity are not covered, we did not incorporate a weather website, but if we did incorporate one this would easily match this information need. Drainage systems is the only information need which is not covered, none of the data sources provide us any information on that.

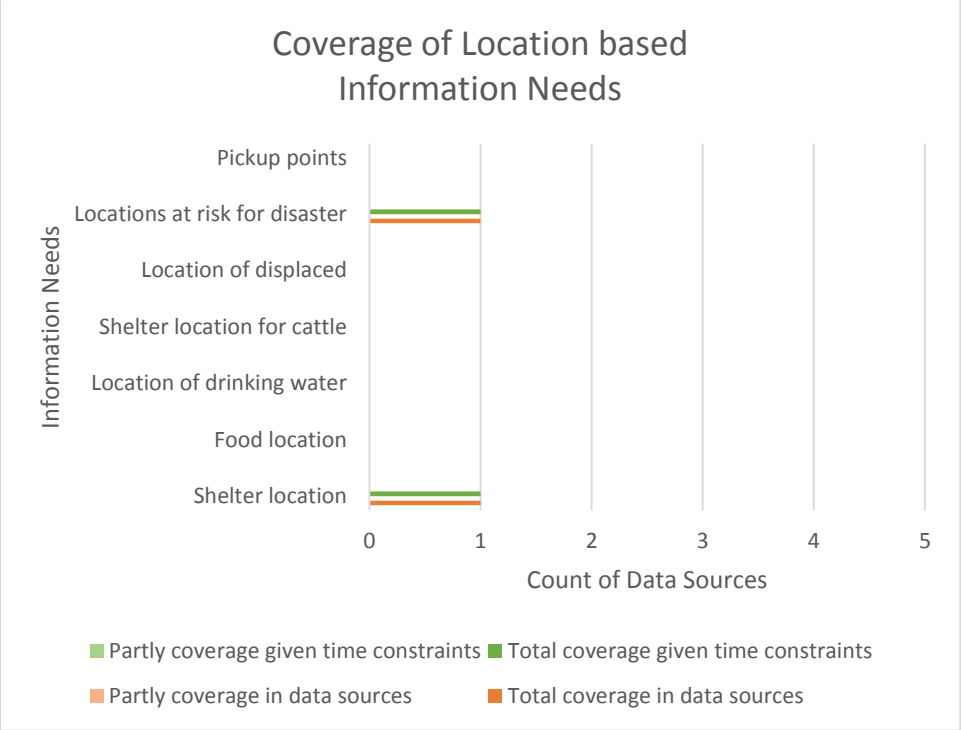


Figure 37 Coverage of location based information needs

Figure 37 shows the coverage of location related information needs. There is a big absence of location information in the disaster context, even if there is location related information, it is mostly only on a high level (like district or upazila). A lot of additional data collection should be considered here. Some of the information needs mentioned in this figure are covered by available data sources, but do not point to an exact geographic location, rather to only a certain area. There are two information needs which are actually fulfilled: Shelter locations and locations at risk for disaster.

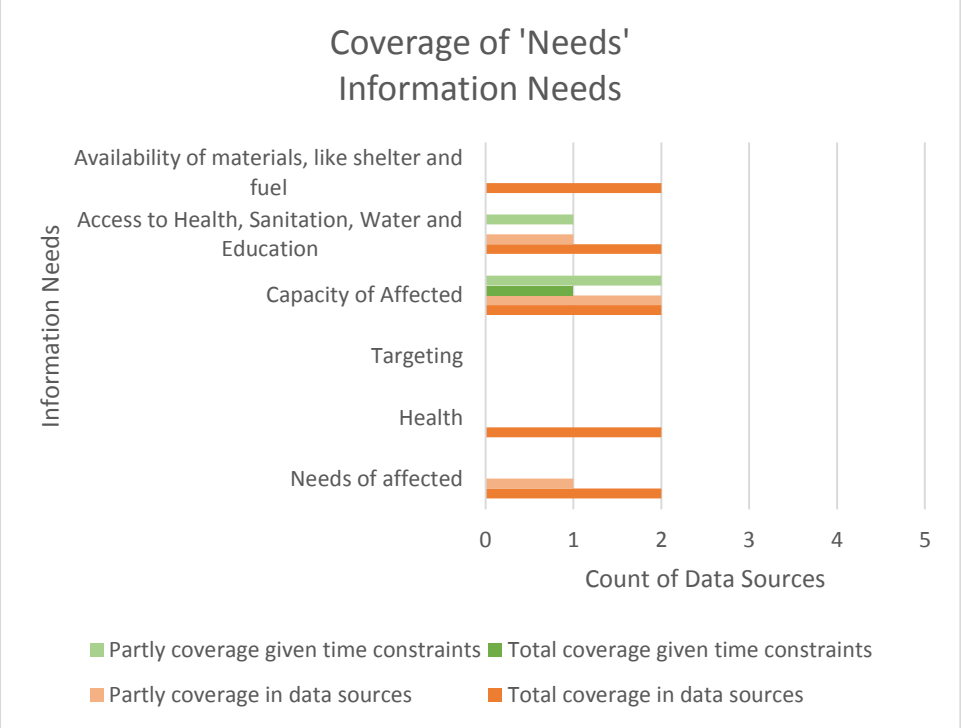


Figure 38 Coverage of 'Needs' information needs

Figure 38 shows the coverage of the information needs around the humanitarian needs during a disaster. We observe only one completely fulfilled information need (Capacity of the affected). Next to this, access to health, sanitation, water and education is partly covered by one data source. On the other hand, if we don't incorporate the time constraints we get a different picture. Most of the information needs are fulfilled in this case, this leads us to the conclusion that the information collection or sharing process should be done more rapidly.

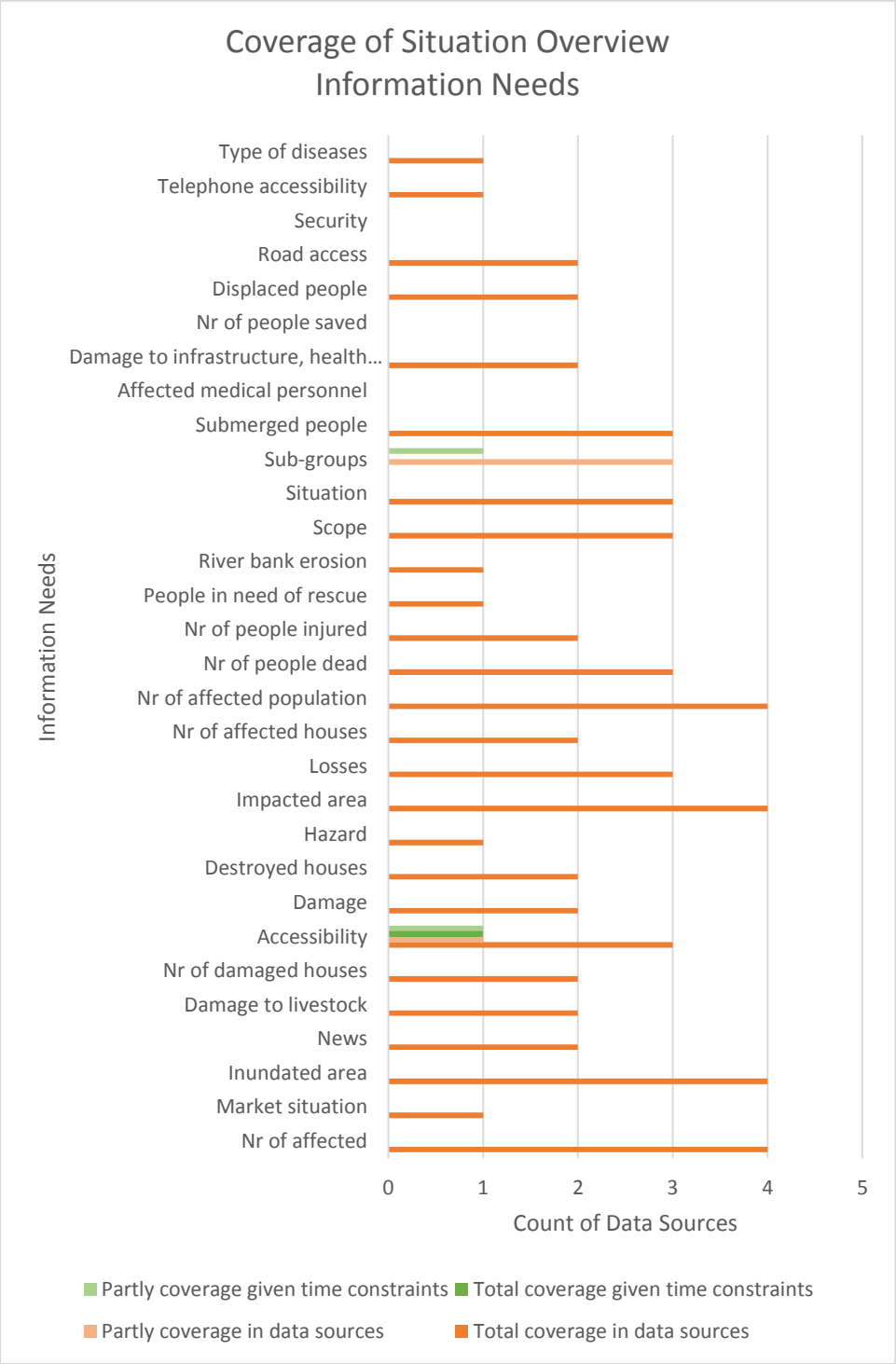


Figure 39 Coverage of situation overview information needs

Figure 39 shows us the coverage within the situational overview information needs, it shows a similar view like Figure 38, where most information is covered but unfortunately not 'in time'. This is highly related to the results we found in the academic literature regarding the difficulties in data management in a disaster context.

Below we provide a summarization of the gap analysis, by combining all (not fulfilled) information needs from the graphs above, we get two lists of information needs. One which depicts the information needs which are not covered in time (Table 20), and the other one depicts the information needs which are not covered at all. So even when timing constraints are not a factor, none of the 15 data sources contain information about these information needs (Table 21).

Table 20 Information needs not covered by data sources (given time constraints)

Coordination	**Locations**	**Situation overview (continued)**
Gap	Doctor location	Hazard
Burying strategies	Food location	Impacted area
Capacity (boats)	Labour location	Losses
Relief goods stock	Location of drinking water	Nr of affected houses
Staff skills	Medicine location	Nr of affected population
Helicopter capacity	Shelter location for cattle	Nr of people dead
Incidents registration	Location of people	Nr of people injured
Response activities of entrepreneurs or companies	Location of displaced	People in need of rescue
Staff training	Location of relief goods	River bank erosion
Stock of emergency items	Meeting points	Scope
Storage location of relief goods	Pickup points	Situation
Flood news	**Situation Overview**	Submerged people
Drainage systems	Nr of affected	Affected medical personnel
Temperature, humidity	Market situation	Damage to infrastructure, health facilities, public buildings
Needs	Inundated area	Nr of people saved
Needs of affected	News	Displaced people
Health	Damage to livestock	Road access
Targeting	Nr of damaged houses	Security
Availability of materials, like shelter and fuel	Damage	Telephone accessibility
	Destroyed houses	Type of diseases

We also listed the information needs which are not covered even when we don't apply the time constraints (Table 21). This means the following list is not found in any of the data sources. These are a high priority for additional data collection. If we look at this table we can easily see the biggest problems arise at the locations based information. Next to this coordination related information is urgent.

Table 21 Information needs not covered in the data sources (without timing constraints)

Coordination	Locations	Needs
Gap	Doctor location	Targeting
Burying strategies	Food location	**Situation Overview**
Capacity (boats)	Labour location	Affected medical personnel
Staff skills	Location of drinking water	Nr of people saved
Helicopter capacity	Medicine location	Security
Incidents registration	Shelter location for cattle	**Flood news**
Response activities of entrepreneurs or companies	Location of people	Drainage systems
Staff training	Location of displaced	Temperature, humidity
Stock of emergency items	Location of relief goods	
Storage location of relief goods	Meeting points	
	Pickup points	

We can easily see the difference between the two tables, Table 21 shows a lot less information needs as compared to Table 20. This means we could fulfil a lot of the information needs when we collect the data faster or in an earlier phase of the disaster. We also concluded this based on the graphs above.

6.1.2.3 Selecting the 'to-be integrated' data sources

After this overview analysis, we get more into depth, by determining a way to select the most efficient set of disaster data sources. We want to determine a list of data sources that should be integrated in order to maximize the amount of fulfilled information needs. However, integration of data sources is a very costly task, therefore we should not incorporate too much data sources in order to stay cost effective. We came up with several approaches:

1. "Integrate them all"- approach
2. "Sorting based on coverage"- approach
3. Optimization approach

6.1.2.3.1 "Integrate them all"- approach

The first option is obviously directly rejected since it involves a very big integration effort. However, all information needs will be fulfilled with this option.

6.1.2.3.2 "Selection based on sorting the coverage"- approach

The second option seems a good choice, therefore we analysed the matrix presented in Table 17 with some excel formulae and visualisations to determine which data sources should be integrated if we choose this method. Since the timing constraints are very important in this context we first sorted the data sources on the coverage of information needs with timing constraints. This gave us the following list:

Table 22 Sorting coverage approach

District Disaster Management Plan	17.86%
Secondary data assessment (ACAPS/HCTT)	7.14%
FFWC (Flood Forecasting and Warning Centre)	7.14%
DMIC portal - Situation Reports (Inundation)	5.95%
National Water Resources Data	4.76%
News	3.57%
DMIC union fact sheets	3.57%
DMIC hazard map	2.38%
BBS (Bangladesh Bureau of Statistics)	1.19%
Flood shelter list	1.19%
JNA	0.00%
D-Form	0.00%
Geonode WFP	0.00%
DMIC portal - 4W DB	0.00%
DMIC disaster incident database	0.00%

We could just choose to integrate all data sources that cover at least a percentage of information needs. But obviously some data and information needs are overlapping between these sources. We need to control for this, therefore we created Table 23. The data source with the highest coverage is in the leftmost column. Since the district disaster management plan is the data source with the largest coverage we start with that one. The second is the secondary data assessment from ACAPS, whilst it has the second largest coverage of information needs, it does not fill any blank spots from the district disaster management plan. The third most fulfilling data source is the FFWC, which fulfils an additional five information needs.

Table 23 Deciding on "to-be-integrated" data (Most coverage sorting approach) with timing constraints

	District Disaster Management Plan	Secondary data assessment (ACAPS/HCTT)	FFWC (Flood Forecasting and Warning Centre)	DMIC portal - Situation Reports (Inundation)	National Water Resources Data	News	DMIC union fact sheets	DMIC hazard map	BBS (Bangladesh Bureau of Statistics)	Flood shelter list	JNA	D-Form	Geonode WFP	DMIC portal - 4W DB	DMIC disaster incident database
Additional information needs fulfilled	15	0	5	0	2	0	0	0	0	1	0	0	0	0	0
Cumulative information needs fulfilled	15	15	20	20	22	22	22	22	22	23	23	23	23	23	23
% Cumulative information needs fulfilled	18%	18%	24%	24%	26%	26%	26%	26%	26%	27%	27%	27%	27%	27%	27%

We observe a low percentage of total coverage of information needs when controlling for time constraints, the maximum coverage never goes beyond 27%. This was also found when analysing the data as a whole in Table 19. So, we need to integrate the following four data sources to get the most fulfilled information needs: District Disaster Management Plan, FFWC, National Water Resources database, and the Flood Shelter List. This analysis is also visualized in Figure 40.

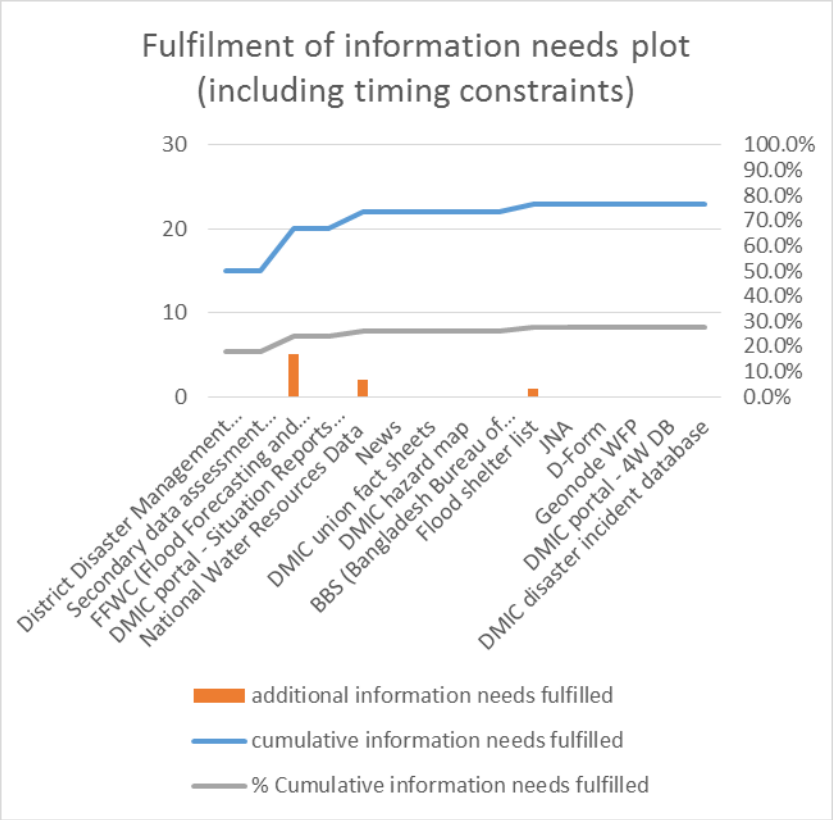


Figure 40 Fulfilment of information needs plot (including timing constraints)

We could also perform the same analysis whilst not controlling for time constraints, as sort of an envisioned or "to-be" state when all data is collected and shared in time. We conduct the same steps as described above, whilst not incorporating the timing constraints. This led to the following table with the sorted coverage of information needs.

Table 24 Sorting approach without timing constraints

JNA	34.52%
D-Form	30.95%
District Disaster Management Plan	17.86%
DMIC portal - Situation Reports (Inundation)	14.29%
News	13.10%
Secondary data assessment (ACAPS/HCTT)	7.14%
FFWC (Flood Forecasting and Warning Centre)	7.14%
National Water Resources Data	4.76%
DMIC union fact sheets	3.57%
DMIC hazard map	2.38%
BBS (Bangladesh Bureau of Statistics)	1.19%
Flood shelter list	1.19%
Geonode WFP	0.00%
DMIC portal - 4W DB	0.00%
DMIC disaster incident database	0.00%

When we calculated how many information needs every data source additionally covered, we got the following table:

Table 25 Deciding on "to-be-integrated" data (Most coverage sorting approach) without timing constraints

	JNA	D-Form	District Disaster Management Plan	DMIC portal - Situation Reports (Inundation)	News	Secondary data assessment (ACAPS/HCTT)	FFWC (Flood Forecasting and Warning Centre)	National Water Resources Data	DMIC union fact sheets	DMIC hazard map	BBS (Bangladesh Bureau of Statistics)	Flood shelter list	Geonode WFP	DMIC portal - 4W DB	DMIC disaster incident database
additional information needs fulfilled	29	3	11	4	0	0	2	1	0	0	0	1	0	0	0
cumulative information needs fulfilled	29	32	43	47	47	47	49	50	50	50	50	51	51	51	51
% Cumulative information needs fulfilled	35%	38%	51%	56%	56%	56%	58%	60%	60%	60%	60%	61%	61%	61%	61%

If we analyse Table 25 we obviously get a much higher maximum of the cumulative information needs fulfilment. If we follow this method, we could get a coverage of 61% whilst integrating the following data sources: JNA, D-Form, District Disaster Management Plan, Situation Reports, FFWC, National Water Resources Database, and the Flood Shelter List. We can also use a visualisation like Figure 41 to manually select an optimal amount of data sources.

This involves integrating seven data sources, which is a very costly operation. Therefore we must seek some kind of optimization to minimize the amount of "to-be-integrated" data sources whilst maximizing the amount of fulfilled information needs. This is the third approach we mentioned earlier.

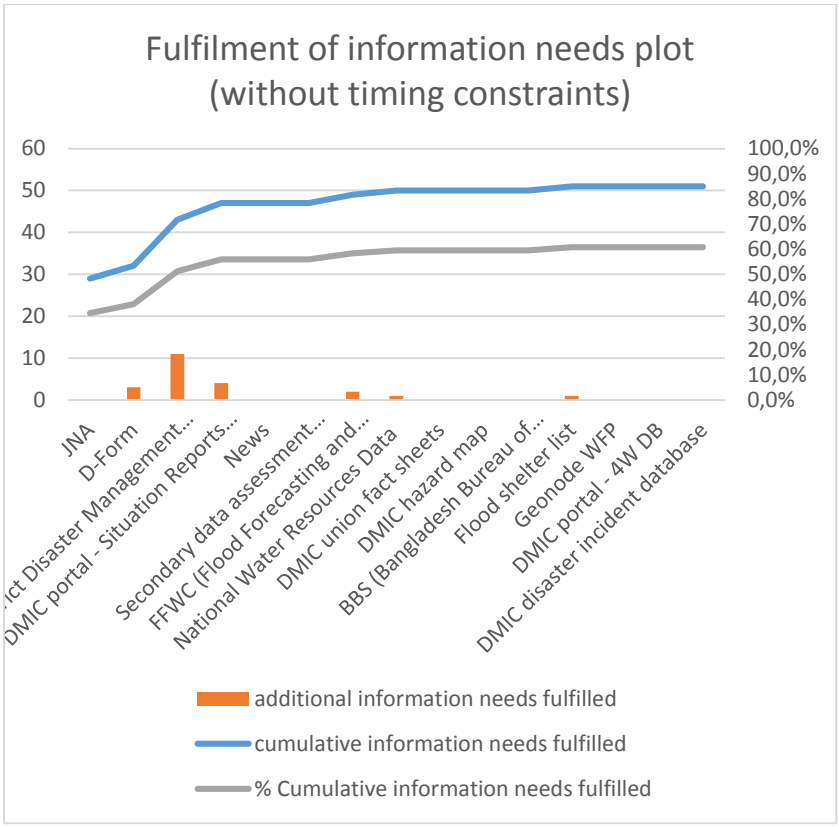


Figure 41 Fulfilment of information needs plot (excluding timing constraints)

6.1.2.3.3 Optimization approach

The goal of this optimization is to select the data sources which cover the most information needs, whilst reducing the overlap between the data sources. This way we minimize the amount of data sources required for the integration. The difference with the approach in the previous section (6.1.2.3.2) is that the current approach re-evaluates the gain in fulfilment after every iteration. We choose to do the analysis without timing constraints because it paints a clearer picture of the value we can get from integrating data sources.

This optimization starts with choosing the first data set, we choose the one with the highest coverage of information needs without timing constraints, which is the JNA data source (29 information needs). Starting from the JNA we pick the one which fills in the most information needs gaps, this is the District Disaster Management Plan (12 additional fulfilled information needs). Then we select the data sources which fills in the most gaps left by the District Disaster Management Plan *and* the JNA. This results in the selection of FFWC data source (5 additional fulfilled information needs). If we conduct one more iteration of this optimization we get the D-form as most interesting for adding to the list, this data source only delivers two additional information needs. There are other contenders for adding but these only add one additional information need. We stop the optimization here because integration of a whole data source just to fulfil one information need is not feasible. We could argue that adding only 2 additional information needs is comparable with 1 information need, so for convenience reasons we choose not to add additional data sources. In this stage, additional information could be gathered about the importance of these information needs, to decide about adding the last data source. This paragraph is summarized in Table 26.

If we compare the number of 48 fulfilled information needs by the optimization approach, with the maximum attainable by integrating all data sources (53), we see a difference of five information needs that are not fulfilled. If we compare it with the approach focussed on sorting the coverage (51) we see a difference of only three information needs whilst we only need to integrate four sources instead of

seven. It can also be argued to drop the D-form from integration so we only have to cover the cost of integrating three data sources with a resulting 46 covered information needs. This should be determined by incorporating the disaster responders opinions and see what additional value these 2 information needs yield.

6.1.3 Compiling Disaster Data list

Table 26 Summarization of results from Optimization approach

Data sources	Additional covered information needs	Cumulative	Percentage	Cumulative Percentage
JNA	29	29	34,5%	34,5%
District Disaster Management Plan	12	41	14,3%	48,8%
FFWC	5	46	6,0%	54,8%
D-Form	2	48	2,4%	57,1%

The data sources we pick to analyse based on their integration criteria, are the ones from the optimization approach, where we remove the D-form because it is not publicly available. This gives the following list: JNA, District Disaster Management Plan and FFWC.

6.2 Selecting an integration method

6.2.1 Determining 'disaster data'-type integration issues and heterogeneity issues

To get a complete picture of the situation, we need to analyse every combination of 2 data sources for integration. Analysis is done based on the criteria defined before (section Table 4). A simple calculation gives us 6 combinations (Table 27) to analyse. The direction of integration is not important, due to the requirements we set earlier, therefore we do not need to analyse both ways.

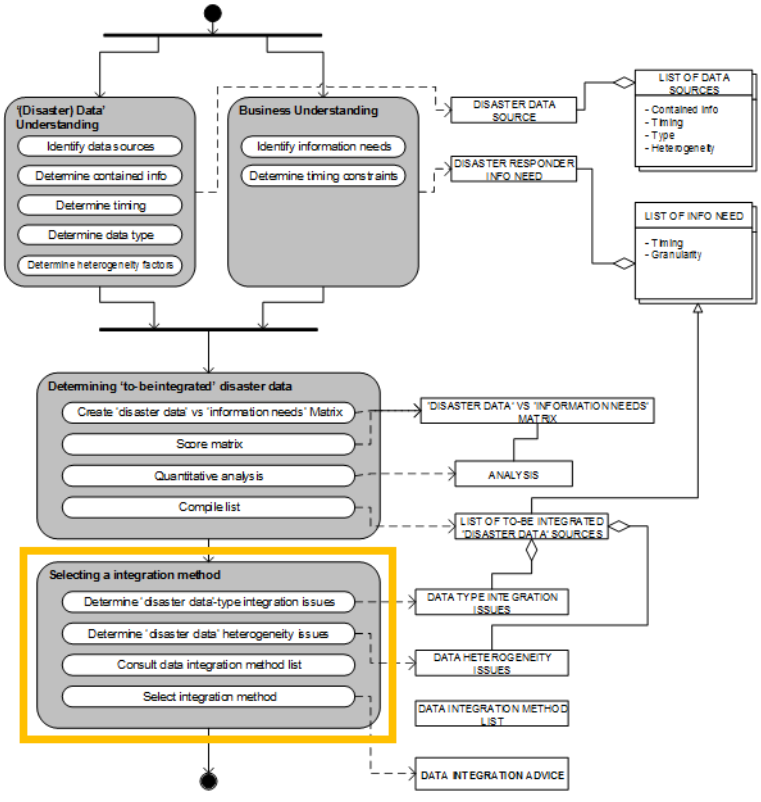


Figure 42 Relation between section, main research question and proposed method

The JNA data source consist of 2 sub sources: an excel sheet with the results of the Needs Assessment, and a written report in PDF. This raises the question whether we should divide it into 2 separate data sources. There will be a lot of overlap between the data but also information which is only mentioned in one source, since the pdf is (among others) based on the excel sheet. We choose to separate them to be sure to grab all available information in both sources.

Table 27 to-be-integrated data combinations to analyse

Combination	Data source 1	Data source 2
1	JNA excel	FFWC
2	JNA pdf	FFWC
3	JNA excel	District Disaster Management Plan
4	JNA pdf	District Disaster Management Plan
5	JNA excel	JNA pdf
6	District Disaster Management Plan	FFWC

In the tables below we present for every combination the analysis of dissimilarity between the data sources.

Table 28 Analysis of dissimilarity between JNA excel and FFWC

1. JNA excel vs FFWC		
Location of data		Both are located online on different URLs
Structure of data		JNA is a structured data base, FFWC is a website with pdf reports, jpg images and structured data
Data model		The models are not equal, due to the difference in data types.
Data schema		The only overlap between the two sources is the location of measurement/forecast/assessment in the FFWC and the JNA. The schemas are not equal, there might be record based integration possible by using the upazila names.
Hetero-geneity	Inconsistency of syntax	There is no inconsistency found, this is mainly because there is not much overlap between the sources (they focus on a totally different subject).
	Different measurement units	JNA is on Union level, and FFWC is on Upazila/district level.
	Inconsistency of representation	There is no inconsistency found, this is mainly because there is not much overlap between the sources.
	Redundancy of entities	There is not much overlap on the statistics (flood news vs damage assessment) of the entities (geographical areas), so no redundancy
	Violation of Cardinality	The JNA excel has cardinality, while the FFWC has no explicit cardinality due to its various sources, this could lead to problems.
	Semantic	Due to the low amount of overlap in the indicators there is no semantic inconsistency

Table 29 Analysis of dissimilarity between JNA PDF and FFWC

2. JNA PDF vs FFWC		
Location of data		Both are located online on different URLs
Structure of data		JNA is a pdf report, the FFWC is an online website with PDFs, JPGs and structured data.
Data model		The models are not equal, due to the difference in data types.
Data schema		The JNA has no schema due to its complete unstructured nature, therefore the schemas are not overlapping.
Hetero-geneity	Inconsistency of syntax	There is no inconsistency found, this is mainly because there is not much overlap between the sources.
	Different measurement units	Most reporting in JNA pdf is on upazila level, FFWC also reports on upazila level.
	Inconsistency of representation	No inconsistency found due to limited overlap
	Redundancy of entities	There is not much overlap on the statistics (flood news vs damage assessment) of the entities (geographical areas), so no redundancy
	Violation of Cardinality	Since the JNA PDF is unstructured it has no explicit cardinality, while the FFWC has no explicit cardinality due to its various sources.
	Semantic	Due to the low amount of overlap (the sources focus on a totally different subject) there is no semantic inconsistency

Table 30 Analysis of dissimilarity between JNA excel and District Disaster Management Plan

3. JNA excel vs District Disaster Management Plan (DDMP)		
Location of data		Both are located online on different URLs
Structure of data		JNA is a structured excel file, whilst the DDMP is a pdf report
Data model		The models are not equal, due to the difference in data types.
Data schema		The DDMP is an unstructured pdf, therefore there is no schema overlap.
Hetero-geneity	Inconsistency of syntax	Due to both sources unstructured nature there is a very high probability for a different syntax usage
	Different measurement units	Most information in DDMP is on upazila or district level, whilst the JNA is on union level
	Inconsistency of representation	High inconsistency, for example: JNA has scales for damage to crops (high-low amount of damage), whilst DDMP reports on crops number based (hectares)
	Redundancy of entities	There is not much overlap on the statistics (assessment vs pre disaster data) of the entities (geographical areas), so no redundancy
	Violation of Cardinality	The JNA excel has cardinality whilst the DDMP has no cardinality due to its unstructured format, this will lead to cardinality issues.
	Semantic	There is no predefined ontology which both sources use, so semantic issues are very likely. For example: what entails a road in both sources (does a dirt road qualify)?

Table 31 Analysis of dissimilarity between JNA PDF and District Disaster Management Plan

4. JNA pdf vs District Disaster Management Plan		
Location of data		Both are located online on different URLs
Structure of data		The JNA and the DDMP are pdf reports
Data model		Both sources don't have an explicit data model, due to the unstructured nature
Data schema		The DDMP is an unstructured pdf, therefore there is no schema overlap.
Hetero-geneity	Inconsistency of syntax	Due to both sources unstructured nature there is a very high probability for a different syntax usage
	Different measurement units	Most information in DDMP and JNA are on upazila level
	Inconsistency of representation	High inconsistency, for example: JNA has scales for damage to crops (high-low amount of damage), whilst DDMP reports on crops number based (hectares)
	Redundancy of entities	There is not much overlap on the statistics (assessment vs pre disaster data) of the entities (geographical areas), so no redundancy
	Violation of Cardinality	Both sources are unstructured, and don't have an explicit cardinality. This could lead to problems
	Semantic	There is no predefined ontology which both sources use, so semantic issues are very likely. For example: what entails a road in both sources (does a dirt road qualify)?

Table 32 Analysis of dissimilarity between JNA excel and PDF

5. JNA Excel vs. JNA PDF		
Location of data		Both are located online on different URLs
Structure of data		JNA excel is a structured excel file, whilst the JNA PDF is a pdf report
Data model		The models are not equal, due to the difference in data types.
Data schema		The JNA PDF is an unstructured pdf, therefore there is no schema overlap.
Hetero-geneity	Inconsistency of syntax	The syntax between these two sources is equal because they are compiled from the same raw data
	Different measurement units	JNA pdf is on upazila level, while the excel is on union level
	Inconsistency of representation	Representation is equal because it has the same source
	Redundancy of entities	There is a lot of redundancy because the sources describe the same indicators and are compiled from the same raw data
	Violation of Cardinality	The JNA excel has cardinality whilst the JNA PDF has no cardinality due to its unstructured format, this will lead to cardinality issues.
	Semantic	Semantic issues are non-existent because these two sources are compiled from the same raw data

Table 33 Analysis of dissimilarity between District Disaster Management Plan and FFWC

6. District Disaster Management Plan vs FFWC		
Location of data		Both are located online on different URLs
Structure of data		The DDMP is a pdf report, whilst the FFWC is a website with PDFs, JPGs and structured data.
Data model		Both sources don't have an explicit data model, due to the unstructured nature
Data schema		The DDMP is an unstructured pdf, which has no schema, therefore there is no schema overlap.
Hetero-geneity	Inconsistency of syntax	There is no inconsistency found, this is mainly because there is not much overlap between the sources.
	Different measurement units	Most reporting in de DDMP pdf is on upazila level, FFWC also reports on upazila level.
	Inconsistency of representation	No inconsistency found due to limited overlap
	Redundancy of entities	There is not much overlap on the statistics (flood news vs pre disaster data) of the entities (geographical areas), so no redundancy
	Violation of Cardinality	Since the JNA PDF is unstructured it has no explicit cardinality, while the FFWC has no explicit cardinality due to its various sources.
	Semantic	Due to the low amount of overlap (the sources focus on a totally different subject) there is no semantic inconsistency

If we combine all tables and score them on the heterogeneity per factor (high, medium or low), and determine the differences (in location, structure, schema and model) we get Table 34 as an overview.

Table 34 overview of analysis between data source dissimilarity

	Difference in:				Heterogeneity type					
	Location	Structure	Model	Schema	Syntax	Measurement	Representation	Entity redundancy	Violation of Cardinality	Semantic
1. JNA excel vs FFWC	high	high	high	high	High	medium	low	low	medium	low
2. JNA pdf vs FFWC	high	high	high	high	High	low	low	low	medium	low
3. JNA excel vs District Disaster Management Plan	high	high	high	high	high	medium	high	low	medium	high
4. JNA pdf vs District Disaster Management Plan	high	high	high	high	high	low	high	low	medium	high
5. JNA Excel vs. JNA pdf	high	high	high	high	High	medium	low	high	medium	low
6. District Disaster Management Plan vs FFWC	high	high	high	high	High	low	low	low	medium	low

The differences in location, structure, model and schema are very clear, none of the sources is only remotely alike one of the other sources. This is obviously easily explained since all sources are produced by other institutions and aren't adjusted towards each other. We need to select an integration method which can overcome these differences in structure, location, model and schema. Most schema issues are encountered because most of the data is unstructured (pdf files), which does not have an explicit data schema.

When we analyse the heterogeneity factors, we perceive some interesting things, one of which is the relative low amount of difficulties within the "entity redundancy" factor, which is only a "High" score once. This is due to the relative low overlap between the sources and therefore also low redundancy between the data. However, issues arise when we look at the syntax, representation and semantic factors, these all give problems, mostly between the JNA and the District Disaster Management Plan, because these two sources describe the same entities and the same information, but are not developed by the same organization, which leads to differences in these factors. On the highest level, the JNA and the DDMP are focussed on either the situation on the ground or the contextual factors. Where the DDMP has pre disaster data and the JNA shares a damage and needs assessment, they both describe the same entities (like the affected population, the crops, the farmers, the infrastructure etc.), this leads to a higher heterogeneity. They also describe these on a different level of abstraction, with regards to geography

(union vs upazila). On the other hand, the differences between either the JNA or DDMP versus the FFWC does not lead to much problems. This is mainly because these sources describe the same entities but on a whole different spectrum of the information needs (flood information vs a more situational overview information).

6.2.2 Select integration method

Which data integration method is the most useful? Based on the data sources which should be integrated and their integration heterogeneities and the data integration method's capabilities.

From the high level perspective we choose for a Peer 2 Peer integration method (see section 3.2 for an elaboration of all methods), because this method serves most of the requirements set within our evaluation framework. For example, the heterogeneity issues are easily solved as compared to others, because every actor integrates his data only once with a neighbour. Since the owner of the data is the most knowledgeable of the data this should also yield less effort. This high level architecture is also capable of incorporating all types of information (structured and unstructured), which is not possible with data warehouses for example. A data space also seems a very good option, however, due to the very low amount of integration, this option is not applicable for our case, since we already determined three (from the 15) sources that should be incorporated. A data space is more applicable in a context where the users don't yet know what sources should be integrated. Collaborative systems is a strong second preference, however, P2P is slightly better in the ease of integration, since the collaborative approach incorporates the knowledge of all users, where P2P uses the owners of the data (who possess the most knowledge of the data and the domain). The collaborative approach is related to the coordinated data scramble recently used in the humanitarian aid context.

On the low level perspective we have six data sources which need integration as followed from our analysis of the data source heterogeneity and information needs fulfilment (section 6.1). For every combination we propose an integration method which best suits the criteria.

There are several methods not applicable for our case due to the unstructured nature some of our sources have. The methods which we directly drop are: Generating mapping schemas, HXL, Adaptive Query processing and Model Management. We also drop XML because this method only solves the syntactic heterogeneity of the data sources.

Nr.	Combination	Integration Method	Page Nr.	Text Reference
1	**JNA excel vs FFWC**	linking text documents to structured information	44	(13)
		integrating unstructured data into relational databases	44	(14)
2	**JNA pdf vs FFWC**	Ontology guided information extraction from unstructured text	44	(15)
		Unstructured information integration through data-driven similarity discovery	44	(16)
3	**JNA excel vs District Disaster Management Plan**	linking text documents to structured information	44	(13)
		integrating unstructured data into relational databases	44	(14)
4	**JNA pdf vs District Disaster Management Plan**	Ontology guided information extraction from unstructured text	44	(15)
		Unstructured information integration through data-driven similarity discovery	44	(16)
5	**JNA Excel vs. JNA pdf**	linking text documents to structured information	44	(13)
		integrating unstructured data into relational databases	44	(14)
6	**District Disaster Management Plan vs FFWC**	Ontology guided information extraction from unstructured text	44	(15)
		Unstructured information integration through data-driven similarity discovery	44	(16)

For every combination we have two applicable integration methods. In further research we will test both of them to see which gives the most information gain.

Every combination of a structured source and an unstructured source has 2 applicable methods, we can either go for: “linking text documents to structured information”, where the results is linked text fragments to the entities in the database, or we can choose for: “integrating unstructured data into relational databases”, where the information is transformed and added to the database schema. The latter is most usable for analysis and visualisation afterwards, but could also result in difficulties for interpreting, it might be more useful for a disaster responder to just read some of the relevant text fragments.

Every combination of an unstructured with unstructured data source also results in 2 possibilities. Where “Ontology guided information extraction from unstructured text” is using an ontology the other method is solely relying on the data itself. We should test in further research whether a good ontology exists within our problem domain. We might be able to convert our information needs framework to a domain ontology, which can then be used for an information extraction. One specific question is whether the disaster management ontologies (Babitski, Probst, Hoffmann, & Oberle, 2009; Clark, Keßler, & Purohit, 2015; Liu, Brewster, & Shaw, 2013; W. Xu & Zlatanova, 2007) are applicable in our specific problem environment.

A possible result for our stakeholders is a fully integrated set of the 3 data sources, which can be used for visualisation and subsequently decision making. We envision a dashboard that combines and relates all information of the 3 sources with each other. This way the disaster responders do not need to check multiple sources and draw conclusions themselves. They would be able to see the current status of the affected (from the JNA), in the same view as the pre disaster situation (District Disaster Management

Plan). This view could help them take more informed decisions. Next to this, the retrieval of information from the large PDF documents, like the DDMP and the JNA report will be fastened. Since people focussed on a single group of affected, or on a specific can easily extract all relevant information for their own cause, without having to scan through these large documents.

7 Discussion

Below we share the discussion of our results, we divided the discussion into five sections, each relates to a section of the research. First we discuss the results around our main research question (7.1), then we discuss our results around the information needs (7.2). The last three sections (7.3, 7.4 and 7.5) are about the results around the disaster data sources, the integration methods and the challenges around the field trip.

7.1 Our approach works!

We validated our approach works. We highlight some of scientific contributions.

It saves a tremendous integration effort to take the time to optimize the amount of data sources and the coverage of information needs. For example: for a coverage of 53/84 information needs we should integrate fifteen data sources, but when we used some optimization, we only needed to integrate four for a coverage of 46/84. Furthermore, we used an extended the CRISP-DM process. Which is not often applied in the disaster management literature (only 270 hits on google scholar for the combination of the terms).

Our approach to ask disaster responders not only directly about their information needs but also about their decisions and activities definitely paid off. By not only focussing on the main goal (information need), but also by asking questions indirectly related to that goal so we could deduce the accompanying information needs, we got higher quality results. This is mainly explained by the fact that people sometimes don't recollect memories or can't express their needs because they don't know they have these needs. A clear example of this is the additional information needs we collected when validating the results with two domain experts based on the activities and the decision lists.

Evaluating the integration methods gave us a very clear picture of the applicability of the methods to our problem. The approach to just quickly gather (download) all data and conduct analysis on every individual sources is clearly sub optimal compared to our approach. When we take the time to decide which data to integrate, and which method to use we get a more optimal integration effort.

Our reusable method, which helps the user with selecting the data sources, and an applicable integration method for their specific situation, can be found in Figure 23. This approach can be used in other contexts as well. Different contexts for usage should at least entail to the following constraints: the disaster should be recurring, and the perspective of the responders should be a local/national.

A shortcoming of our research is the fact that we did not incorporate the granularity of the data, and the importance of the information needs. We decided to omit these because we did not want to make the analysis too complex. Every data source is evaluated solely on timing and contained information (indicators). We would suggest to other researchers to extend the information needs with an importance score and a difference in granularity (For example: Upazila, union or district level data). This way we could base our decisions for the optimal combination of data sources on multiple factors, which would increase the quality of the decision. However, we feel that we made an excellent first step by our proposed method.

7.2 Information needs

We have selected the respondents based on the criteria shared in the research design section. However, we did not manage to speak to all NGOs in the specified context. For example, Oxfam Novib and Practical Action are also active around flooding in Bangladesh but could not be approached for an interview. We used a snowballing approach to get in contact with our respondents, this has the advantage of getting quality respondents, but could also bias the research. However, we are confident that we talked to a representative sample of the NGOs in the Bangladesh flood context, since we talked to employees of 4 different NGOs.

The information needs we determined are on a number of points different from the information requirements as proposed by Gralla et al (2013). Our information needs have a very clear local focus, whereas the one proposed by Gralla et al (2013) has a strong international focus, which follows from the respondent list and the research design. We validated that it is very useful and relevant to determine the local information need, it saves a lot of time because we do not have to collect information which is already known by all responders. For example, we don't have to go into great depth in public and media perception information because this is not relevant for the local responder. On the other hand, we should not solely use a local framework for a big international response. Our approach would be useless for a disaster like Haiti, because for example, in this case the public and media perception is of very high importance. However, we would suggest to use a hybrid approach, by combining our local focussed framework, with the international focussed framework of Gralla et al (2013).

The recurring nature of our disaster makes our information needs also different from the Gralla-framework, but also from other literature which describes the problems around data management in a disaster context. We should keep in mind that we can collect some of the important data beforehand because we know the recurring nature of our disaster, this releases some of the time pressure as compared to an unanticipated disaster. Nonetheless we still encounter issues like the heterogeneity of data, and the issue to combine the flow of information from the ground with the stationary data we collect beforehand.

Because we collected 84 information needs, and validated it with experts, we are confident that our framework approaches an exhaustive list. Next to this it tells us that this concept is very complex and is highly variable. The variability can be explained by the preferences of individuals or by the problems people have with expressing their needs. However, we clustered the low level information needs to high level clusters, we can conclude that every cluster is very important and almost not subject to this variability.

We conducted a gap analysis to determine which information needs are not satisfied, the results can be found in Table 20 and Table 21. Interesting to see is that the influence of time constraints is very big, most operational related information is not in time. Also a lot of location based information is missing completely. Cordaid and TNO could use these lists to enhance the functionality of their app which they are currently building for the TamTam project. This app can be used to fill these gaps, this could yield very big returns. Since mobile phones could have functionalities like a GPS sensor, this could be used to get very location specific data. Information is power in the current economy, it could be a business model to monetize on this information by supplying it to the government or fellow NGOs.

7.3 Data sources

The list of identified data sources is not exhaustive, we made some selections when we found a data source online which did not have anything to do with our problem domain or where unusable for other reasons. For example, we dropped sources which only served information related to earthquakes, or contained baseline information which is not relevant to the responder in a flood context. We also dropped information sources which were completely in Bangla. In follow up research we could try to get these sources translated for incorporation in our analysis, we might even get a higher coverage of information needs.

7.4 Integration methods

A reusable artefact from our research is: the integration method evaluation frameworks, which are created from a literature review can be used as a standalone tool to evaluate multiple integration methods. Next to this, other researchers could use our literature review which identified 16 integration methods that could be applied in a disaster data setting. However, since we did not perform a full-fledged structured literature review, we cannot be confident that we produced an exhaustive list.

Nonetheless, because we carefully designed the study based on validated approaches, we can be confident about the results.

7.5 Field trip challenges

There are multiple practical challenges the researches had to overcome when interviewing the respondents.

- Focus group setting – conducting research in focus groups has some advantages, these include: people can interact and respond to each other's comments, less time is used because all participants are in the same place at once. However, during the actual focus groups we experienced some challenges: a low amount of interaction between the participants was observed, probably due to the nature and culture of the participants and the speed of the discussion was pretty slow, due to the usage of an interpreter.
- Language barrier – most people on the level of grass-root affected and lower government level did not have a sufficient English proficiency to successfully complete the interview. Therefore an interpreter was used. First off, the researchers are very grateful for the opportunity to use a skilled interpreter, but it also leads to some challenges. These challenges have possible biases as a result. For example, the interpreter knows the contents of the interview and could try to steer the respondents to an acceptable answer. Secondly it's difficult for the researcher to ask relevant follow up questions because you cannot jump into the discussion at the critical point. Thirdly, whilst trying to exactly translate the response of the respondents, in practice the interpreter is summarizing the answers of the respondent, with possible mistakes as a result.
- Culture – Bangladeshi people are very proud, friendly and open people, they perceive talking to people from Europe as an honour, and therefore it was not hard to get people for an interview. This has both upsides and downsides, if people are honoured to talk to you they are probably more inclined to provide answer they deem "wanted". Secondly since the Bangladeshi are very proud of their country they might not share too much negativity around the information situation in disaster response. Bangladesh people are not very bold and direct, therefore it was challenging to get them to speak freely from time to time. This was a complete different situation than the interviews in the country's capital Dhaka, where everybody was more bold and direct.
- Low conceptual understanding – Most people on the grass root level and the lowest government level are not highly educated. One of the most difficult things was to get the respondents to understand the concept of information. Before the interviews the researchers assumed that when asking about 'information needs', the respondents could tell a clear story about their needs. This was not the case at the lowest level, people found it hard to grasp the concept and came with unsatisfactory answers at first, but when the researchers explained what was meant by giving a few examples the respondents came up with some additional 'information needs'. However, giving examples could direct the respondents to a certain perspective or view, so this might result in a bias in the results. On the other hand, we also asked about activities and decisions, these were answered more satisfactory. We used these results to deduce the information needs.
- Travel times – the infrastructure at the field visit area was sometimes very challenging which resulted in long travel times, which resulted in a lower amount of interviews we could complete as compared to a situation where we could travel faster to different locations. However, since we used time efficient focus groups we still managed to interview a sufficient amount of people.

8 Conclusions

Our main research question entails: “How can one determine and integrate required disparate datasets to fulfil the information needs of disaster responders?”. Which leads us to our main scientific contribution: we created a reusable method (Section 4.1) to help guide our research, but it can also be applied by other researchers and practitioners in a similar context. Our method emerged from the CRISP-DM process and is also extension to the process, specifically for the “select data” and “integrate” tasks within the data preparation phase. Our method didn’t solely emerge from the CRISP-DM process, it is adjusted to the problem domain (Floods in Bangladesh) and could also be used in different related environments. Our method is applicable for local and recurring disasters. We have validated our approach by applying it in a case study around the flooding in 2014.

Next to the method, we created multiple reusable artefacts. The first one is the disaster responder information needs framework (Section 4.2), which can be applied in different disaster contexts, where the disaster is natural and recurring. The second is the identification and analysis of the disaster data sources, which can be used in the future around the Bangladesh floods (4.3). The third reusable artefact is the list of data integration methods (Section 3.2 and 4.4), which can be applied across sectors. The fourth artefacts are the evaluation frameworks for the data and the data integration methods (Section 3.1), where the data evaluation framework can be applied across sectors, whilst the data integration framework can be applied in a developing world context.

We conducted an analysis (Section 6.1.2) between the information needs and data sources. Which lead us to the conclusion that only 27% of the information needs are covered in time, whilst 63% is covered in the data sources if we do not take timing constraints into account. We also conducted a gap analysis, which resulted in a list of information needs which are not covered (Table 20 and Table 21). These lists can be used to create a new business model or new project for TNO and Cordaid.

Based on our evaluation frameworks, list of information needs, disaster data sources and integration methods, we have provided TNO and Cordaid with an answer to the main research question: “how to determine and integrate required disparate datasets to fulfil the information need of disaster responders?”. We have provided a list of “to-be integrated” data sources (Section 6.2.2) and a suggestion of integration methods to use (Section 6.1.3). This advice for specific data sources and data integration methods is directly applicable for Cordaid and TNO in their current project, they can hire a software company to implement these choices.

We also did some interesting observations in the field (section 5) which can be used for future research by TNO and Cordaid. Some examples are the difficulties in the data quality, data collection and the target audience’s conceptual understanding.

Next to the reusable artefacts and the conclusions specifically for our partners, we have multiple scientific conclusions which can be used in different contexts. Our main conclusion is: It saves a tremendous integration effort to take the time to optimize the amount of data sources (Table 26) and the coverage of information needs which is documented in our reusable method (Section 4.1). Result from our case study validate this: for a coverage of 53/84 information needs we should integrate 15 data sources, but when we used some optimization, we only needed to integrate four sources for a coverage of 46/84 information needs. Our method shows to be a useful extension to the CRISP-DM process.

A second conclusion for our specific context is the big influence of timing constraints on the information needs (Table 19). We conclude that a big proportion of the data sources is not available on time for effective usage in the field, we assume this conclusion holds for related environments.

Our third conclusion is that the local disaster responder information needs differs highly from the international perspective, this should be taken into consideration when developing information products with a local perspective. Lastly we conclude our method definitely fulfils a need by helping users select relevant data sources and an applicable integration method.

9 Recommendations

We recommend several integration methods which should be applied to solve the integration issues in our problem domain, these can be found in section 6.2.2. We will start an experiment in the future to test whether the proposed methods are effective in our case. This experiment will result in a prototype that integrates the data sources, with a complete set of integrated data. We recommend TNO and Cordaid to hire a software development company that can implement the validated methods.

With a broader perspective, we recommend Cordaid to use our general method (Figure 23) in different disaster locations. It is validated to be really helpful to identify the information needs, and the data sources beforehand. If these are identified, they can use our integration methods list to decide on an integration method.

For our case study, we recommend to keep an up-to-date inventory of the data sources. Changes and addition of data sources could lead to a different implementation of integration methods. Also keep up-to-date with data integration methods, there could surface more effective or improved ways to integrate the data sources in our case.

Timing of the data sources is of utmost importance. Focussing on an early delivery of the data sources could yield a very high amount of satisfied information needs. There are also a lot of information needs which are not fulfilled at all, TNO and Cordaid should try to fill these gaps.

When the data sources are integrated by a software company, TNO and Cordaid should keep a close eye on the probable change of the information needs. We learned from the validation exercise that seeing information needs sparks the creativity of responders. Cordaid and TNO could get reactions like: “wow, if this is possible, we would also like X information”. It is important to keep up with these suggestions, otherwise the information product they are building will become outdated and useless.

For other projects around information products we recommend to thoroughly identify the information needs of the target audience. It is tempting to just build something and distribute it, but our approach validated the time investment is definitely worth its while.

At last we would like to stress that data and information are very valuable. If TNO and Cordaid will be able to develop an information product usable in the disaster response to floods in Bangladesh, they should come up with a business model to ask fellow NGOs and government a (financial) compensation for usage of their data.

10 Future research

Our approach entails a high amount of manual labour. Examples are: gathering all indicators from every data source, determining whether a data source fulfils an information needs, determining which difficulties arise from the "to-be-integrated" data sources, determining whether an integration method solves the integration issues and fulfils the required functionality from the problem environment. We see opportunity to automate this process, maybe it is possible to automatically analyse the datasets and extract their indicators (for example by applying text mining). Afterwards we could automatically determine whether a data source fulfils an information needs by applying semantic principles. This would make the optimization approach even more effective.

We want to motivate other researchers to determine the information needs of responders in a different environment and see where the differences are. This way we could come to a more definitive list.

We will start a new research project in which we conduct an experiment to evaluate the suggested integration methods.

We have given advice on the integration of the data sources, these methods can be implemented by a software development team. But, by just integrating the data we are not at our final destination. We still need to analyse and visualise all these sources to really fulfil the information needs of disaster responders. This could be a subject for further research. In this further research we should also take into account the cognitive capabilities of people, since the interpretation of results is not equal for every individual. We also need to take into account the cooperation between different NGOs and the government, it will not be feasible to share the sources to everybody and to let everybody decide on their response. This could lead to a lot of NGOs responding in the most critically hit areas, whereas the less severe hit areas do not get the attention they deserve.

11 Acknowledgements

First of: a big thanks to the creator of the "spark" for this project, Roderick Besseling, who introduced me (after eight months of searching for the right project) to Marc van den Homberg who had the perfect project for me.

This project was funded and supported by TNO and Cordaid. Whereas TNO offered me an internship position with a monthly salary, a working space, and access to academic literature, as well as the possibility to talk to experienced researchers to steer my research. Cordaid offered me a very interesting subject to research and supported me with an allowance for survival during my trip to Bangladesh (a place to stay and nutrition). Next to this Cordaid let me use their network to find relevant interview subjects and helped me get around in Bangladesh.

Two of Cordaid's partner NGOs (Concern Universal Bangladesh and MMS) also supported me during my trip to Bangladesh, where Concern Universal supported me in getting into contact with the right people in Dhaka, and MMS helped me around in my research on the Char islands (transportation and organization of interviews and focus groups). I owe a very big thanks to both of them for helping me without any direct benefit for themselves.

Marco Spruit was my Business intelligence teacher during my Master programme. As I was heavily interested in the concept of data mining and the sheer goldmine it could be for humanity (both NGOs and Corporates) I followed his course with great pleasure. Afterwards I made an appointment with him and told him about my idea to apply these data mining techniques in the humanitarian sector. Marco was directly very supportive, while in fact he did not have any room in his hourly budget for MSc thesis support, he accepted my request for his help as my first supervisor. I know from stories of my fellow students that it is sometimes hard to find a supervisor due to their busy schedules. I am very lucky to have Marco as a supervisor, not only because we are both interested in data mining but also because we had very interesting discussions which gave the project a very nice impulse.

At last I would like to thank Marc van de Homberg, who was by first supervisor from TNO side of the project. He was the first person in the research field, who had the same vision as me around the usage of data for the "greater good". He was very inspiring to work with and a great supervisor since he has so much experience as a researcher. Not much students get a supervisor from their internship company with so much research experience, this really greatly enhanced the quality of my research.

On a more personal note I would like to thank my family and girlfriend for their support. My girlfriend was always there for me when I hit a rough patch and got me right back on track with her supporting words. Also when everything was going well she was there to make me realise the great work I've accomplished and would accomplish if I would stay on track, which made me even more motivated to bring this project to a successful end. Next to that I would like to thank my Mom for always keeping a "work-spot" and a lunch or coffee-break place available for me when I was in the neighbourhood. It is also very inspiring to have such a proud mother, this makes me even more motivated to accomplish an MSc thesis like this. At last I want to thank my father for helping me with the finishing touches of this thesis.

12 References

Abiteboul, S. (1997). Querying Semi-Structured Data. *Database Theory—ICDT'97*, *1186*, 1–18. http://doi.org/10.1007/3-540-62222-5_33

ACAPS. (2010). *Review of information needs after natural disaster – Key findings*.

Access to Information Programme. (2013). *Global e-Indices ' Rankings and Bangladesh : Indicators for Measuring Digital Bangladesh*. Retrieved from http://203.112.218.66/WebTestApplication/userfiles/Image/Other Reports/ICT4DIndicators_DigitalBD.pdf

Ackoff, R. L. (1989). From Data to Wisdom. *Journal Of Applied Systems Analysis*, *16*, 3–9. http://doi.org/citeulike-article-id:6930744

Agrawal, R., & Batra, M. (2013). A Detailed Study on Text Mining Techniques. *International Journal of Soft Computing and Engineering*, (26), 2231–2307.

Altay, N., & Green, W. G. (2006). OR/MS research in disaster operations management. *European Journal of Operational Research*, *175*(1), 475–493. http://doi.org/10.1016/j.ejor.2005.05.016

Ananthanarayanan, R., Reinwald, B., Balakrishnan, S., & Yee, Y. (2009). Unstructured information integration through data-driven similarity discovery. In *International Joint Conference on Artificial Intelligence - Workshop on Information Integration on the Web*.

Anantharangachar, R., Ramani, S., & S, R. (2013). Ontology Guided Information Extraction from Unstructured Text. *International Journal of Web & Semantic Technology*, *4*(1), 19–36. http://doi.org/10.5121/ijwest.2013.4102

Ashish, N., Lickfett, J., Mehrotra, S., & Venkatasubramanian, N. (2009). Rapid Information Integration for emergency response.

Ashish, N., & Mehrotra, S. (2010). Community Driven Data Integration for Emergency. *Proceedings of the 7th International ISCRAM*, (May), 1–6. Retrieved from http://www.iscram.org/ISCRAM2010/Papers/212-Ashish_etal.pdf

Babitski, G., Probst, F., Hoffmann, J., & Oberle, D. (2009). Ontology Design for Information Integration in Disaster Management. *Informatik 2009.*, 3120–3134. Retrieved from http://www.loria.fr/~hoffmanj/papers/ast09.pdf

Bakillah, M., Mostafavi, M. A., Brodeur, J., & Bédard, Y. (2007). Mapping between dynamic ontologies in support of geospatial data integration for disaster management. *Geomatics Solutions for Disaster Management*, 201–224. http://doi.org/10.1007/978-3-540-72108-6_14

Bangladesh Bureau of Statistics. (2011). Census data. Retrieved March 25, 2015, from http://www.bbs.gov.bd/PageWebMenuContent.aspx?MenuKey=445

Bangladesh Telecommunication Regulatory Commision. (2015). Mobile Phone Subscribers in Bangladesh January 2015. Retrieved March 25, 2015, from http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-january-2015

BBS. (2015). Bangladesh Bureau of Statistics. Retrieved July 3, 2015, from http://www.bbs.gov.bd/Home.aspx

Bergamaschi, S., Castano, S., & Vincini, M. (1999). Semantic integration of semistructured and structured data sources. *SIGMOD Rec.*, *28*(1), 54–59. http://doi.org/10.1145/309844.309897

Boufares, F., & Ben Salem, a. (2012). Heterogeneous data-integration and data quality: Overview of conflicts. *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2012*, 867–874. http://doi.org/10.1109/SETIT.2012.6482029

Bryman, A., & Bell, E. (2007). *Business Research Methods*. *Methods* (Vol. 3). http://doi.org/10.4135/9780857028044

Campbell, H. (2015). Coordinated data scrambles. Retrieved December 3, 2015, from https://docs.google.com/presentation/d/1weoUvekKWEyRE7EiX9CmHGE-9MGrRrvCTQ2Bf64fY48/edit#slide=id.p4

CDMP II. (2014). Situation report landing page. Retrieved July 2, 2015, from http://www.cdmp.org.bd/modules.php?name=Situation

Chakaravarthy, V. T., Gupta, H., Roy, P., & Mohania, M. (2006). Efficiently linking text documents with relevant structured information. In *Framework*. http://doi.org/10.1145/1014052.1014058

Clark, T., Keßler, C., & Purohit, H. (2015). Feasibility of Information Interoperability in the Humanitarian Domain. *AAAI Spring Symposium Series*.

Cordaid. (n.d.). Cordaid - About us. Retrieved July 22, 2015, from https://www.cordaid.org/en/about-us/

Cumiskey, L. (2014). *Mobile Services for Flood Early Warning in Bangladesh: Final Report*.

Cutter, S. L., Barnes, L., Berry, M., Burton, C., Evans, E., Tate, E., & Webb, J. (2008). A place-based model for understanding community resilience to natural disasters. *Global Environmental Change*, *18*(4), 598–606. http://doi.org/10.1016/j.gloenvcha.2008.07.013

the Daily Observer. (2014). Search for: Flood Sirajganj 2014. Retrieved July 28, 2015, from http://www.observerbd.com/search.php?cx=partner-pub-2946490312175476%3A8048617947&cof=FORID%3A10&ie=UTF-8&q=flood+siraganj+2014&sa=&siteurl=www.observerbd.com%2F&ref=www.google.nl%2F&ss=8j64j2

the Daily Star. (2014). Search for: Flood Sirajganj 2014. Retrieved July 28, 2015, from http://www.thedailystar.net/google/search

Dhaka Tribune. (2014). Search for: Flood Sirajganj 2014. Retrieved July 28, 2015, from http://www.dhakatribune.com/search-result?query=flood+siraganj+2014&op=Search

District Level Disaster Management Comittee Sirajganj. (2014). *District Level Disaster Management Plan Development - Sirajganj District*. Sirajganj, Bangladesh.

DMIC. (2015a). 4W database. Retrieved July 2, 2015, from http://www.dmic.org.bd/4w/

DMIC. (2015b). Disaster Hazard Maps. Retrieved July 2, 2015, from http://www.dmic.org.bd/inmap/

DMIC. (2015c). Disaster Incident Database. Retrieved July 2, 2015, from http://www.dmic.org.bd/didb/didb.php

DMIC. (2015d). *DMIC Products and Services*. Dhaka.

DMIC. (2015e). Union fact sheets. Retrieved July 2, 2015, from http://www.dmic.org.bd/factsheet/

Fahland, D., & Quilitz, B. (2007). HUODINI – Flexible Information Integration for Disaster Management. *Proceedings of the 4th International ISCRAM Conference*, (May), 1–8. Retrieved from http://www.ki.informatik.hu-berlin.de/wbi/research/publications/2007/huodini_final.pdf

FFWC. (2015). FFWC. Retrieved July 3, 2015, from www.ffwc.gov.bd

Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspaces. *ACM SIGMOD Record*. http://doi.org/10.1145/1107499.1107502

Gharehchopogh, F. S., & Khalifehlou, Z. A. (2012). Study on Information Extraction Methods from Text Mining and Natural Language Processing Perspectives. *AWERProcedia Information Technology & Computer Science*, *1*(2), 1321–1327.

Gralla, E., Goentzel, J., & Van de Walle, B. (2013). *Field-Based Decision Makers' Information Needs in*

Sudden Onset Disasters.

Grele, R. (1991). Movement Without Aim: Methodological and Theoretical Problems in Oral History. In R. Perks & A. Thomson (Eds.), *Envelopes of Sound The art of oral history* (pp. 38–52). Routledge; Psychology Press.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*. http://doi.org/10.4304/jetwi.1.1.60-76

Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data Integration Teenage Years. In *Proceedings of the 32nd international conference on Very large data bases* (pp. 9–16). VLDB endowment.

HCTT. (2014a). *JNA final report*. Dhaka. Retrieved from http://www.lcgbangladesh.org/HCTT/JNA materials update/2014 Flood Assessment/0809_NW_Flooding_JNA_FinalFINAL.pdf

HCTT. (2014b). *JNA Phase 1 – Initial Days Union Level Assessment Format*. Dhaka.

HCTT. (2015). About Humanitarian Coordination Task Team(HCTT). Retrieved July 1, 2015, from http://www.lcgbangladesh.org/HCTT.php

Hendrix, C., & Keßler, C. (2009). The Humanitarian eXchange Language: Coordinating Disaster Response with Semantic Web Technologies. *Unpublished: Semantic Web Journal*, *1*, 1–5.

Homberg, M. van den, & Neef, M. (2015). Towards novel community-based collaborative disaster management approaches in the new information environment : an NGO perspective. *GRF Davos Planet@Risk*, *3*(1), 185–191.

Hristidis, V., Chen, S.-C., Li, T., Luis, S., & Deng, Y. (2010). Survey of data management and analysis in disaster situations. *Journal of Systems and Software*. http://doi.org/10.1016/j.jss.2010.04.065

Huang, W., Chen, K., & Xiao, C. (2014). Integration on heterogeneous data with uncertainty in mergency system, *211*, 483–490. http://doi.org/10.1007/978-3-642-38667-1

Inter-Agency Standing Committee, & Committee. (2012). Multi-Cluster/Sector Initial Rapid Assessment, (March).

the JNA consolidation project. (2014). *Pre-Disaster secondary data review* (Vol. 2014). Retrieved from http://www.lcgbangladesh.org/HCTT/JNA materials update/Pre-Disaster Secondary Data/Pre-Disaster Secondary Data - River Flooding - March 2014.pdf

Kou, G., Lou, C., Wang, G., Peng, Y., Tang, Y., & Li, S. (2010). A heterogeneous information integration framework for emergency management. *Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on*. http://doi.org/10.1109/ICICIS.2010.5534684

Kovavisaruch, L., Kamolvej, P., Prommoon, G., & Iamrahong, N. (2013). Data Integration for Thailand Disaster Risk and Response Management System. *Proceedings of PICMET: Technology Management for Emerging Technologies*, 1239–1248.

Kunii, O., Nakamura, S., Abdur, R., & Wakai, S. (2002). The impact on health and risk factors of the diarrhoea epidemics in the 1998 Bangladesh floods. *Public Health*, *116*(2), 68–74. http://doi.org/10.1038/sj.ph.1900828

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(01), 1. http://doi.org/10.1017/S0269888906000737

Lee, J., Bharosa, N., Yang, J., Janssen, M., & Rao, H. R. (2011). Group value and intention to use - A study of multi-agency disaster management information systems for public safety. *Decision Support Systems*. http://doi.org/10.1016/j.dss.2010.10.002

Li, J., Li, Q., Liu, C., Ullah Khan, S., & Ghani, N. (2014). Community-based collaborative information system for emergency management. *Computers and Operations Research*, *42*, 116–124. http://doi.org/10.1016/j.cor.2012.03.018

Liu, S., Brewster, C., & Shaw, D. (2013). Ontologies for Crisis Management: A Review of State of the Art in Ontology Design and Usability.

Mansuri, I. R., & Sarawagi, S. (2006). Integrating unstructured data into relational databases. In *Proceedings - International Conference on Data Engineering*. http://doi.org/10.1109/ICDE.2006.83

Meisner, R., Lang, S., Jungert, E., Almer, A., Tiede, D., Sparwasser, N., … Silvervarg, K. (2009). Data integration and visualization for crisis applications. *Remote Sensing from Space: Supporting International Peace and Security*, 141–160. http://doi.org/10.1007/978-1-4020-8484-3_10

Mescherin, S. A., Kirillov, I. A., & Klimenko, S. V. (2013). Principles of Information Integration of the Mono-profile Situational Centers for Effective Disaster Management, (Icaicte), 815–819.

Ministry of Flood and Disaster Management. (2010). *Standing Orders on Disaster*.

Morton, M., & Levy, J. L. (2011). Challenges in Disaster Data Collection during Recent Disasters. *Prehospital and Disaster Medicine*. http://doi.org/10.1017/S1049023X11006339

Naumann, F., & Raschid, L. (2006). Information Integration and Disaster Data Management (DisDM). *Workshop on Information Integration*, 25–27.

New Age. (2014). Search for: Flood Sirajganj 2014. Retrieved July 28, 2015, from http://newagebd.net/?s=flood+sirajganj+2014#sthash.G0PZgLpM.dpbs

Open Street Map. (2015). Open Street Map. Retrieved July 28, 2015, from https://www.openstreetmap.org/node/3360222393#map=11/24.2798/89.8349

Preece, G., Shaw, D., & Hayashi, H. (2013). Using the Viable System Model (VSM) to structure information processing complexity in disaster response. *European Journal of Operational Research*. http://doi.org/10.1016/j.ejor.2012.06.032

Rishe, N., & Yesha, Y. (2011). Geospatial Data Integration for Disaster Mitigation with TerraFly.

Samuel, R. (1976). Local history and oral history. *History Workshop Journal*, 191–208.

Sarker, M., Huque, I., Alam, M., & Koudstaal, R. (2003). Rivers, chars and char dwellers of Bangladesh. *International Journal of River Basin Management*, *1*(1), 61–80.

Stair, R., Reynolds, G., & Chesney, T. (2008). *Fundamentals of Business information Systems*. Thomson Learning.

Stancalie, G., Craciunescu, V., & Irimescu, A. (2009). SPATIAL DATA INTEGRATION FOR EMERGENCY SERVICES OF FLOOD MANAGEMENT. *Threats to Global Water Security*, 155–165.

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. http://doi.org/10.1.1.38.7672

TNO. (n.d.). TNO - Mission and Strategy. Retrieved July 22, 2015, from https://www.tno.nl/en/about-tno/mission-and-strategy/

Tran, M. C., Labrique, A. B., Mehra, S., Ali, H., Shaikh, S., Mitra, M., … West Jr, K. (2015). Analyzing the Mobile “Digital Divide”: Changing Determinants of Household Phone Ownership Over Time in Rural Bangladesh. *JMIR mHealth and uHealth*. http://doi.org/10.2196/mhealth.3663

Tsui, E. (2002). *Symposium on Best Practices in Humanitarian Information Exchange*.

Turban, E., Sharda, R., & Aronson, J. (2008). *Business intelligence: a managerial approach*. (S. Yagan & E. Svendsen, Eds.) (Second). New Jersey: Pearson. http://doi.org/10.1109/HICSS.2012.138

Twitter (various). (2014). Twitter search. Retrieved July 28, 2015, from https://twitter.com/search?q=flood sirajganj since%3A2014-07-01 until%3A2014-10-31&src=typd

Van De Weerd, I., Brinkkemper, S., Souer, J., & Versendaal, J. (2006). A situational implementation

method for web-based content management system-applications: method engineering and validation in practice. *Software Process Improvement and Practice*, *11*(5), 521–538. http://doi.org/10.1002/spip.294

Vatseva, R., Solakov, D., Tcherkezova, E., Simeonova, S., & Trifonova, P. (2013). Applying GIS in seismic Hazard assessment and data integration for disaster management. *Intelligent Systems for Crisis Management in Geoinformation and Carthography*, 349–355. http://doi.org/10.1007/978-3-642-33218-0

Verschuren, P., & Doorewaard, H. (2010). *Designing a Research Project*. *Chemistry & …*.

WARPO. (n.d.-a). Brief on NWRD. Retrieved July 6, 2015, from http://www.warpo.gov.bd/nwrd_brief.html

WARPO. (n.d.-b). *List of Data Layers for National Water Resources Database (NWRD)*. Retrieved from http://www.warpo.gov.bd/pdf/nwrd_Data_list.pdf

WFP. (2015). Geonode WFP. Retrieved July 1, 2015, from http://geonode.wfp.org/search/?q=bangladesh

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39. http://doi.org/10.1.1.198.5133

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2000). *Experimentation in software engineering: an introduction*. *Springer Netherlands* (Vol. 15).

Wolbers, J., & Boersma, K. (2013). The Common Operational Picture as Collective Sensemaking. *Journal of Contingencies and Crisis Management*, *21*(4).

Xiang-hong, W., Ji-ping, L., Sheng-hua, X., & Yong, W. (2011). Research on Integration and Analysis of Yushu, 625–632.

Xu, J. (2011). A Framework for Multiple and Heterogeneous Earthquake Disaster Information Fusion *, (March), 44–50.

Xu, W., & Zlatanova, S. (2007). Ontologies for Disaster Management Response. *Discovery*, 185–200. http://doi.org/10.1007/978-3-540-72108-6_13

Yang, Y., Lu, W., Domack, J., Li, T., Chen, S.-C., Luis, S., & Navlakha, J. (2012). MADIS: A Multimedia-Aided Disaster Information Integration System for Emergency Management. *Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, 233–241. http://doi.org/10.4108/icst.collaboratecom.2012.250525

Ying, W., Daoping, W., Guangli, L., & Di, L. (2010). Data Integration Platform for Village Emergency. *International Journal of Digital Content Technology and Its Applications*, *4*(4), 215–217. http://doi.org/10.4156/jdcta.vol4.issue4.22

Zhang, D., Zhou, L., & Nunamaker Jr, J. F. (2002). A Knowledge Management Framework for the Support of Decision Making in Humanitarian Assistance/Disaster Relief. *Knowledge and Information Systems*, *4*(3), 370–385. http://doi.org/10.1007/s101150200012

Ziegler, P., & Dittrich, K. R. (2007). Data Integration — Problems, Approaches, and Perspectives. In *Conceptual Modelling in Information Systems Engineering* (pp. 39–58). http://doi.org/10.1007/978-3-540-72677-7

13 Appendixes

13.1 List of appendixes

Appendix 1 Causes of floods in Bangladesh 107
Appendix 2 Responsibilities of the LDRCG 107
Appendix 3 Responsibilities of a Local Disaster Incident Manager 108
Appendix 4 Activities of disaster management in different phases of disaster 108
Appendix 5 D-form template 110
Appendix 6 JNA template 111
Appendix 7 Best practices for developing data and information products 117
Appendix 8 Best practices for information "preparedness" 117
Appendix 9 Respondent List 118
Appendix 10 Interview guide 119
Appendix 11 Data source table 120
Appendix 12 Overall implementation criteria evaluation integration methods 133
Appendix 13 Analysis of data integration capabilities by integration methods 139
Appendix 14 Interviews 147

Appendix 1 Causes of floods in Bangladesh

Causes of floods and its intensity

a) Heavy rainfall in the Himalayas and immense water volume due to melting of snow

b) Heavy rainfall in Assam valley and Northern Assam

c) Local heavy rainfall

d) Due to landslides upstream the alluvial soil raises the bed of rivers and canals

e) Large-scale tree felling in and around source of river, streams and water source

f) Blockade of natural drainage of water due to unplanned population settlement and construction of embankments

g) Formation of shoals in rivers and sand beds

h) Increase of sea level and its effect in low-lying areas

(Ministry of Flood and Disaster Management, 2010)

Appendix 2 Responsibilities of the LDRCG

Responsibilities of LDRCG

Establish a local emergency operation centre.

Liaise with the higher authority in order to inform the situation and obtain fed back.

Organize a directory with listing of local and national resources (human, infrastructural and financial) that can be used for this disaster incident.

Coordinate with armed forces team (if they are in).

Evaluate disaster situation and activate systems and procedures for disaster response and early recovery.

Mobilize resources and team for disaster response.

Ensure effective dissemination of warning signals.

Coordinate response and early recovery activities.

Supervise operations conducted by Urban Search and Rescue Taskforces.

Coordinate relief operations in post-impact recovery period.

Ensure the rapid supply of additional equipment/materials to places where telecommunication has been disrupted.

Determine priorities and issue instructions regarding relief materials, funds and transports

(Ministry of Flood and Disaster Management, 2010)

Appendix 3 Responsibilities of a Local Disaster Incident Manager

Responsibilities of Local Disaster Incident Manager

Taking control of the disaster incident and establishing a Disaster Incident Management Point.

Assess the situation and advising the appropriate authorities and agencies.

Liaise with the media, and developing/implementing a media plan for the disaster incident.

List local responder and resources availability.

Organize the task and conduct response programme as a team.

Determine priorities and time constraints.

Formulate a Disaster Incident Management Team (DIMT) made up of personnel to assist in the management of the disaster incident.

Determine the structure and role of the Disaster Incident Management Team.

Develop a disaster incident plan in conjunction with members of the DIMT.

Assign tasks to DMIT, response agencies and supporting services.

Coordinate resources and support.

Monitor events and responding to changing circumstances.

Report actions and activities to the appropriate agencies and authorities.

Ensure safety of all personnel at the disaster incident.

Establish media liaison procedures

Initiating recovery actions

(Ministry of Flood and Disaster Management, 2010)

Appendix 4 Activities of disaster management in different phases of disaster

Activities of disaster management in different phases of disaster

Mitigation

• Zoning and land use controls to prevent occupation of high hazard areas

• Barrier construction to deflect disaster forces

• Active preventive measures to control developing situations

• Building codes to improve disaster resistance of structures

• Tax incentives or disincentives

• Controls on rebuilding after events

• Risk analysis to measure the potential for extreme hazards

• Insurance to reduce the financial impact of disasters

Preparedness

• Recruiting personnel for the emergency services and for community volunteer groups

• Emergency planning

• Development of mutual aid agreements and memorandums of understanding

• Training for both response personnel and concerned citizens

• Threat based public education

• Budgeting for and acquiring vehicles and equipment

• Maintaining emergency supplies

• Construction of an emergency operations center

• Development of communications systems

• Conducting disaster exercises to train personnel and test capabilities

Recovery

• Disaster debris cleanup

• Financial assistance to individuals and governments

• Rebuilding of roads and bridges and key facilities

• Sustained mass care for displaced human and animal populations

• Reburial of displaced human remains

• Full restoration of lifeline services

• Mental health and pastoral care

Response

• Activating the emergency operations plan

• Activating the emergency operations center

• Evacuation of threatened populations

• Opening of shelters and provision of mass care

• Emergency rescue and medical care

• Fire fighting

• Urban search and rescue

• Emergency infrastructure protection and recovery of lifeline services

• Fatality management

Appendix 5 D-form template

FORM-D: ASSESSMENT OF LOSS AND DAMAGE

1	2	3	4	5			6	7	8	9	10	
Name of Upazila	Total Union (nos)	Total areas (Sq. km)	Char Areas (if) (sq km)	Total population (No)			Total families/ households	Cost of house Tk/Unit	Repairing Cost of house Tk/Unit	Other information (housing materials used)	Total disaster shelter (under LGED, DRR and other institutes)	Information source
												Baseline data/ Basic statistics
Name of Upazila	Affected Union (No)	Affected Area (Sq. km)	Affected char areas (sq km)	Affected population (No)	No. of dead buried/ burnt	No. of injured	Number of affected families	No. of house Fully damaged	No. of houses partially damaged	No of pacca house damaged	Shelter used during disaster (if)	

11		12		13		14				15		16		17	
Sheep and goat Population (No)		Cattle and buffalo Population (No)		Poultry Population (Chicken and Duck) (No)		Total crop land				Other farms (Pond fisheries, shrimp and other farms)		Total Power lines and accessories (unit)		Other infrastructure (if any) telecom Towers	
Death/washed out sheep and goats		Death and washed out cattle and buffalo including farms		Death and washed out poultry including farms		Fully damaged		Partially damaged		Other farm (Pond fisheries, shrimp, *Gher, fish fingerlings*		Damaged Power lines and accessories		Damaged other infrastructure (if any)	
Number	Taka/unit	Number	Taka/unit	No.	Taka/unit	ha	Taka/ha	ha	Taka/ha	ha	Taka/ha	Fully (cost tk)	Partially (cost tk)	Fully (cost tk)	Partially (cost tk)

Form for Assessment of damage and loss

18				19				20				21				22				23				24			
Total Mosques/Temples No				Carpeted roads (KM)				Other roads (Km)				Embankments (KM) River, Coastal, *Haor*				Total forest and nursery areas (ha)				Total Educational Institutes (College, primary and high schools, madrasha and other)				Total Telecom-communication means (tk)			
Number of damaged mosques/temples				Destroyed carpeted roads (KM)				Damaged other roads (Km)				Destroyed embankments (KM)				Damaged forests and nursery (ha)				Damaged educational institutes (College, primary and high school, madrasha and other)				Damaged Telecom-communication means (tk)			
Fully		Partially		Fully		Partially		Fully		Partially		Fully		Partially		Fully		Partially		Fully		Partially		Fully		Partially	
no	Tk/ unit	no	tk/ unit	km	Tk/ km	km	Tk/ km	km	Tk/ km	km	tk/km	km	Tk/ km	km	Tk/ km	ha	tk/ ha	ha	Tk/ ha	no	Tk/ unit	no	Tk/ unit	no	tk/ unit	no	tk/ unit

25				26						27				28				29				30			
Other Industry (garments, agro-processing, dry fish, salt etc)				Tube-wells (Shallow and deep)						Pond/Water reservoir (Nos)				Hospital/clinic/heath centre and accessories				Fishing nets boats/ Trawlers				Looms/ hand looms (No)			
Damaged others industry (garments, agro-processing)				Damaged tube-wells						Pond/Water reservoir (Nos)				Damaged Hospital/clinic/heath centre and related accessories				Lost /damaged boat/trawlers/fishing nets				Damaged Looms/handlooms			
Fully		Partially		Deep		Shallow		Hand driven		Fully		Partially		Fully		Partially		Fully		Partially		Fully		Partially	
no	tk/unit	no	tk/ unit	no	tk/unit	no	tk/ unit	no	tk/ unit	no	tk/ unit	no	Tk unit	no	Tk /unit	no	tk/ unit	no	tk/unit	no	tk/ unit	no	tk/unit	no	tk/ unit

Appendix 6 JNA template

JNA Phase 1 – Initial Days
Union Level Assessment Format

The objective of phase 1 assessment is to provide a rapid overview of the disaster and the need for assistance. Information collected through this format is the basis for decision making in the initial stages of a disaster, including the need for more detailed assessments. **This format is designed to assist in the completion of a Phase 1 joint report which is intended to be ready within three days of a sudden onset disaster. The format compiles information from several sources.**

Once the assessment has been agreed, the following steps will be taken by the Assessment Team to complete the phase 1 format:

- Visit the Union officials in all disaster affected Unions
- Fill in one format for each Union visited
- Consult government officials and elected representatives (Union Chairmen, Union Secretaries) to fill out this format
- Consult other GoB officials, UN agencies, and I/NGOs operating in the Union to confirm findings and/or address gaps in knowledge
- Complete the format using a combination of key informant interviews, field visits, and direct observation:
    a. Q 1-7 - completed by the assessment team
    b. Q 8-52 - asked of the authorities
    c. Q 53-56 - for the assessment team's judgment (with authorities)
    d. Q 57-60 - VERY important to ask authorities
    e. Q 61 - combination of assessment team and authorities' observations
    f. Q 62 - note sources of information.
- Visit a number of different locations in the affected Union
- When there is conflicting information, the Assessment Team should fill the format in using their best idea at the time, based on their understanding of the disaster's impact, their own professional experience, secondary data, and lessons learned from similar disasters
- When accessibility is challenged, but where phone communication is possible, the format can be completed over the phone with the permission of the organization/team coordinating the assessment.

This format is not a survey, rather it:

- Provides a standard format for recording and comparing information about a disaster
- Presents an understanding of the disaster's impact by the GoB officials and other stakeholders in the affected area
- Outlines how the disaster is likely to unfold in the days and weeks to follow
- Uses information based on local knowledge and past experiences
- Uses ***estimations*** of the numbers/percentages of people affected in different ways (providing an estimate is challenging, but local authorities are in the best position to do this).

Some information in this format relates to the pre-disaster situation in the location. This can and should be completed from pre-crisis sources (either in Dhaka or in the district). Do not waste the time of Union officials by asking questions that could be answered easily using other sources of information.

Basic Information

1. Name of the District:	2. Name of the Upazila:	3. Name of the Union:	4. Total population of the Union

5. Contact information for person completing this report (Team Leader):			6. Date of this format (DD/MM/YY):
Name:	Designation/ organization:	Contact number:	

7. Category of the area affected by the disaster (Predominantly):	☐ City Corporation ☐ Paurashava ☐ Rural/Union	8. Description of the area affected by the disaster (Predominantly):	☐ Char area ☐ Haor ☐ Coastal ☐ Hilly ☐ Island ☐ Flood plain

Disaster Event

9. Date of disaster/start of disaster (If it can be specified):		10. Time of disaster (If it can be specified):	

11. Type of disaster: *(If other, please specify)*	☐ Flood	☐ Earthquake	☐ Wind storm/Tornado
	☐ Landslide	☐ Water-logging	☐ Epidemic/Outbreak
	☐ Cyclone	☐ Cold Wave	☐ Other.....

12. As a result of the disaster are people thought to be (Please tick one per category):

0% = None
1-25% (Up to approximately ¼ of the population = A few
26-50% (Between ¼ and ½ of the population) = Some
50% - 100% (More than ½ of the population) = Many

12.01. Dead	12.02. Missing	12.03. Injured	12.04. Displaced
☐ None	☐ None	☐ None	☐ None
☐ A few	☐ A few	☐ A few	☐ A few
☐ Some	☐ Some	☐ Some	☐ Some
☐ Many	☐ Many	☐ Many	☐ Many
☐ Do not know	☐ Do not know	☐ Do not know	☐ Do not know
Approximately how many people are dead? *Only fill out if known*	Approximately how many people are missing? *Only fill out if known*	Approximately how many people are injured? *Only fill out if known*	Approximately how many people have been displaced? *Only fill out if known*

12.05. Where are people living in this Union since the disaster? *(Tick all that apply; If other, please specify)*	☐ Spontaneous settlement ☐ Pre-disaster location (original home) ☐ Collective center/public building ☐ Pre-disaster location (original village, but not original home, house damaged) ☐ Formal Camp ☐ Other.....
13. Has accessibility to the affected area been reduced by the disaster?	14. Type of accessibility reduced: *(If other, please specify)*
☐ Yes ☐ No ☐ Do not know	☐ Road ☐ Telecommunications ☐ Bridge ☐ Other.....

Specific Location of Affected Population				
15. Total number of Wards?	16. Number of affected Wards?	17. Estimated % of overall population affected?	18. Estimated population affected? *(Indicate the answer using # of affected persons OR # of affected households)*	
			Individuals	Households
19. Which are the three worst affected Wards? *(Please write their names)*	1. 2. 3. OR all Wards are equally badly affected ☐ (If so, tick box)			

WASH	
20. Has water supply been damaged/adversely affected? If No or Do not know, skip to Q21.	☐ Yes ☐ No ☐ Do not know
21. Approximate % of total population of the Union without access to safe drinking water? 0% = None 1-25% (Up to approximately ¼ of the population = A few 26-50% (Between ¼ and ½ of the population) = Some 50% - 100% (More than ½ of the population) = Many	☐ None ☐ A few ☐ Some ☐ Many ☐ Do not know
22. Has sanitation been damaged/adversely affected? *(If No or Do not know, skip to Q23)*	☐ Yes ☐ No ☐ Do not know
23. Approximate % of total population of the Union without access to safe sanitation? 0% = None 1-25% (Up to approximately ¼ of the population = A few 26-50% (Between ¼ and ½ of the population) = Some 50% - 100% (More than ½ of the population) = Many	☐ None ☐ A few ☐ Some ☐ Many ☐ Do not know
24. Are there or are there likely to be problems related to any of the following? *(Please tick if yes. If no, leave blank)*	☐ Safe and private latrines for women and girls ☐ Safe and private latrines for men and boys ☐ Safe and private spaces to bath for women and girls ☐ Safe and private spaces to bath for men and boys ☐ Sufficient hygiene materials for women
Shelter and Essential Non-food Items	
25. Is shelter an issue as a result of the disaster? *(If No or Do not know, skip to Q26)*	☐ Yes ☐ No ☐ Do not know
26. Approximate number of households in need of immediate shelter? 0% = None 1-25% (Up to approximately ¼ of the population = A few 26-50% (Between ¼ and ½ of the population) = Some 50% - 100% (More than ½ of the population) = Many	☐ None ☐ A few ☐ Some ☐ Many ☐ Do not know
27. Are people are likely to be without sufficient bedding/blankets?	☐ Yes ☐ No ☐ Do not know
28. Are alternative places available to people who require shelter (e.g. communal shelters or buildings that can be used as collective centers)?	☐ Yes ☐ No ☐ Do not know

Food and Livelihoods	
Explanation on how to interpret the severity criteria in the food security questions: Less than 20% = Low damage; 20-50% = Moderate damage; 50% - 100% = Severe damage	
29. Are people likely to have had their food stores destroyed or damaged as a result of the disaster? *(If No or Do not know, skip to Q30)*	☐ Yes ☐ No ☐ Do not know
30. If yes, estimate the severity of the damage:	☐ Severe ☐ Moderate ☐ Low ☐ Do not know
31. Are markets in the affected area generally functioning?	☐ Fully ☐ Partly ☐ Not functioning ☐ Do not know
32. Do markets have stocks of food?	☐ Yes ☐ No ☐ Do not know
33. Are markets generally accessible by the local community?	☐ Yes ☐ No ☐ Do not know
34. Is the disaster likely to have an effect on long term food security for those affected?	☐ Yes ☐ No ☐ Do not know
35. Is immediate food security affected because of the impact of the disaster on livelihoods?	☐ Yes ☐ No ☐ Do not know
36. Are households able to cook food/boil water since the disaster?	☐ Yes- majority can cook ☐ No, few can cook ☐ Do not know
37. Which livelihoods are likely to be most affected? *(If others, please specify)*	☐ Non-agricultural day labour ☐ Agricultural day labour ☐ Small and marginal farmers ☐ Medium and big farmers ☐ Others (Specify).....
38. What is the severity of damage of the major crop/crops?	☐ Severe ☐ Moderate ☐ Low ☐ No damage ☐ Do not know
39. Have there been losses to agricultural inputs and equipment?	☐ Severe ☐ Moderate ☐ Low ☐ No loss ☐ Do not know
40. What is level of death or loss of livestock (animals and poultry)?	☐ Severe ☐ Moderate ☐ Low ☐ No loss or death ☐ Do not know
41. What is the damage to fisheries?	☐ Severe ☐ Moderate ☐ Low ☐ No damage ☐ Do not know
Education	
42. What is the total number of schools/education institutions in the Union *(this information will be provided from pre-crisis data; do not ask at Union level)*	
43. Is education an issue as a result of the disaster? *(If not or do not know, skip to Q45)*	☐ Yes ☐ No ☐ Do not know
44. How many schools/education institutions are not functioning because of the disaster? 0% = None 1-25% (Up to approximately ¼ of the population = A few 26-50% (Between ¼ and ½ of the population) = Some 50% - 100% (More than ½ of the population) = Many	☐ None ☐ A few ☐ Some ☐ Many ☐ Do not know

45. What are the main reasons for schools/education institutions not functioning? (tick all that apply)	☐ No teachers ☐ No students ☐ Infrastructure damage ☐ Study materials damaged ☐ School not accessible ☐ Schools used as shelter ☐ Do not know

Health and Nutrition	
46. What is the total number of health facilities in the Union *(this information will be provided from pre-crisis data; do not ask at Union level)*	
47. Is health an issue as a result of the disaster? *(If not or do not know, skip to Q49)*	☐ Yes ☐ No ☐ Do not know
48. How many health facilities are not functioning because of the disaster? **0% = None** **1-25% (Up to approximately ¼ of the population = A few** **26-50% (Between ¼ and ½ of the population) = Some** **50% - 100% (More than ½ of the population) = Many**	☐ None ☐ A few ☐ Some ☐ Many ☐ Do not know
49. What are the main reasons for health facilities not functioning? *(If other, please specify)*	☐ Staff shortage ☐ Damage to Building ☐ Supplies/medicine shortage ☐ Do not know ☐ Not applicable/all are functioning ☐ Other.....
50. Is/are health facilities accessible?	☐ Yes ☐ No ☐ Do not know
51. Are there any underlying health concerns in Union? *(information should come from pre-disaster knowledge; if other, please specify)*	☐ Malnutrition ☐ Communicable diseases ☐ Do not know ☐ No underlying concerns ☐ Other.....
52. Are there any health concerns as a result of the disaster? *(If other, please specify)*	☐ People injured ☐ Dead bodies (people/animals) ☐ Communicable disease ☐ Anti-natal Care ☐ No ☐ Do not know ☐ Other.....
53. Are there safe and private places for women to breastfeed?	☐ Yes ☐ No ☐ Do not know

Post Disaster Scenario Development	
54. What are the present weather conditions: *(If other, please specify)*	☐ Normal ☐ Heavy Rain ☐ Very cold ☐ Windy ☐ Hot ☐ Humid ☐ Cloudy ☐ Other.....
55. In the coming two weeks, the situation in the disaster affected area is most likely to:	☐ Stay the same ☐ Improve ☐ Worsen
56. What factors could make the situation for affected people worse? *(If other, please specify)*	☐ Continuous heavy rain ☐ Water level rising ☐ Aftershock ☐ Disease Outbreak ☐ No rain ☐ Other.....
57. In the worst case scenario, how many people might this affect?	

Available Resources, Coping Strategies and Support Required		
58. Outline resources available at the Union level in the following sectors:		
Sector	Is extra assistance required?	Comment on what assistance is required.
58.01. WASH	☐ Yes ☐ No ☐ Do not know	
58.02. Shelter and non-food items	☐ Yes ☐ No ☐ Do not know	
58.03. Food	☐ Yes ☐ No ☐ Do not know	
58.04. Livelihoods	☐ Yes ☐ No ☐ Do not know	
58.05. Education	☐ Yes ☐ No ☐ Do not know	
58.06. Health	☐ Yes ☐ No ☐ Do not know	
59. How would the **assessment team** describe the immediate overall relief needs in this Union *(needs in coming days and weeks)*:	☐ Serious need of assistance ☐ Some need of assistance ☐ Needs can be managed with resources available at Union level	
60. Which appear to be the **highest priority** for immediate assistance? *(rank up to, but no more than three)*	☐ Water	☐ Sanitation
	☐ Shelter	☐ Bedding and blankets
	☐ Clothing	☐ Food
	☐ Livelihoods	☐ Education
	☐ Health	
61. How would you describe the recovery needs in this Union*(needs in coming three or more months)*:	☐ Serious need of assistance ☐ Some need of assistance ☐ Union and communities coping strategies will be enough	
62. Any further comments or observations:		

63. Information sources (please indicate the sources of information used in compiling this report)	
☐ Union Parishad Chairman	☐ UN (please name agency)
☐ UP Secretary	☐ Affected community member (male)
☐ Union Parishad/ Ward Member	☐ Affected community members (female)
☐ I/NGOs (please name organization)	☐ Direct Observations of assessment team
☐ Other.....	

Appendix 7 Best practices for developing data and information products

Best Practices for Developing Data and Information Products

Define user needs and utilise data sets and formats that directly support decision-making at the field and headquarters levels. Identify user groups, conduct user requirement analysis, inventory information resources and define core information products based on user input.

Develop and implement information products on operationally and strategically relevant themes, such as the location and condition of the affected population, "who is doing what, where?" and assessments of needs.

Consult data providers as well as affected populations when designing information products. Without direct feedback to and from the data providers and affected populations, community information collection strategies will fail. The humanitarian community, governments and beneficiaries must be thought of as customers as well as data providers.

Create and disseminate templates for sectoral assessments and analysis, such as the Rapid Village Assessment (RVA) tool or the Microsoft/Save the Children assessment tool for personal digital assistants (PDA), to speed data collection. Create maps and graphic presentations to effectively communicate information to decision-makers.

Appendix 8 Best practices for information "preparedness"

Best Practices for Preparedness

Maintain preparedness "toolboxes" for online and offline distribution. These toolboxes provide guidelines and reference tools for the rapid-deployment of HICs or the establishment of Web sites, intranets and databases under a variety of field conditions. Toolboxes should include data standards, operating procedures, training materials, database templates and manuals.

Develop surge capacities for rapid deployment. Maintain rosters of experienced information professionals, equipment caches and training programmes.

Preserve institutional operational memory. Define and adhere to sound data and information management policies and techniques for handling large volumes of information. Document datasets with metadata. Maintain quality control and preserve organisational learning to avoid starting from scratch with each new emergency.

Define an exit strategy. Develop a clear, phase-out strategy, including transitioning to development activities and creating archiving systems to maintain access by current and future stakeholders after the project is closed.

Appendix 9 Respondent List

Role	Name	Level	Interview Technique	Date
Government disaster manager	Latif Khan (DMIC)	National	Semi structured interview	26/4/15
NGO disaster manager	Mahbubur Rahman (CARE)	National	Semi structured interview	27/4/15
UNDP management	Ahmadul Hassan (UDNP)	National	Semi structured interview	27/4/15
Database specialist JNA	Liam Costello (FAO)	National	Semi structured interview	28/4/15
Government Project implementation officer (DNA)	Ismail Biddut (Keraniganj Upazila)	Upazila	Semi structured interview	29/4/15
JNA involved	Clody Ayliffe	National	Semi structured interview	29/4/15
NGO Disaster management	Sanjukte (Concern)	National	Semi structured interview	30/4/15
NGO Disaster management	Abdul Hamid (Concern)	National	Semi structured interview	30/4/15
NGO Disaster management	MMS Directors (Amir Hossain + Habib + Motaker Hossen)	Local	Focus group	2/5/15
NGO Disaster responders	MMS Disaster responders (Atique + Chairo + Shamim + Regina + Hamid + Hilal + Pharid)	Local	Focus group	2/5/15
Local Disaster responders	Gorzan Union Local disaster responders (TamTam volunteers, Imam, Teachers, Entrepreneurs)	Union	Focus group	4/5/15
Government local Disaster responders	Chowhali Upazila (Upazila DMC, Union DMC, AnserVDP, Digital Entrepreneur)	Upazila	Focus group	5/5/15
Government local Disaster responders	Belkuchi Upazila (Upazila Chairman, Union Chairman, Project implementation Officer)	Upazila	Semi structured interviews	6/5/15
Responding Community	Fisherman and Farmer	Local	Focus group	7/5/15

Appendix 10 Interview guide

Intro: We are conducting research towards the role of information in disaster response and preparation. You have been selected as respondent because you have experience as an actor in preparation or response to a disaster. With the results we aim to build a prototype which assists disaster responders. Examples of prototype functionality could be: fusing all data available in disaster context (D-form, JNA etc.), so we'll be able to: visualise needs of affected, predict best shelter area or prioritize response actions based on data.

Context

1. What is your job title?
2. In how many disaster responses did you participate?
3. How long are you active in this field?
4. Is there any other relevant experience you wish to share?

Activity

1. Can you describe the activities you're performing in preparation of a disaster?
    a. What challenges occur?
    b. What are your most important activities?
    c. What activities are most time consuming?
2. Can you describe the activities you're performing in the response to a disaster?
    a. What challenges occur?
    b. What are your most important activities?
    c. What activities are most time consuming?

Decisions

1. Can you describe decisions you need to make (while performing your activity)?
    a. What are the most important decisions?
    b. How do you make these decisions?

Information Need

1. When you're in *preparation* of a disaster, can you list and describe the information you are currently using?
    a. Can you describe your information experience during the 2014 flooding?
2. When you're in *response* of a disaster, can you list and describe the information you're currently using?
    a. Can you describe your information experience during the 2014 flooding?
3. What information could make your activities more effective in preparation phase?
4. What information could make your activities more effective in response phase?
5. Please don't think of any constraints, what would make your actions in the field even more effective?

Appendix 11 Data source table

Data Source	Description	Indicators	Format and type	Timing
JNA	Joint needs assessment. Set up by the HCTT (Humanitarian Coordination Task Team), which goal is to strengthen the collective capacity of the actors in Bangladesh (HCTT, 2015). From the official template of the first phase JNA: *"The objective of phase 1 assessment is to provide a rapid overview of the disaster and the need for assistance. Information collected through this format is the basis for decision making in the initial stages of a disaster, including the need for more detailed assessments."*(HCTT, 2014b) An overview of the first phase assessment questionnaire is shown in Appendix 6**Error! Reference source not found.**.	**Union disaster statistics:** Name of district Name of Upazila Name of union Total population of union Contact info enumerator Date of the report Category of area (city, pourashava, rural) Description of area affected (char, coastal, island, haor, hilly, flood plain) Date of disaster Time of disaster Type of disaster (flood, landslide, cyclone, earthquake etc.) No. of death No. of displace No. of missing No. of injured Where are the people living since disaster (spontaneous settlement, original home, public building, formal camp, other house) Accessibility area Accessibility of roads affected by disaster? Accessibility of Telecommunications affected by disaster? Accessibility of brigdes affected by disaster? Accessibility of other affected by disaster? No. of wards No. affected wards % of population affected	• PDF report (unstructured text) • Excel file	• 8 September (24 days after start of disaster)

		No. people affected		
		No. households affected		
		Which are the worst affected wards		
		Is the water supply affected?		
		No. people without safe drinking water		
		Is sanitation affected?		
		No. people without sanitation		
		Are there problems with latrines/bath for woman/girls/man/boys (and hygiene materials for woman)		
		Is shelter an issue due to disaster?		
		No. households in need of shelter?		
		Are there enough beds/blankets?		
		Are there enough alternative shelter places available?		
		Are the foodstores affected by the disaster		
		Severity of damage to foodstores		
		Are the markets affected or still functioning		
		Do markets have stocks of food		
		Are markets generally accessible by the local community		
		Is the long term food security affected		
		Immediate food security affected		
		Are the households able to cook food since disaster		
		Most affected livelihoods (non agri day labour, agri day labour, small/medium/big farmers)		
		Severity of damage to crops		
		Lost agricultural equipment		
		Level of death/loss livestock		
		Level of damage to fisheries		

		Total education institutes Education problems due to disaster? Affected education institutes Reasons for not functioning education institutes Total health facilities Health problems due to disaster? Affected health institutes Reasons for not functioning health facilities Accessibility health facilities Underlying health concerns? Health concerns due to disaster? Safe places to breastfeed? Present weather conditions Outlook disaster situation What factor would worsen situation for people Worst case scenario, how many people will be affected WASH resources Shelter resources Food resources Livelihoods resources Education resources Health resources Relief need overview Highest priority for immediate assistance Recovery needs overview Information sources		
D-Form	Damage and needs assessment performed by the government based on the Standing Orders on Disasters (Ministry of Flood and Disaster	**Upazila information on:** Name of Upazila Death Amount of unions Amount of affected unions Total area (sq km)	• PDF situation reports (unstructured text) • Excel (Bangla)	• Final results not published, only summarized in situation reports.

	Management, 2010). See Appendix 5 for the template of the questionnaire.	Affected area (sq km) Char area (sq km) Affected char area (sq km) Total population of Upazila Affected population No. of injured Total amount of families Affected families Cost of house (tk/unit) No. of house fully damaged Repairing cost of house (tk/unit) No. of house partially damaged Other info (housing materials needed) No. of pacca house damaged Total disaster shelter (LGED, DRR and other) Shelter used No. of sheep and goat Death/lost sheep and goat Cost per lost sheep or goat No. of cattle and buffalo Death/lost cattle and buffalo Cost per lost cattle or buffalo No. of poultry Death/lost poultry Cost per lost poultry Total crop land ha fully damaged crop land cost/ha fully damaged crop land ha partially damaged crop land cost/ha partially damaged crop land other farms (pond fisheries, shrimp etc.)		

		ha damaged other farms (pond fisheries, shrimp etc.)		
		cost/ha damaged other farms (pond fisheries, shrimp etc.)		
		No. of power lines and accessoiries		
		cost/No. Fully damaged power lines and accessoiries		
		cost/No. partially damaged power lines and accessoiries		
		No. other infrastructure (telecom towers)		
		cost fully damaged other infrastructure (telecom towers)		
		cost partially damaged other infrastructure (telecom towers)		
		No. of mosques/temples		
		No. of fully damaged mosques/temples		
		No. of partially damaged mosques/temples		
		cost/No. of fully damaged mosques/temples		
		cost/No. of partially damaged mosques/temples		
		km of carpeted roads		
		km of fully destroyed carpeted roads		
		km of partially destroyed carpeted roads		
		cost/km of fully destroyed carpeted roads		
		cost/km of partially destroyed carpeted roads		
		km of other roads		
		km of fully destroyed other roads		
		km of partially destroyed other roads		
		cost/km of fully destroyed other roads		
		cost/km of partially destroyed other roads		
		km embankements		

		km fully destroyed embankements		
		km partially destroyed embankements		
		cost/km fully destroyed embankements		
		cost/km partially destroyed embankements		
		ha forest and nursery		
		ha fully destroyed forest and nursery		
		ha partially destroyed forest and nursery		
		cost/ha fully destroyed forest and nursery		
		cost/ha partially destroyed forest and nursery		
		No. educational institutes (college, primary and high schools, madrasNo. etc.)		
		No. fully destroyed educational institutes (college, primary and high schools, madrasNo. etc.)		
		No. partially destroyed educational institutes (college, primary and high schools, madrasNo. etc.)		
		cost/No. fully destroyed educational institutes (college, primary and high schools, madrasNo. etc.)		
		cost/No. partially destroyed educational institutes (college, primary and high schools, madrasNo. etc.)		
		No. telecom-communication means		
		No. fully destroyed telecom-communication means		
		No. partially destroyed telecom-communication means		
		cost/No. fully destroyed telecom-communication means		
		cost/No. partially destroyed telecom-communication means		

		No. other industry (garments, agro-processing, dry fish, salt etc.)		
		No. fully destroyed other industry (garments, agro-processing, dry fish, salt etc.)		
		No. partially destroyed other industry (garments, agro-processing, dry fish, salt etc.)		
		cost/No. fully destroyed other industry (garments, agro-processing, dry fish, salt etc.)		
		cost/No. partially destroyed other industry (garments, agro-processing, dry fish, salt etc.)		
		No. tube wells (shallow and deep)		
		No. damaged tube wells (deep)		
		No. damaged tube wells (shallow)		
		No. damaged tube wells (hand driven)		
		cost/No. damaged tube wells (deep)		
		cost/No. damaged tube wells (shallow)		
		cost/No. damaged tube wells (hand driven)		
		No. pond/water reservoir		
		No. fully destroyed pond/water reservoir		
		No. partially destroyed pond/water reservoir		
		cost/No. fully destroyed pond/water reservoir		
		cost/No. partially destroyed pond/water reservoir		
		No. hospital/clinic/health centre		
		No. fully destroyed hospital/clinic/health centre		
		No. partially destroyed hospital/clinic/health centre		
		cost/No. fully destroyed hospital/clinic/health centre		

		cost/No. partially destroyed hospital/clinic/health centre		
		No. fishing nets/boats/trawlers		
		No. fully destroyed fishing nets/boats/trawlers		
		No. partially destroyed fishing nets/boats/trawlers		
		cost/No. fully destroyed fishing nets/boats/trawlers		
		cost/No. partially destroyed fishing nets/boats/trawlers		
		No. looms/hand looms		
		No. fully destroyed looms/hand looms		
		No. partially destroyed looms/hand looms		
		cost/No. fully destroyed looms/hand looms		
		cost/No. partially destroyed looms/hand looms		
Geonode WFP	A web application for creating and sharing geospatial data and maps designed for non-GIS experts. This resource has several layers and maps about Bangladesh (WFP, 2015).	Flood prone areas	Maps	• Available before disaster
		Coast boundaries		
		Mayor towns		
		Roads		
		Power plants		
		Dams		
		Landuse		
		Inland waters		
		Main rivers		
		River lines		
		Admin boundary levels (union, upazila, district, division levels)		
		Bangladesh population 2011		
		Railways		
		Electricity lines		
DMIC portal – 4W DB	Web-based GIS system to show the actors within the field of Disaster risk reduction activities.	Who	• Database • GIS	• Available before disaster (updated two times a year)
		What		
		Where		
		When		

	(DMIC, 2015d) (DMIC, 2015a)			
DMIC portal – Situation Reports (Inundation)	Government issued disaster situation reports (every day during the flood of 2014). (CDMP II, 2014)	Highlights (including predictions) Stations above water danger level (+- cm) Rise or Fall of water level at stations Rainfall (last 24h) Rainfall (coming day) map Water level of measure stations map Affected upazilas (per district) Effected family (partially, full) (per district) Effected people (per district) Distributed relief (rice & cash) (per district) Disaster Risk Reduction officer Telephone number (per district)	PDF (unstructured text)	• Issued the first on 20 august (and then weekly till 10 september)
Community Risk Assessments	Not available for our focus area.		PDF	
District Disaster Management Plan	Large document with multiple perspectives, current situation, disaster risk, risk reduction, emergency response, rehabilitation plan. (District Level Disaster Management Comittee Sirajganj, 2014)	Infrastructure (km of road, amount of bridges) Social resources (number of homes, nr of toilets) Active NGOs (target audience and amount of beneficiaries) Weather and climate Hazards and Vulnerabilities per region and sector Seasonal calendar of livelihood Seasonal calendar of disaster risk Influences of climate change Disaster risk reduction Identification of risks Means of reduction Development plans of NGO	• PDF report (unstructured text)	• Available before disaster (issued in 2014)

		Disaster management work plan Emergency response Contingency plans (who does what etc.) List of shelters incl telephone numbers Telephone numbers of members disaster management committees For every upazila: Families, Males, Females etc. Schools Literacy % Nr of farms, roads etc. Table with locations of infrastructure like: embankment and bridges Names and addresses of health centres All bank branches (including village location) Arsenic pollution Capacity of safety (shelter) places		
Secondary data assessment (ACAPS/HCTT)	A review of secondary data to provide an adequate overview of the situation in a river flooding. Most numbers are on a national scale. (the JNA consolidation project, 2014)	Climate and geography Population statistics Administrative divisions Disaster management roles Health indicators Birth rate death rate Economy and markets Social cultural statistics Mobile phones Poverty profile Disaster management structure and responsibilities Flooding profile (history, warning systems,locations) Baseline livelihood and food security Infrastructure/Logistics	• PDF report (unstructured text)	• Available before disaster (conducted in March 2014)

		Nutrition Shelter		
DMIC disaster incident database	Maps and numbers on past disasters. (DMIC, 2015d) (DMIC, 2015c)	Time Duration Type Location Damage info (short)	• Database • GIS	• Updated frequently
DMIC hazard map	JPEG files with inundation maps on upazila level. (DMIC, 2015d) No data on Sirajganj yet. (DMIC, 2015b)	Surge height	• JPEG images	• Available before disaster
DMIC union fact sheets	1700+ union factsheets available online. (DMIC, 2015d) No data on Sirajganj yet. (DMIC, 2015e)	General information	•	• Available before disaster
FFWC (Flood Forecasting and Warning Centre)	A Bangladesh institution responsible for the forecasting and monitoring of floods. (FFWC, 2015)	Indicators of current flood level Forecasts Flood summary Flood bulletin 5-day forecast (experimental) 10-day forecast Inundation map Rainfall map	• PDF (unstructured text) • JPG (maps)	• Before and during disaster
BBS (Bangladesh Bureau of Statistics)	Government institution responsible for providing statistical information for decision-making. (BBS, 2015)	High level census data Reports on the cultivation and productivity of certain crops Poverty maps Literacy study Household income and expenditure survey (mostly on division level) Analysis of vital statistics (mortality rates etc.)	• PDF	• Available before disaster (Census is done every couple of years)
Flood shelter list	Excel file with all flood shelters. Provided by the DMIC	Location of shelter (union and 'name of facility')	• Excel	• Available before disaster

National Water Resources Data	Organization that manages a large amount of data with different angles, mostly around water. Due to the high amount of datasets mentioned in the data list, we only share the once applicable to our case. NWRD holds more than 400 data layers, out of which 125 layers are spatial data. Data in the NWRD are organized in several main groups which are: Base data, Surface water, Groundwater, Soil and Agriculture, Fisheries, Forest, Socio-economic, Meteorological, Environment and Images (WARPO, n.d.-a)	Crop areas (shapefiles) Amount of catched fish per region and type of water Socio economic Char information Flood and riverbank erosion (shape file) Bankline of main and major rivers (shapefile) (WARPO, n.d.-b)	• Various	• Available before disaster
Water points data base (Department of public health and engineering)	Tried to access but got only data with question marks. Send an email for help. Hoping for response.			
Community Risk Assessments & Baseline reports	Provided by NARRI consortium, but the reports aren't accessible online (I send a request)			
Social media (twitter)	In the rural areas of Bangladesh not really used, as is found when searching for flood related tweets in the flood period. Mostly found are tweets from news agencies or from international NGOs. (Twitter (various), 2014)	None		

News	Several online newspapers report on ongoing disasters in Bangladesh. Some of the leading sources are: The Daily Star, BD News 24, Dhaka Tribune, Daily Observer and The New Age. (the Daily Observer, 2014; the Daily Star, 2014; Dhaka Tribune, 2014; New Age, 2014)	Flood news Amount of affected Current situation Damage	• Unstructured web documents	• Directly after disaster and daily updates during disaster
OSM	Open street map. This is a map build by the community with their local knowledge. Unfortunately, the char islands and most areas in Sirajganj aren't correctly mapped yet. However, this is a great initiative and we should keep an eye on this. There is an active Facebook community that voluntarily maps requested areas, due to time constraints we did not post a request. (Open Street Map, 2015)			

Appendix 12 Overall implementation criteria evaluation integration methods

Criteria	**Functionality of to-be system**	**Low resource implementation**		**Availability of resources**			**Data management architecture**
		Internet connectivity variability	Processing limits	Skills	Time	Money	
Generating mapping schemas	Low (users should possess knowledge on SQL querying, no easy dashboard building)	Medium (mapping schema is subject to internet connectivity loss)	High (with this relative low amount of data sources it should be possible on a personal computer)	Low (complex language is used for schema matching)	High (generating the mapping of only a few sources should be done in a low amount of time)	x	Medium (data could stay in their own place however, the schema should be changed when the data changes, sensitive to streaming data however)
Adaptive Query Processing	Low (users should possess knowledge on SQL querying, no easy dashboard building)	High (the optimizations incorporate on the availability of the sources)	Low (it is questionable whether we need such a heavy optimization when the data is not that large)	Medium (complex language is used for schema matching)	Low (we probably need to invest a lot of time in optimization)	x	Medium (data could stay in their own place however, the optimization should be changed when the data changes)

XML	Medium (data is not easily analysable due to its semi structured nature, but simple queries are possible))	Medium (for XML integration we should have access to the source)	High (XML does not pose high processing requirements)	Medium (XML is a straightforward language to use)	x	x	Medium (we should first transform the sources and store them somewhere else)
Model management	Low (users should possess knowledge on a model language, no easy dashboard building)	Medium (model management is subject to internet connectivity loss)	High (with this relative low amount of data sources it should be possible on a personal computer)	Low (complex language is used for schema matching)	High (generating the mapping of only a few sources should be done in a low amount of time)	x	Medium (data could stay in their own place however, the schema should be changed when the data changes)
Peer-to-Peer data management	High (everybody share his data and integrates it with a neighbour, therefor analysis is possible)	Medium (connections between peers are internet based)	High (everybody is responsible for their own data, therefore processing is not an issue)	High (everybody does their own data integration, therefore you can use the skills you possess)	High (the total time will be reduced because all actors work at the same time)	x	High (all data stays in the same place and is integrated)

Data Warehousing + ETL	High (easy visualisations and analysis is possible)	Medium (sources could stay in place with resulting difficulties in internet variability, depending on implementation)	Low (data warehouses can be very heavy)	Medium (you need data warehouse architects to model the data)	Low (modelling all data is a time consuming process)	x	Medium (data is extracted and moved to different location)
Personal data integration						x	
Collaborative integration	High (data is integrated on the run by all actors and afterwards accessible for analysis)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	Medium (not a lot of processing power is required, but it might not fit on a PC)	High (data integration is crowdsourced, only skills to set up the system is required, but the real integration is done with regular users)	Medium (integration is not very quick, but depends on the users performance)	x	Medium (all data is extracted and put in a different location)

Dataspace systems	Medium (data is not directly analysable, it should be integrated on a pay as you go principle)	High (data spaces are designed for variable data)	Medium (dataspaces are a thin layer without much functionality so implementation isn't costly processing wise)	Medium (implementing a data space service layer is not standard for developers)	Medium (data is not integrated from the start, you need to decide which data you integrate on a pay as you go principle)	x	High (very loose, data stays in same place)
Humanitarian exchange language (HXL)	High (resulting tagged excel sheets are good for analysis and visualisation)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	High (no additional processing power required for HXL)	High (skills for HXL are easily acquired)	High (tags could be quickly added)	x	Medium (data should be shared in a different place and tagged)

Usage of text mining algorithms	Medium (all information is extracted from unstructured text, it is not connected to the structured sources though)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	Medium (text mining algorithms can be applied on a single PC, however the more pdfs we analyse, the more processing we need given the time constraints)	Medium (people who can apply text mining are required)	Medium (extracting the relevant info could take up some time)	x	Low (we need to download the data and store the extracted info on a server for usage by the responders)
link text documents to structured information	Medium (parts from the text are linked to entities in the database, this is probably hard to analyse, however, it might be easy to visualise)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	Medium (we assume a PC is sufficient for the amount of data in this context)	Low (advanced programming is required)	Medium (the integration can be done quickly, due to the relative low amount of data)	x	Low (we probably need to download the data first and store the results on a different server)
Integrating unstructured data into relational databases	High (information is ready for visualisation and analysis)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	Medium (it should be feasible to apply the algorithm on a PC)	Low (advanced programming is required)	Medium (the integration can be done quickly, due to the relative low amount of data)	x	Low (we probably need to download the data first and store the results on a different server)

Ontology guided information extraction from unstructured text	Medium (information is ready for visualisation and analysis, not linked to structured info though)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	Medium (it should be feasible to apply the algorithm on a PC)	Low (advanced programming is required)	Medium (the integration can be done quickly, due to the relative low amount of data)	x	Low (we probably need to download the data first and store the results on a different server)
Unstructured information integration through data-driven similarity discovery	Medium (information is ready for visualisation and analysis, not linked to structured info though)	Medium (we are dependent on internet for the download of the sources, and afterwards for the access to the integrated data)	Medium (it should be feasible to apply the algorithm on a PC)	Low (advanced programming is required)	Medium (the integration can be done quickly, due to the relative low amount of data)	x	Low (we probably need to download the data first and store the results on a different server)

Appendix 13 Analysis of data integration capabilities by integration methods

Criteria	**Type of information**	**Location**	**Model**	**Schema**	**Heterogeneity**					
					Inconsistency of syntax	Different measure units	Inconsistency of representation	Redundancy of entities	Violation of Cardinality	Semantic
Generating mapping schemas	Relational vs Relational	x	x	High (resolves schema issues)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)

Adaptive Query Processing	Relational vs Relational	x	x	High (resolves schema issues)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)
XML	Unstructured can be made semi structured, and relational can be transformed to semi structured	x	x	Low (does not solve schema issues)	Low (XML is just a layer on top of the data, all heterogeneity stays)	Low (XML is just a layer on top of the data, all heterogeneity stays)	Low (XML is just a layer on top of the data, all heterogeneity stays)	Low (XML is just a layer on top of the data, all heterogeneity stays)	Low (XML is just a layer on top of the data, all heterogeneity stays)	Low (XML is just a layer on top of the data, all heterogeneity stays)

Model management	Relational vs Relational	x	x	High (resolves schema issues)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)
Peer-to-Peer data management	Undefined (but everything should be possible)	x	x	High (resolves schema issues)	High (every actor integrates his own data, at which he is a domain expert, therefore this should be done easily)	High (every actor integrates his own data, at which he is a domain expert, therefore this should be done easily)	High (every actor integrates his own data, at which he is a domain expert, therefore this should be done easily)	High (every actor integrates his own data, at which he is a domain expert, therefore this should be done easily)	High (every actor integrates his own data, at which he is a domain expert, therefore this should be done easily)	High (every actor integrates his own data, at which he is a domain expert, therefore this should be done easily)

Data Warehousing + ETL	Excel and relational	x	x	High (resolves schema issues)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the data architect should possess domain knowledge)
Personal data integration		x	x							
Collaborative integration	High (can be modified to any combination)	x	x	High (resolves schema issues)	Medium (these issues should be overcome manually, therefore the users should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the users should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the users should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the users should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the users should possess domain knowledge)	Medium (these issues should be overcome manually, therefore the users should possess domain knowledge)

Dataspace systems	High (all types are supported)	x	x	Low (does not solve schemas itself, other methods should be applied)	Low (these issues should be overcome with another method)	Low (these issues should be overcome with another method)	Low (these issues should be overcome with another method)	Low (these issues should be overcome with another method)	Low (these issues should be overcome with another method)	Low (these issues should be overcome with another method)
Humanitarian exchange language (HXL)	Excel vs Excel	x	x	High (solves schema issues)	Low (should be done manually)	Low (should be done manually)	Low (should be done manually)	Low (should be done manually)	Low (should be done manually)	Medium (an ontology is integrated in this solution)

Usage of text mining algorithms	Unstructured	x	x	Low (creates a schema for the unstructured sources but doesn't match it with the structured ones)	Low (extracts the info as-is, so no syntax integration)	Low (extracts the info as-is, so no syntax integration)	Low (extracts the info as-is, so no syntax integration)	Low (extracts the info as-is, so no syntax integration)	NA (there is no explicit cardinality in unstructured text)	Low (extracts the info as-is, so no syntax integration)
link text documents to structured information	Unstructured vs Structured	x	x	High (matches the schema of unstructured to structured)	Low (after linking it leaves the interpretation to humans)	Low (after linking it leaves the interpretation to humans)	Low (after linking it leaves the interpretation to humans)	Low (after linking it leaves the interpretation to humans)	Low (after linking it leaves the interpretation to humans)	Medium (uses the context of entities to integrate)

Integrating unstructured data into relational databases	Unstructured vs Structured	x	x	High (matches the schema of unstructured to structured)	High (uses an algorithm to counter syntax differences)	Unknown (example in paper is without measurements)	Unknown (no scales are used in the example)	High (checks for redundancy)	High (cardinality constraints are used from the database)	Medium (uses the context of entities in the db. and text to integrate)
Ontology guided information extraction from unstructured text	Unstructured vs Unstructured	x	x	High (matches the schema of unstructured sources)	Low (only extracts entities does not integrate it with different syntaxes)	Low (only extracts entities does not integrate it with different measurements)	Low (only extracts entities does not integrate it with different representations	Low (extracts all, the redundancy should be removed afterwards)	NA (there is no explicit cardinality in unstructured text)	High (solves the semantic issues by applying ontology

Unstructured information integration through data-driven similarity discovery	Unstructured vs Unstructured	x	x	High (matches the schema of unstructured sources)	Low (discovers similarity between text, but the user should do the final integration step)	Low (discovers similarity between text, but the user should do the final integration step)	Low (discovers similarity between text, but the user should do the final integration step)	Low (discovers similarity between text, but the user should do the final integration step)	Low (discovers similarity between text, but the user should do the final integration step)	Low (discovers similarity between text, but the user should do the final integration step)

Appendix 14 Interviews