Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Bridging the information gap of disaster responders by optimizing data selection using cost and quality

Marc van den Homberg[a,b,c,*], Robert Monné[b,c,d,e], Marco Spruit[d,1]

[a] 510 an initiative of the Netherlands Red Cross, Anna van Saksenlaan 50, 2593, HT, The Hague, the Netherlands
[b] Cordaid, Lutherse Burgwal 10, 2512, CB, The Hague, the Netherlands
[c] TNO, Oude Waalsdorperweg 63, 2597, AK, The Hague, the Netherlands
[d] Utrecht University, Utrecht, Princetonplein 5, 3584, CC, Utrecht, the Netherlands
[e] ORTEC, Houtsingel 5, 2719, EA, Zoetermeer, the Netherlands

ARTICLE INFO

Keywords:
Disaster management
Data fusion
Big data analytics
Decision making
Natural disaster
Information requirements
Humanitarian response
Data preparedness
Integer linear programming
Decision making

ABSTRACT

Natural disasters are chaotic and disruptive events, with compressed timelines and high levels of uncertainty. Comprehensive data on the impact becomes only available well into the response phase and data is scattered across organizations. Data heterogeneity issues are common. Consequently, responding organizations have difficulties finding data that match their information needs. We investigated the information needs of and the disaster management data available to both national and local decision makers during the 2014 floods in Bangladesh. We conducted 13 semi-structured interviews and three focus group discussions, collecting in this way input from 51 people, transcribed and coded them so that themes of information needs emerged. We mapped the information needs on the available data sets and determined which needs were not, partially or completely covered. We identified seven themes of in total 71 information needs and 15 data sets. The mapping revealed a significant information gap of timely and location-based data. Only 40% of the information needs are covered in time and 75% if no time constraints are considered. Instead of using all data sets, we optimized for coverage -with Integer Linear Programming-combinations of data sets against the costs of extracting data from structured versus unstructured data and against the quality in terms of timeliness, source and content rating and granularity. Without time constraints, three data sets yield already a coverage of 68%, whereas adding five extra data sets only gives an improvement of 7%. We recommend executing identification and mapping of available data sets on the information needs as part of Data Preparedness. Determination of the optimal combination of data sets can be used to extract data on information needs more efficiently. Currently, we did this manually, but future research will investigate automatic matching of information needs on data sets, by applying intelligent querying and semantic data matching.

1. Introduction

1.1. Background: information gap in humanitarian decision making

Natural disasters are inherently chaotic and disruptive events, with compressed timelines and high levels of uncertainty. Accurate and comprehensive data on the impact of a disaster becomes only available well into the response phase and is often a rough snapshot rather than a continuously updated detailed operational picture. In addition, data is scattered across organizations. Each organization will hold certain (baseline) data sets in relation to their mandate and will have their own focus in damage and needs assessments. Harmonizing and coordinating these assessments is a difficult task. Heterogeneity issues in the data sets that come out of the assessments are most commonly unavoidable. For example, in the case of the 2014 floods in Bangladesh, the several governmental and humanitarian organizations that worked together on a Joint Needs Assessment stated right after the floods arrived: "*Based on the information that was available for review it is difficult to get an overview*

[*] Corresponding author. 510 an initiative of the Netherlands Red Cross, Anna van Saksenlaan 50, 2593, HT, The Hague, the Netherlands.
E-mail address: mvandenhomberg@redcross.nl (M. van den Homberg).

[1] MvdH proposed the overall research question, wrote the manuscript, was involved in the data collection in Bangladesh and the interpretation of the data. RM developed the research design as part of his MSc thesis with support of MvdH and MS. RM performed the acquisition of the data in Bangladesh together with MvdH and labelled and coded the data. MS supervised the research and reviewed the manuscript. All authors were involved in reviewing and approving the final version of the manuscript.

https://doi.org/10.1016/j.cageo.2018.06.002
Received 31 May 2017; Received in revised form 16 May 2018; Accepted 5 June 2018
Available online 15 June 2018

of the flooding situation across the country because of the quality of information available and because of the differences in the collection, presentation and content of the information. In addition, most of the information is several weeks old." (CARE Bangladesh, 2014). Also baseline data, that could have been collected and collated beforehand, is in many cases not readily available. Early in the disaster data is lacking and later in the disaster there might be too much data. In the case of Typhoon Haiyan, those responding considered the multitude of different information sources and formats generally as an information overload (Comes et al., 2014). In both cases, information gap and overload, disaster responders cannot meet the information needs (Gralla et al., 2015) for the decisions they need to make. This leads to so-called cognitive or motivational biases in their decision making. A motivational bias is induced by the desire for a specific event or result to happen (Montibeller and von Winterfeldt, 2015). For example, local organizations might overreport to make sure sufficient relief items are going their way or national governments might put pressure on assessment teams to underreport to avoid having to scale up their response and funding. A cognitive bias is introduced when responders create a simplified mental model when dealing with complex problems such as a disaster (Comes, 2016). New approaches to bridge the information gap and reduce these biases are required.

1.2. New approaches in Small and Big Data for disaster management

Usually, data on the impact of a disaster consists of Small Data, i.e. data produced in a tightly controlled way using sampling techniques that limit their scope, temporality, size and variety (Kitchin and Lauriault, 2015), such as field surveys organized by an NGO or the government. In the digital age that we have entered, new technologies such as social media, mobile phone technology, the internet of things and satellite technology, create an exponential increase in the data that become available during a disaster (Meier, 2015). On the one hand, this is an increase of Small Data, as social media and the internet, such as geospatial data sharing platforms (Payne et al., 2012), make small data sets from a variety of stakeholders more widely available with an associated drive to harmonize with respect to data standards, formats and metadata (Kitchin and Lauriault, 2015). On the other hand, it is about the creation of new data, so-called Big Data, such as social media or transaction data (Whipkey and Verity, 2015). Big data is -in comparison with Small Data-of high volume and velocity, flexible, very exhaustive and fine-grained in resolution (Kitchin and Lauriault, 2015). New techniques and analytics are required to handle and extract useful information from both Big Data and -the increasing amount-of pooled and linked Small Data. With the increases in computing power, artificial intelligence (such as machine learning) and expert systems can automatically mine and detect patterns and build predictive models (Kitchin, 2014). For example, huge amounts of tweets can be filtered to find only the ones relevant for responders (Sangameswar et al., 2017; Vieweg et al., 2010). However artificial intelligence for disaster response (Imran et al., 2014) is still in its early stages and not yet ready for use in humanitarian operations (Tapia et al., 2011). Socio-technical solutions such as citizen coproduction of crisis communications (Chatfield and Reddick, 2017) and bounded microblogging (Tapia et al., 2011) will be required in parallel. Another important factor to consider is that both responders and affected communities have different levels of data literacy and access to digital infrastructure. The Data Poverty Index (Leidig and Teeuw, 2015) is a metric for evaluating variations in access to digital data and infrastructure and not surprisingly shows higher data poverties for poor countries. Digital inequality can amplify social inequalities, where those with no access to social media have less opportunities to prepare or recover from a disaster (Madianou, 2015), and where they also will leave no trace in Big Data.

The ever-increasing amount of Small and Big Data poses a severe challenge to the disaster responders. Small and Big Data are collected or created with specific and varying objectives in mind and hardly ever solely with the aim to address the information needs of the responders. Basically, disaster responders have to undertake a *knowledge discovery process* (KDP) (Fayyad, 1996) to be able to extract actionable knowledge out of these Small and Big Data sets. An implementation of KDP is the CRoss Industry Standard Process for Data Mining (CRISP-DM) method, which outlines a typical knowledge discovery process with specified sub-steps to successfully complete a data (mining) oriented project (Chapman et al., 2000a,b; Kurgan and Musilek, 2006). CRISP-DM can be considered part of the data layer, that feeds into the decision and business process layer. Horita et al. (2017) modelled how to align the business process and decisions to data sources by developing the observation-aware Decision Model and Notation+ (oDMN+). They focused on the business process of one specific organization (a government center for disaster management). In this research, we focus on the data layer and the matching between data and information requirements. We take a cross-organizational perspective and translate the first three phases of CRISP-DM (Business Understanding, Data Understanding and Data Preparation) to the disaster management domain. All activities in the Business Understanding phase are aimed at understanding the business requirements for the project, i.e. what kind of information does the decision maker require for effective decisions. The activities in the Data Understanding phase include the collecting, describing and exploring of data, plus a check on the data quality. Data Preparation focuses on selecting, cleaning and integrating the data. The first three phases of CRISP-DM can also be captured by the term Data preparedness as has been introduced to the humanitarian domain. Data Preparedness is defined by Raymond and Al Achkar (2016) as "the ability of organizations to be ready to responsibly and effectively deploy data tools before a disaster strikes" (p.3). van den Homberg et al. (2017) go beyond only the aspect of being ready to deploy tools and include the pre-staging of data. They define Data Preparedness as all activities, that can be done before a disaster hits, to pre-stage data with sufficiently high data quality (that matches the prospective information needs of responders) and to develop capacities to collect data on affected communities and areas once a disaster hits to ensure a timely, efficient, and effective response.

1.3. Objectives

This research investigates how to bridge the information gap disaster responders are facing so that they can take better decisions for a faster and more efficient humanitarian response. Three research objectives were defined following a Data Preparedness approach as explained in the previous section. First, regarding the inventory of information needs and data sets: What are the information needs of disaster responders so that they can take appropriate decisions? What are the associated timing constraints? What are available and relevant data sources and when do they become available? Second, with respect to mapping data sets on information needs: How do these data sources meet the information requirements? Third, towards integrating data sets: How can Data Sets be optimally combined to cover the most Information Needs? Our previous work (van den Homberg et al., 2018) addressed the first two questions mostly qualitatively. This paper extends the previous work by including the mathematical approach for the first two questions and by addressing the third research question. In addition, the research is placed in the context of ongoing research at the cross-over of data science and disaster management.

The next section describes the materials and methods used to address these research objectives. The case study on floods in Bangladesh is presented, followed by an explanation of how we calculate the coverage of information needs and how we optimize for costs and quality using Integer Linear Programming. The results section presents first the overview of the information needs and the datasets identified before describing the coverage and the optimal data selection. In the discussion, the results are compared and examined in relation to work and in the recommendations, we explain how the approach of identification

Table 1
Description of the sample for the semi-structured interviews and focus group discussions.

Actor	Administrative level	Role	Number of people involved	
			Semi-structured Interviews	Focus group discussions
Government	National	Disaster manager	1	
	Upazila	Project implementation officer Damage Needs Assessment	1	
	Upazila	Upazila Chairman	1	
	Upazila	Project implementation officer	1	
	Union	Union Chairman	1	
International Organization	National	Program Coordinator	1	
	National	Database specialist Joint Needs Assessment	1	
NGO	National	Disaster manager	1	
	National	Consultant Joint Needs Assessment	1	
	National	Disaster manager	1	
	National	Disaster manager	1	
	District	Disaster managers, director	3	
	District	Disaster responders		7
Various	District	Fisherman and farmer	2	
	Upazila	Upazila and Union Disaster Management Committee, Ansar and Village Development Party (VDP), Digital Entrepreneur		13
	Union	Disaster Management Volunteers, Iman, Teachers, Entrepreneurs		15
			16	**35**

and mapping of available data sets on the information needs as part of Data Preparedness in combination with optimization can be used also beyond the case study of Bangladesh. Finally, future research directions are presented including automatic matching of information needs on data sets, by applying intelligent querying and semantic data matching.

2. Materials and methods

2.1. Case study on floods in Bangladesh

This section describes the case study using the consolidated criteria for reporting qualitative research (Tong et al., 2007). A practice-oriented research approach was selected (Verschuren et al., 2010; Monné, 2016), as the main objective is to develop a solution for the real-world problem NGOs such as Cordaid are facing. We selected a case study on hydrometeorological hazards, and more specifically river floods, as they are the most frequently occurring type of natural disaster (Guha-Sapir et al., 2016) that affect more people globally than any other type of natural hazard. Bangladesh is well known as one of the most flood prone areas of the world. About one-fifth to one third of the country is annually flooded by overflowing rivers during the heavy rainfall of the monsoon (June to September). While normal floods are considered a blessing for Bangladesh providing vital moisture and fertility to the soil, moderate to extreme floods are of great concern, as they inundate large areas (more than 60% of the country is inundated in large flood events), and cause physical damages to agricultural crops, buildings and other infrastructures, social disruptions in vulnerable groups, livelihoods and local institutions, and direct and indirect economic losses. The flood hazard problem in recent times is getting more and more frequent and acute due to growing population size and human interventions/socio-economic activities in the floodplain at an ever-increasing scale (Mozzammel Hoque, 2014). The case study consists of the most recent and severe river flood of the last years, namely the floods of 2014 that from mid-August onwards affected almost two million poor and vulnerable people living in nine districts in North West Bangladesh (Wahed et al., 2014).

We performed 13 oral history semi-structured interviews of which 11 in Dhaka (national NGOs (active in the JNA consortium) and Department of Disaster Management) and two in Sirajganj (one with a farmer and fisherman, and one with the director and his two co-directors of the local NGO, MMS), see Table 1. We held one focus group discussion with seven disaster responders of MMS, one focus group with 15 people living on the chars (imam, teachers, entrepreneurs, part of the volunteer disaster management committees) and one focus group with 13 local government officials (Upazila and Union Disaster Management Committee,[2] civil defense organization (Ansar Village Development Party)). So, in total we got input from 51 people. Both the focus group discussions and the interviews lasted about 1 h and were held in April 2015, so about eight months after the floods. We arranged the first batch of interviewees based on our existing network (purposive sampling) via emailing and calling and such that we would have a representative cross-section. Subsequently we used a snowballing approach once in-country to grow our sample considering availability of respondents and useful references. Although focal point in these sessions was the flooding of 2014, we did allow interviewees also to draw from their earlier or more recent disaster management experiences. All interviews were transcribed. The focus group discussions were done with an interpreter, usually at an open noisy market place and could not be literally transcribed. Instead we used the notes taken.

All interviews and notes were subsequently labelled using NVIVO 10 for Windows and coded based on three themes, i.e. *Activity*, *Decision* and *Information Need*. We used inductive coding to have subthemes emerge from the data. For each of these themes clustering was done based on experience emerging from the familiarization phase, domain knowledge and literature study. In addition, we asked the interviewees to validate our transcribed interviews. We asked two domain experts to validate and expand on the list of needs. The domain experts had between ten and fifteen years of experience in humanitarian response, with a focus on information management, and up to ten deployments to disaster affected areas, including Bangladesh. We also used the lists of *Activities* and *Decisions* as a way to identify possible discrepancies. To obtain the Data Sets, we used in addition to the interviews, internet search and literature study. In that way, we could make an inventory of the data sets that were available during the flooding of 2014.

2.2. Calculating coverage of information needs

To be able to determine which dataset covers the most Information Needs, we calculate a Coverage parameter P_s for each Data Set $s \in S$ by summing the multiplication of the Weight W_i with the match $P_{s,\,i}$ for each subtheme Information Need $i \in I$:

[2] Bangladesh is at the local government level administratively divided into divisions, districts, Upazila, unions, wards and villages.

$$Ps = \sum_{i} W_i \cdot P_{s,\,i} \tag{1}$$

We single out all the indicators per data file s and manually determine the match $P_{s,i}$ with a subtheme information need $i \in I$, i.e. Yes (1), No (0) or Partly (0,5). If for each of the subtheme information needs, the match with the indicators in the Data Set is 1, then Coverage P_s results to 100%. For each subtheme information need $i \in I$ that has no corresponding indicator in the Data Set s, the coverage-score P_s decreases.

To be able to discriminate between the importance of different information needs, we introduce the weight parameter. We can use three ways to determine the weight $W_i \in [0{,}1]$:

(1) The weight of subtheme information needs towards their corresponding higher-level theme is always divided equally. Each theme has the same weight, but an indicator under a theme with more indicators will have a lower weight than an indicator under a theme with less indicators.
(2) Each subtheme information need gets the same weight, i.e. 1/71. This gives more weight to a theme that has more subtheme information needs, which might be justified given that responders mentioned more subtheme information needs.
(3) Each actor can assign a specific weight to a specific information need, where for example damage to houses might be more relevant than damage to public buildings for an NGO than for a government actor.

We can further refine Coverage by adding a timing aspect and granularity level.

Timing: We used approximately the phases as defined in the Multi-Sector Initial Rapid Assessment (MIRA) (MIRA, 2015) to label both the data sets as well as the information needs. The phases consisted of before (1), the first 72 h (2), the first two weeks (3) and the first two months (4). Table 2 gives an example for Data Set with and without timing. Data Set A covers 100% of the information needs if no time constraints are taken into account. With time constraints, only 50% of the information needs is met, since the information need A and C required data to be available already in phase (1), whereas Data Set A became only available in phase (2).

Table 2
Example of mapping Data Sets on Information Needs as a function of timing.

	Timing	Data Set A (no time constraints)	Data Set A (with time constraints)
Timing			2
Information need A	1	Yes	No
Information need B	2	Yes	Yes
Information need C	1	Yes	No
Coverage		100%	33%

Granularity level: Each Data Set contains data collected at a certain administrative level or in some cases at multiple administrative levels. Information needs can also be at a very local or at a more aggregated level depending on the decision-making purpose they have to serve. For example, the number of damaged houses must be known at the lowest administrative level when distributing housing repair kits, whereas it will be useful at a more aggregated level for acquiring donations. Therefore, Coverage can be calculated for each administrative level up to the deepest available granularity level for a specific Data Set. For example, for the Joint Needs Assessment, we can calculate Coverage at District, Upazila and Union level, but not at ward level as at that level no data is available.

2.3. Data sets input optimization using Integer Linear Programming

The next step is to select a combination of Data Sets with the highest coverage of Information Needs, while keeping in mind the relative importance (hence weight) W_i of the information needs, the costs C_s of adding a dataset and the quality of the dataset Q_s. We chose to model this complex optimization problem with the Advanced Interactive Multidimensional Modelling System (AIMMS) software, which can be used for Integer Linear Programming (ILP) as follows:

$$\max \sum_{\mathrm{i} \in \mathrm{I}} \sum_{s \in S} W_i \cdot Q_s \cdot P_{s,i} \cdot y_{s,i} \tag{2}$$

Subject to:

$$\begin{aligned} &\textstyle\sum_{s \in S} y_{s,i} \le 1 \;\; \forall \;\; \mathrm{i} \in \mathrm{I} \\ &y_{s,i} \le x_s \;\; \forall \;\; i \in I, \;\; \forall \;\; s \in \mathrm{S} \end{aligned} \tag{3}$$

$$\begin{aligned} &\textstyle\sum_{s \in S} C_s \cdot x_s \le B \\ &x_s \in \{0{,}1\} \;\; \forall \;\; s \in S \\ &y_{s,i} \in \{0{,}1\} \;\; \forall \;\; s \in S, \;\; \forall \;\; i \in I \end{aligned} \tag{4}$$

The ILP problem uses parameters (input data) and variables to reach a solution. Below we explain the parameters and variables of the model that were not yet explained in the previous section:

2.3.1. Parameters

2.3.1.1. Cost of a data set. Cost of a dataset $C_s \in \mathbb{R}^+$ (set of all positive real numbers) is determined by how easy or difficult it is to extract data on the information needs from the dataset. A relative low cost is involved for a structured file (such as Excel or a relational database) and a high cost for an unstructured file (such as PDF or Word). The budget available to cover for the cost to select and use a source is set by parameter $\mathrm{B} \in \mathbb{R}^+$.

2.3.1.2. Quality of a dataset $\boldsymbol{Q_s}$*.* As some datasets have a higher quality than others, we would like to give these datasets priority. To do so, we introduce the quality parameter Q_s. Quality of a Data Set: $Q_s \in [0{,}1]$, where 1 is the highest score. Quality is determined by combining the scoring of four components:

(a) *Timeliness*: Timeliness is a combination of when the data set was last updated and how long a data set remains representative of the reality (van den Homberg et al., 2017; de Vries, 2002). How long a data set s remains representative can also be termed *retention period*$_s$ and this differs per information need. For example, geographic boundaries of lower administrative units can generally considered to be stable for a long period (several years). However, needs of those affected can change within weeks. Mathematically, *Timeliness*$_s$ is defined as follows for Data Set s:

$$Timeliness_s = Max\left(\frac{retention\ period_s - time_{passed}}{retention\ period_s},\, 0\right) \tag{5}$$

The first part of the formula resolves to 1 if $time_{passed} = 0$. When $time_{passed} \ge retention\ period_s$ the $Timeliness_s$ score resolves to 0. We note that the *retention period*$_s$ has a relationship with the phases (1) to (4) that were introduced in the previous section. For example, information needs in the first 72 h, such as people in need of rescue, have a very short retention period.

(b) *Source reliability* [A to F]: where an A score means a reliable source, where there is no doubt of authenticity, trustworthiness or competency; has a history of complete reliability (US Intelligence, 2017). It is important to properly distinguish which organization is the source. IASC defines three categories of organizations when it comes to how datasets are governed (IASC Guidelines, 2010):

- Guardian is responsible for facilitating distribution of datasets and information products (in emergencies for example).
- Sponsor is responsible for identifying and liaising with relevant sources to analyze, collate, clean and achieve consensus around a

Table 3
List of information needs.

CRISIS IMPACT	OPERATIONAL ENVIRONMENT
Baseline context	**Coordination**
Livelihoods	Coordination groups at local and national level
Vulnerabilities	Response Activities NGOs and government
Hazard identification (location, timing)	*Response activities private sector*
Socioeconomic context	Community leaders
Political (local governance) and religious context	Gap analysis between capacities and needs
Community Preparedness (such as Security/evacuation plans)	Presence of NGO workers
Preparedness of people	Staff skills
Village and ward boundaries (location of households)	Telephone numbers
Damage and needs	*Communication channels*
WASH needs	*Incidents registration*
Health needs	*Evacuation routes*
Education needs (closed schools)	**Capacity**
Food security needs (stoves, firewood)	Stock of emergency items
Shelter needs (including non-food items)	Coping mechanisms of affected communities
Needs of subgroups (elderly, children)	Local agricultural and fishery situation
Number of people affected	Local market situation
Livestock affected	Institutional capacity
Type of damage to houses	Staff skills and training
Number of damaged houses	Burying strategies
Number of destroyed houses	**Service locations (during the flooding)**
Losses of private belongings	Shelters for humans
Number of people dead	Shelters for cattle
Number of people injured	Doctors
People in need of rescue	Medicine distribution points/shops
Submerged houses	Food buying and selling places
Damage to infrastructure	Labor opportunities
Damage to health facilities	Drinking water locations
Damage to public buildings	Emergency items
Affected medical personnel	*Meeting points*
Number of people saved	*Pickup points*
Displaced people	**Security and access**
Impacted area	News
Flood situation	Accessibility
Flood news	*Security*
Flood duration	*Mobile phone coverage*
Earlier predictions	
Time of inundation	
Inundated area	
Drainage and irrigation systems	
Flood trend analysis	
Water quality	
River embankment erosion	

specific dataset or information product.

- Source: Designated source or owner of a dataset, fully responsible for the development, maintenance and metadata associated with a dataset and control distribution restrictions. We note that this position is highly similar to the Steward role in the related field of data warehousing (e.g. Kimball, 1998).

(c) *Content accuracy* [1 to 6]: A 1 score corresponds to Confirmed by other independent sources; logical in itself; consistent with other information on the subject (US Intelligence, 2017).

(d) *Granularity*: Quality of a Data Set is considered higher if the Data Set goes down to a lower administrative level. We note that for ILP, we have not calculated coverage $P_{s,i}$ as a function of granularity, but rather calculated an overall coverage.

2.3.2. Variables

To determine the combination of datasets that adds the most value, we introduce two binary variables $x_{s,i}$ and y_s which the software uses to reach a solution (see Formula 3 and 4). $x_s \in \{0,1\}$ indicates whether a dataset $s \in S$ is used (1) or not (0). $y_{s,i} \in \{0,1\}$ indicates whether dataset $s \in S$ is used for information need $i \in I$ (1) or not (0).

The ILP can be solved for different values of B (budget) to realize a scenario (where each scenario contains a selection of data sources given the constraints above). The ILP produces output which is analyzed to select the best scenario given the other requirements emerging from the context of this problem, further described in the result section.

3. Results

3.1. Information Needs

A small group of interviewees, especially at the local level, had difficulties expressing their information needs and identifying the type of decisions they had to take when directly asked for it. However, when interviewees where asked to describe their role in the flooding of 2014, it was possible for us to derive these. Information needs varied as well from one responder to the other, which could usually be attributed to differences in the organization they were working for, their specific expertise and level of education. Table 3 summarizes the needs as emerged from the coding and clustering of the transcribed interviews in normal text. The list is not exhaustive given our limited sample size. In italic text we have added the needs that two domain experts contributed. We decided not to aggregate the information needs to a too large extent, given that we want to map later on the information needs to the information in the available data sets, but also to reflect the needs as they were expressed. We defined seven clusters for in total 71

Table 4
Data and information products. Source Reliability is given on a scale of A (reliable) to E (unreliable), Content Accuracy on a scale of 1 (confirmed) to 5 (improbable). Costs range from 500, 1000 to 1500 (arbitrary units), going from structured to unstructured data sets. Overall score for the Quality is a score between 0 and 1, determined by taken the qualitative and quantitative scores on the four attributes of Quality together.

Data Set & Information Product	Quality						Costs		
	OVERALL QUALITY	Timeliness		Source reliability	Content accuracy	Granularity	OVERALL COSTS	Type	Level of structuredness
		Date of source	Retention period						
Geonode WFP (mostly maps)	0.7	Multiple	Varies, usually a few years	A	1	Multiple	1500	Multiple	Unstructured
Joint Needs Assessment.	1	Sep-2014	A few weeks	A	1	Union	1000	Excel	Semi-structured
News on websites	0.5	Daily during floods	A few weeks	B	2	Multiple	1500	Word	Unstructured
D Form	0.9	Aug-2014	A few weeks	B	2	Ward	1000	Excel	Semi-structured
4 W Database (Who's doing what where and when)	0.9	Monthly	A few weeks	A	2	Upazila	500	GIS file	Structured
Situation Report 20th of August 2014 (makes use of D-form)	0.9	Aug-2014	A few weeks	A	2	Upazila	1500	PDF	Unstructured
District Disaster Management Plan	0.9	Jul-2014	A year	A	1	Mostly at Upazila	1500	PDF	Unstructured
Pre-disaster secondary data assessment	0.8	Mar-2014	From one year up to a few years	A	1	Mostly national, some at Upazila	1500	PDF	Unstructured
Disaster Incident Database (on past disasters)	0.7	After disaster	N/a	A	1	Multiple	500	Relational database	Structured
Hazard map and predicted inundation map.	0.7	Pre-disaster	Lead time of forecast	A	1	Upazila	500	GIS file, JPEG	Structured
Union fact sheet	0.8	Multiple	A few years	B	1	Union	1500	PDF	Unstructured
Flood bulletin (including forecasts, inundation and rainfall maps)	1	Daily during floods	Lead time of forecast	A	1	Union	500	Excel file, Word	Structured
National Census data	0.8	2011	A few years	A	1	Ward	1000	Excel file, Word	Semi-structured
Flood shelter list	0.9	2008–2012	A few years	A	1	Union	500	Excel file	Structured
National Water Resources Data	0.7	Multiple	Varies	A	1	Multiple	1500	GIS data	Unstructured

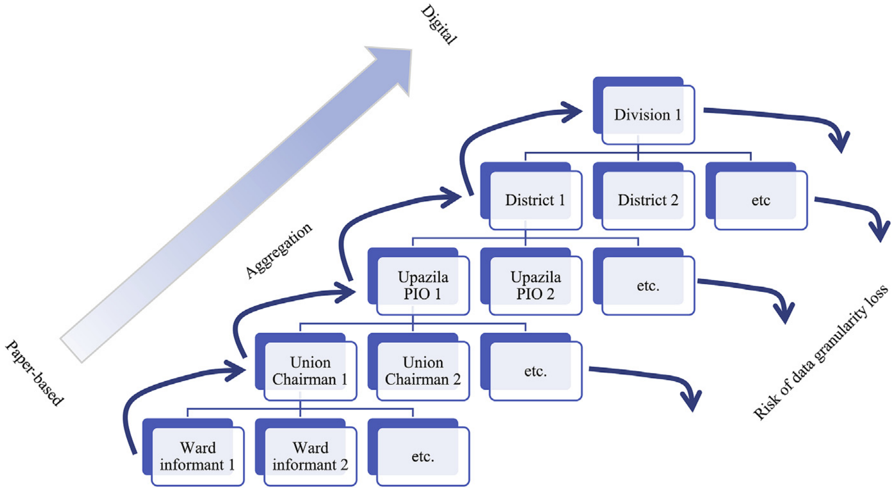

Fig. 1. Data granularity loss throughout the government hierarchy.

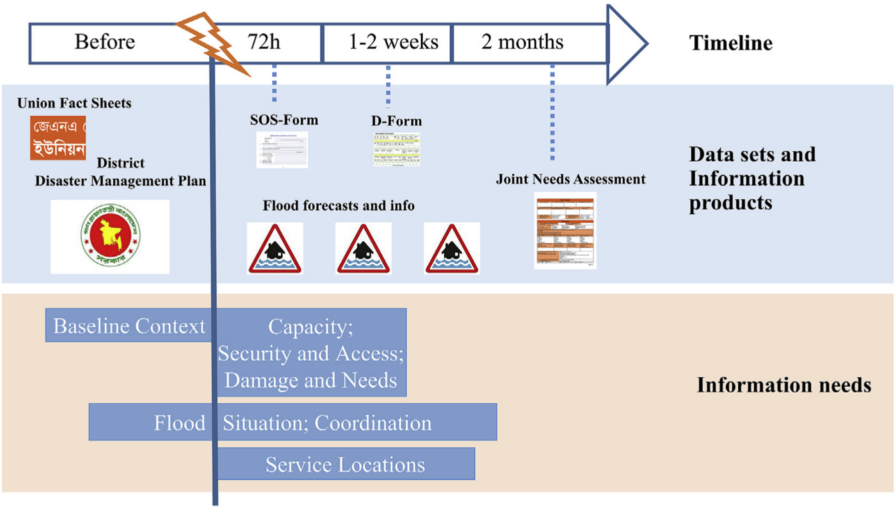

Fig. 2. Timing of information needs and availability of data sources. There is an increasing level of detail, quality and accuracy of the data becoming available.

information needs. We have put in Table 3 on the left clusters that relate to the Crisis Impact and on the right those that relate to the Operational environment, in line with the MIRA Analytical Framework (MIRA, 2015). The cluster *Damage and needs* scored highest in terms of amount of times mentioned in all interviews and in terms of in how many interviews it was mentioned. This cluster of information needs matched also with what the interviewees mentioned as the most difficult decisions for them to take, i.e. determining which beneficiaries to support where and with what kind of support. Next comes the need for information around *Coordination*, especially among government and NGOs. Specifically, it was mentioned in many interviews that it was important to have a gap analysis between the capacities available and the needs to be fulfilled. *Capacity* encompasses the response capacities of the responding communities and professionals and the coping capacity of the affected community. Knowing how to protect one's livelihood (such as agriculture, fishery and hand looming) increases the coping capacity. Interviewees mentioned for example the importance of knowing when to harvest just before the flood arrived and which crop to cultivate when the flood started to recede. Similarly, it was important to know how well the local market was still functioning. Key is also a readily accessible and suitable emergency stock (IFRC Emergency Items Catalogue, 2016). Specifically, information about boat capacity was mentioned as a need in the interviews. The *Baseline* cluster focuses on the context of people before the disaster hit. *Flood news* groups the needs in relation to the arrival and duration of the flood. The *Location Services* cluster refers to locations for essential services such as water, health, food and shelter, but also to places where there are opportunities for labor. *Security and access* refer to access for the responders to the affected community.

3.2. Data Sets

Table 4 shows the different data sets and information products. For example, for the Joint Needs Assessment there is a data set (an extensive excel file that compiles answers to 62 questions for several unions) and an information product (a pdf report discussing and describing the survey results). It also describes the collection and aggregation level of the data. For example, the Project Implementation Officers (PIO) phone representatives from different wards and aggregate the data they get per ward into a consolidated damage needs assessment form for their Upazila. The way the data is aggregated does not allow to go back to the ward level. This type of data granularity loss we encountered in more data sources. Fig. 1 depicts this data granularity loss, where the downward pointing arrow symbolizes data granularity loss at each step up in the local government hierarchy.

Important data providers are the Department of Disaster Management of the Government of Bangladesh and the Humanitarian Coordination Task Team (HCTT), consisting of UN, NGO and government representatives. For each file, we singled out all the indicators and determined the data type (Excel sheets, Relational databases, PDF, Text, Websites and Geographic Information). Fig. 2 depicts the different flood related data sets and information products and at which point in time they were collected and became available. The data sets and information products contain on the order of 40–60 indicators per source.

3.3. Mapping data sets on the information needs

Following the methodology described in section 2.2 we mapped the 71 information needs on the 15 data sets and information products as can be seen in Tables 4 and 5.

As explained in formula x, we could also calculate a coverage per granularity level, given that especially in an information product, the granularity level for each indicator can differ. For example, the District Disaster Management plan has data mostly at district level, but also -for certain indicators-on union level. However, we observed that commonly most of the data in a Data Set is aggregated to one administrative level (ward, union, upazila, district), so we decided not to do a coverage analysis at granularity level.

Overall, we see that 49 information needs are fully covered and eight information needs (that were not fully covered) are partly covered by one or the other data set. This means that 75% of the information needs is covered. We can draw the following conclusions per theme level. *Service locations* is not well covered at all with a cumulative

Table 5
Matrix mapping Data Sets and Information Products on Information Needs.

	Timing	Total coverage in data sources	Partly coverage in data sources	Total coverage given time constraints	Partly coverage given time constraints	JNA	D-Form	Geodash	DMIC portal - 4W DB	DMIC portal - Situation Report 20th of August 2014	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
Timing						4	3	1	1	3	1	1	1	1	1	1	1	1	1	3
Totally covered information needs						38%	34%	0%	0%	13%	20%	7%	0%	1%	4%	8%	1%	1%	4%	14%
Partly covered information needs						6%	4%	4%	6%	3%	17%	1%	0%	0%	0%	0%	3%	3%	1%	0%
Totally covered information needs given time constraints						0%	0%	0%	0%	4%	20%	7%	0%	1%	4%	8%	1%	1%	4%	4%
Partly covered information needs given time constraints						0%	1%	4%	6%	1%	17%	1%	0%	0%	0%	0%	3%	3%	1%	0%
BASELINE CONTEXT		**18.3%**	**5.8%**	**16.7%**	**4.2%**															
Livelihoods	1	5	2	4	2	No	Yes	Partly	No	No	Yes	Yes	No	No	Yes	No	Partly	No	Yes	No
Vulnerabilities	1	2	1	2	1	No	No	Partly	No	No	Yes	Yes	No	No	No	No	No	No	No	No
Hazard identification	1	4	0	3	0	Yes	No	No	No	No	No	No	No	Yes	No	Yes	No	No	Yes	No
Socioeconomic context	1	5	3	5	1	Partly	Partly	Partly	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No
Political and religious context	1	3	0	3	0	No	No	No	No	No	Yes	Yes	No	No	Yes	No	No	No	No	No
Community preparedness	1	2	0	2	0	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No
Preparedness of people	1	1	0	1	0	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
Village and ward boundaries	1	0	1	0	1	No	No	No	No	No	No	No	No	No	No	No	Partly	No	No	No
DAMAGE AND NEEDS		**13.3%**	**2.6%**	**0.0%**	**2.0%**															
WASH needs	2	2	1	0	1	Yes	Yes	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Health needs	2	2	1	0	1	Yes	Yes	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Education needs (closed schools)	2	2	1	0	1	Yes	Yes	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Food security needs (stoves, firewood)	2	2	1	0	1	Yes	Yes	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Shelter needs (including non-food items)	2	2	2	0	2	Yes	Yes	No	No	No	Partly	No	No	No	No	No	No	Partly	No	No
Needs of subgroups (elderly, children)	2	0	3	0	1	Partly	Partly	No	No	No	No	Partly	No	No	No	No	No	No	No	No
Number of people affected	2	4	0	0	0	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
Livestock affected	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Type of damage to houses	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Number of damaged houses	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Number of destroyed houses	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Losses of private belongings	2	3	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes
Number of people dead	2	3	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes
Number of people injured	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
People in need of rescue	2	1	0	0	0	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Submerged houses	2	3	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes
Damage to infrastructure	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Damage to health facilities	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Damage to public buildings	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Affected medical personnel	2	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Number of people saved	2	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Displaced people	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Impacted area	2	4	0	0	0	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
FLOOD SITUATION		**12.6%**	**0.7%**	**8.9%**	**0.7%**															
Flood news	3	3	0	3	0	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes
Flood duration	3	3	0	3	0	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes
Earlier predictions	1	1	0	1	0	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No
Time of inundation	3	3	0	3	0	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes

(*continued on next page*)

Table 5 (*continued*)

Inundated area	2	4	0	0	0	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
Drainage and irrigation systems	3	1	0	1	0	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
Flood trend analysis	3	1	0	1	0	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No
Water quality	3	0	1	0	1	No	No	No	No	No	No	No	No	No	No	No	No	No	Partly	No
River embankment erosion	2	1	0	0	0	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
COORDINATION		**4.8%**	**4.2%**	**3.6%**	**3.0%**															
Coordination groups at local and national level	1	1	3	1	1	Partly	No	No	Partly	Partly	Yes	No	No	No	No	No	No	No	No	No
Response activities NGOs and government	3	1	2	1	2	No	No	No	Partly	Partly	Yes	No	No	No	No	No	No	No	No	No
Response activities private sector	3	0	1	0	1	No	No	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Community leaders	1	2	0	1	0	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No
Gap analysis between capacities and needs	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Presence of NGO workers	3	1	0	1	0	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
Staff skills	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Telephone numbers	2	2	0	1	0	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No
Communication channels	2	0	1	0	1	No	No	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Incidents registration	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Evacuation routes	2	1	0	1	0	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No
CAPACITY		**4.8%**	**6.7%**	**1.0%**	**5.7%**															
Stock of emergency items	2	0	1	0	0	Partly	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Coping mechanisms of affected communities	2	2	2	1	2	Yes	No	No	Partly	No	Yes	No	No	No	No	No	No	Partly	No	No
Local agricultural and fishery situation	3	1	2	0	2	Yes	Partly	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Local market situation	3	1	1	0	1	Yes	No	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Institutional capacity	3	1	1	0	1	Yes	No	No	Partly	No	No	No	No	No	No	No	No	No	No	No
Staff skills and training	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Burying strategies	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
SERVICE LOCATIONS (DURING THE FLOODING)		**0.7%**	**2.2%**	**0.7%**	**2.2%**															
Shelters for humans	2	1	1	1	1	No	No	No	No	No	Partly	No	No	No	No	No	No	Yes	No	No
Shelters for cattle	2	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Doctors	3	0	1	0	1	No	No	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Medicine distribution points/shops	3	0	1	0	1	No	No	No	No	No	Partly	No	No	No	No	No	No	No	No	No
Food buying and selling places	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Labor opportunities	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Drinking water	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Emergency items	3	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Meeting and pickup points	2	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
SECURITY AND ACCESS		**8.3%**	**0.0%**	**0.0%**	**0.0%**															
News	2	2	0	0	0	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes
Accessibility	2	2	0	0	0	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
Security	2	0	0	0	0	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Mobile phone coverage	2	1	0	0	0	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No

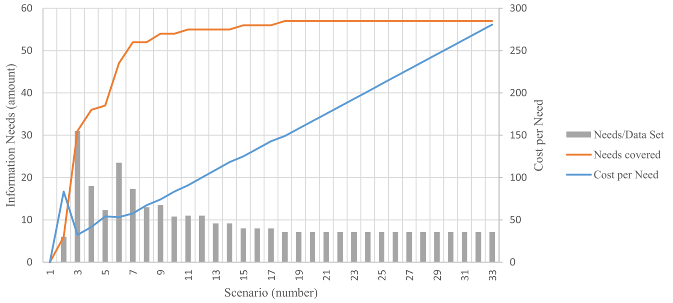

Fig. 3. Optimizing different combinations of data sets for cost, quality and coverage according to weights subdivided to the higher information need theme.

coverage of 0.7%.[3] One of the reasons for this might be that information is often collected by phoning people and asking them to give an overview for their ward or by conducting a paper-based survey. We did not come across local responders that use an app or GPS to map locations during the floods. *Capacity* was also not well covered, varying from relatively easily to monitor capacities such as the number of boats up to the more difficult to assess coping mechanisms of affected communities. *Damage and needs* were reasonably covered largely by only two out of the 15 data sets (JNA and D form). The following data sources match well the information requirements: JNA (38%), D-form (34%), District Disaster Management Plan (20%), the (online) News (14%) and the Situation Report (13%). We note that the 13% is based on the first Situation Report that became available; later ones yield a higher coverage. We also must take into account that these data sources overlap on some indicators (and hence also their coverage of information needs), and the coverage percentage cannot be summed.

If we look at the timing constraints, then it becomes apparent that most operationally related information is not available in time. In this case, 28,5 information needs are covered, or in other words 40%. It is also known from literature that certain response information has to be available within 48 or 72 h after the disaster strikes (Meier, 2015).

3.4. Optimizing for coverage, quality and costs

We determine which combination of data sets yields an optimal coverage against quality and costs for four different parameter sets. The parameters that we vary are (1) without or (2) with timing constraints, where we only take those data sets into account when it is available in time to meet the information needs, (3) the same weights for all information needs or (4) weights subdivided according to the higher theme as explained in 2.2. The costs and quality are given in Table 4. We gave values of 500, 1000 or 1500 (arbitrary units) to Costs of the Data Sets, going from structured to unstructured. Overall score for the Quality is a score between 0 and 1, determined by taken the qualitative and quantitative scores on the four attributes of Quality together. We did not use the *Timeliness* formula to calculate the overall quality score, as the *retention period* was only known within a broad bandwidth.

The effect of changing the weights without timing constraints does not result in a significant difference. In both cases, three data sets, i.e. JNA, District Disaster Management Plan and the Flood bulletin, yield a coverage of 68%. Adding five extra data sets only gives an improvement in coverage of 7%, equivalent to respectively 2, 1, 0, 1 and 1 additional covered information needs per added data set. An important result is hence that a very good coverage of information needs can already be reached by the three most important data sets out of the total 15. The increase of 0 information needs whilst still adding a data sources can be explained by an increase in for example overall quality of the selection.

Fig. 3 shows the results for weights subdivided according to the higher theme and with no timing constraints. The x-axis shows the different scenarios, where a scenario is a combination of data sets. For example, scenario one corresponds to one data set and scenario 33 to eight data sets. There are in total 33 scenarios equivalent to one

[3] Defined as the number of times there is total coverage of one of the service location information needs in one of the data sources divided by (the number of information needs within service locations) x (the number of data sources).

scenario as a starting point and 32 adding up to the total costs as given in the column costs in Table 4. The blue curve shows the increase of budget B needed (y-axis on the right). The orange curve shows the number of needs covered (y-axis on the right). The grey column diagram represents needs/data set and shows that scenario 3 (equal to the JNA data set) and scenario 5 (the three data sets mentioned before) render the most needs per data set. The needs covered curve is quickly reaching a plateau afterwards, whereas the costs keep on linearly increasing. Integrating additional data sets beyond scenario 6 is hence not easily justified, given that the budget required for an additional integration of a data set does not or barely result in an increase in coverage. However, a case could be made to select scenario 7, which adds another 5 information needs and only requires addition of 1 source.

We have to remember that the above results are without taking timing constraints into account. If we do take these into account, we can redo the same calculation also for 33 scenarios but without JNA as a data set, since JNA becomes available too late for all information needs. In this case, we can get a coverage of 45% by using the District Disaster Management Plan and the Flood bulletin. Also in this case, we can get a marginal improvement of 5% by adding five data sets.

4. Discussion

We compared our framework of information needs with the one from Gralla et al. (2015). *Context and scope of the disaster*, *Coordination* and the *Humanitarian Needs* are themes which are overlapping, and which are the most important factors in the earlier response. Several other information requirements are not mentioned in our interviews such as *Relevant laws and policies* as part of *Coordination* and *Looking forward*. The Gralla et al. framework emerged from consultation with mostly responders from the international humanitarian community, whereas our framework emerged from consultation with only national and local responders. Also, the type and scale of disasters looked at was different. We looked at small scale disasters, whereas Gralla et al. focused on large scale disasters, where international response is requested by the nation affected. Floods as in our case study have severe impacts on livelihoods but usually less in terms of loss of life. In many cases, there can be also a difference of opinion between the NGOs on the one hand and the government on the other hand as to whether to declare a flood an official disaster. One interviewee mentioned encountering in some cases political pressure to underreport the impact of a disaster. Nevertheless, it is widely acknowledged that the role of national and local responders is of utmost importance also in large scale disasters. Local responders have more local context knowledge and -in case of recurring disasters like the annual floods in our case-they also usually have more response experience than the international community. This leads to a different level of information needs regarding the *Baseline* theme between local, national and international responders (Van Den Homberg et al., 2014).

For international responders, the *Public and media perception* turned out to be a separate theme. In our interviews media perception did not come forward as an important issue, probably related to the fact that national and local responders usually are not directly applying for funding themselves (but through their supporting international NGOs) and that the local communities affected usually do not have access to a lot of media channels. We did not find much information needs in relation to *Recovery*. This might have to do with the relatively limited possibilities for the responders in our interview group to extent their activities beyond response. In sum, it is important for each type of context and hazard to develop a tailor-made information needs framework. We have now developed a first version for a hydrological hazard in one of the poorest countries in the world. A comprehensive framework with a generic set of themes can be used as a starting point and for each actor there will be differences as to which category is the most important to them given his or her organizational mandate, where for example some NGOs focus on women empowerment and others on disability. Such a comprehensive framework should include both the local, national and international perspective. These tailor-made frameworks can be used as input to the ILP modelling to select an optimal set of sources for one specific responder, but also for all involved responders, weighing everybody's information needs equally. This can lead to a better coverage for all involved.

Subsequent mapping of available data sources on the information needs in the framework is key for identifying the data gaps that currently exist. It is clear from the mapping we did that both the responding and the professional community lack information to effectively dimension and target their response. Data that was available before and during the floods can be largely characterized as Small Data, although some of the data and information products were based on satellite data and can hence be termed Big Data. We did not include other Big Data sources such as Call Detail Records or social media data. Currently, social media penetration among the affected communities is still very low. In the whole of Bangladesh, the percentage of people using the internet is 9.6% in 2014 (ITU Statistics, 2015) and most of these are in dense urban areas. Mobile phone use is more elevated, and analysis of Call Detail Records can be insightful to supplement survey-based data, as a study by Lu et al. (2016) demonstrated by unveiling mobility patterns in climate stressed regions. However, both access to this kind of data and its analysis is still complex and near real-time use during a response for the time being impossible. Our approach of characterizing a Data Set in terms of Cost and Quality will make these considerations transparent as the Costs of extracting information from a Big Data set will be higher than from a Small Data set. We did not yet include in our analysis data sets or information products that were available only in Bengali. However, to our current understanding, based on the interviews and our literature and internet search, this seems to be a minor fraction.

The government works with the SOS form and D form for damage and needs assessments. The D form has 30 questions which are –usually without clear guidelines-filled in by the Union secretary/chairman for on average 5000 to 6000 families, based on very little or no capacity in the field of sampling, data collection, and recording. The system is still largely a paper-based system, whereby forms are manually summarized at each of the administrative levels, before they are passed on to central level, leading to the granularity loss described before. NGOs that are part of the Local Consultative Group often do their own assessments, such as in 2014 via a Joint Needs Assessment, creating in fact a new process with different indicators that is only aligned with the government process to a very limited extent. Once the information is collected at central level, support is mobilized for the response, making the response largely a top-down mechanism. Both the NGO and Government information architecture are not specifically geared towards coordination and action planning at Community, Union and Upazilla level, forming a stumbling block for effective local response. The same observations hold true in many other developing countries, given that a digital transformation of government branches involved in disaster management is especially at the lower administrative levels not yet occurring due to a lack of resources and capacity.

The framework can be implemented as part of an overall data preparedness framework (van den Homberg et al., 2017). First, we propose to organize regular multi-institutional mapping cycles of data sets on information requirements. These cycles should not only consist of keeping an up-to-date inventory of available data sources and providers, but also of regular consultations with responders as to what their information needs are. When the interviewees validated the information needs framework, this sparked their creativity. We got reactions like: “wow, if this is possible, we could also really benefit from X information”. It is important hence to keep on evolving the requirements and to use these requirements to shape the information products that providers are creating so that they meet the decision maker's needs.

Secondly, coordination needs to be improved in the data ecosystem in which humanitarian responders operate. Hereby, an ecosystem is defined as "the people and technologies collecting, handling, and using the data and the interactions between them" (Parsons et al., 2011: p.557). There is an emerging literature on how to characterize and govern data ecosystems. Haak et al. (2018) developed a framework of criteria for a successful data ecosystem specifically for humanitarian purposes, including data supply, user characteristics and governance criteria. Instead of the overarching data ecosystem approach, one can also segment a data ecosystem into data collaboratives, i.e. cross-sector (and public-private) collaboration initiatives aiming at data collection, sharing, or processing for the purpose of addressing a societal challenge (Susha et al., 2017). Susha et al. developed a taxonomy to characterize the data supply and data demand side of a data collaborative. A more technical perspective focuses on the role data infrastructures can play in terms of curating and sharing data among stakeholders. Data infrastructure includes collaborative platforms for managing and sharing data (Payne et al., 2012) that can also be supported by digital collaborative (work) spaces. The Global Facility for DRR has deployed several platforms based on Geonode, a web-based application and platform for developing geospatial information systems (GIS) and for deploying spatial data infrastructures (SDI). For example, Geodash is such a Geonode platform that was the started up by the World Bank and is now taking over by the Government of Bangladesh (Geodash, 2016). UN OCHA deploys the Humanitarian Data Exchange (HDX), more specifically targeting humanitarian data (Keβler and Hendrix, 2015).

Thirdly, to facilitate the sharing and exchange of data, standards are being developed and used –to varying degrees-ranging from P-codes for unique geographic identification codes up to the Humanitarian Exchange Language (HXL). Lastly, it will be key to develop capacities of the different stakeholders in parallel to the above activities enhancing their data literacy and access to digital technologies. Especially at the local level many respondents were for example not aware of all the existing data sets nor were they trained in data collection and analysis. The proliferation of mobile devices that can record the location of features, and access to satellite imagery and online maps (Kitchin and Lauriault, 2015) -as satellite data is becoming more widely and openly available (in resolution and across frequency bands)- will facilitate this development process. Citizens and local organizations are more and more involved in collaborative and participatory mapping and spatial data collection (Liu et al., 2018), whereas it was previously primarily done by only a few specific organizations such as Departments of Survey. Hence, information needs with a location component will become more easily fulfilled.

Our extensive Data Preparedness procedure is part of a broader *knowledge discovery process* (KDP) perspective, where it can be straightforwardly mapped to the first three phases of CRISP-DM as discussed in the introduction: Domain Understanding, Data Understanding and Data Preparation. More specifically, our research provides a proven recipe for operationalizing the often ignored third layer of the CRISP-DM framework. The recipe describes the actual best practice steps for a domain-specific instantiation. To formalize such an approach, we envision a meta-algorithmic model that further specifies a for all stakeholders maximally transparent analytical process for selecting and configuring the appropriate activities deterministically based on disaster-specific data input characteristics and local preferences (Spruit and Lytras, 2018). The last three phases of the CRISP-DM model are subject for further research. Note that properly addressing these steps was simply not feasible yet, as the Data Preparedness procedure with its available data and information needs identification has to be performed precedingly. Nevertheless, the next KDP phases (Modelling, Evaluation and Deployment) are currently being prepared. For example, we are currently experimenting with machine learning approaches in combination with natural language processing techniques for large-scale text classification, sentiment analysis and topic mining (Sangameswar et al., 2017; Syed et al., 2018).

Our research focused on the relation between available data and information needs. Although we inventoried *Decisions*, *Activities* and *Information needs*, we did not investigate the relationship between these three elements into depth and the system dynamics between the different stakeholders including the political and financial dimension. These dimensions played out for example in the still largely separate data collection processes between NGOs and government and in when a flood is declared an official disaster. As Lars-Peter Nissen, director of Assessment Capacities Project (ACAPS), said "*Very little is known about how decisions are made. Examining decision-making forces us to recognize that decisions are political. It makes us ask what may be influencing decisions, other than the needs on the ground. This is a hard question, but it is vital that we ask it, if we are to improve our capacity*." (Nissen, 2015). A political analysis of the stakeholders and the financial flows will strengthen the information management research approach.

5. Conclusions and recommendations

To bridge the information gap decision makers are facing in the aftermath of a disaster, data preparedness activities should become an integral part of the preparedness phase. The main contribution of this paper is that it developed a framework to create an inventory of available data sources and information needs and a way to calculate the coverage of these information needs. Our research shows how a selection of all the different data sets can be made by optimizing for coverage, the costs of extracting data from structured versus unstructured data and quality in terms of timeliness, source and content rating and granularity. The optimal combination of data sets can be used to extract data on information needs more efficiently. Specifically, for the context of Bangladesh and for floods, our results show seven themes of in total 71 information needs and a poor, timely coverage of these information needs by the data sets available. Three data sets are covering already 68% of the information needs, whereby adding more data sets gives only a very limited increase.

We recommend a focus on two future research directions.

First of all, a refinement of the relation between available data and information needs could lead to an enhanced understanding of the information gap. It will be worthwhile to determine the time dependency of the information needs into more detail, including a more detailed calculation of timeliness, and to do the mapping on the data products in a more automated fashion, for example, using a combined forecasting method (Maaβ et al., 2014). For large organizations, it is possible to map through which information channels (email, mobile, fax, chat) information consumers get information products from internal information producers. This kind of mapping does however not take into account the degree to which information needs are covered. Furthermore, it is much more difficult to do this kind of mapping between organizations and even more so if certain work flows are still paper based. It might be possible to log data file usage on the main websites that are used by responders and for example how the app and dashboard are used (Pachidi et al., 2014). In addition, an after-action review with the responders in a focus group setting could be used to have the responders categorize their needs according to the four phases.

A second direction consists of leveraging Artificial Intelligence for Disaster Response (AIDR). Currently, we extracted the data on the information needs manually from the data sets, but automatic matching of information needs on data sources might be possible by applying intelligent querying and semantic data matching (McNeill et al., 2014; Spruit and Vlug, 2015; Giunchiglia et al., 2007). Obviously, also for these technologies, a structured database will be easier to parse and to process. It will not be necessary to do this on all data sets; we showed that a very good coverage of information needs can already be reached by focusing on the most important data sets out of all data sets available. One could set up so-called data spaces which are loosely integrated sets of data sources where integration happens only when needed (Hristidis et al., 2010). This could become an essential extension to the earlier mentioned data exchange platforms so that these platforms offer –to a certain degree-sense making to the multiple organizations using the platform of all the data sets that are shared through them. An individual organization could use such an automatic matching mechanism as part of oDMN+, linking decision-makers' tasks to data sources (Horita et al., 2017).

Although our framework is developed for a case study on floods in Bangladesh, the findings can be applied also to other hazards. Ultimately, we aim at mainstreaming data preparedness with intelligent querying and semantic data matching into information management and policy frameworks that govern the work of disaster management professionals as well as to pilot and replicate this approach operationally for different hazards and in different developing countries.

Acknowledgements

This work was supported by Cordaid as part of the People Centered Interactive Risk and Information Gateway project and by 510 an initiative of The Netherlands Red Cross. TNO offered internship support to Robert Monné. Sander Vlot from ORTEC supported us in the design of the mathematical model for the optimization and its implementation in the AIMMS software. The authors gratefully acknowledge Marlou Geurts, MA, the program leader from Cordaid for her invaluable support and our interviewees who made so generously time available for providing us with their insights. We thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.cageo.2018.06.002.

References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000a. CRISP-dm 1.0 Step-by-step Data Mining Guide. IBM.

Care Bangladesh, 2014. Plan for Arriving at a Shared Understanding of Flooding in Bangladesh. (Unpublished report).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000b. Crisp-Dm 1.0. Cris. Consort, vol. 76. https://doi.org/10.1109/ICETET.2008.239.

Chatfield, A.T., Reddick, C.G., April 2018. All hands on deck to tweet #sandy: networked governance of citizen coproduction in turbulent times. Govern. Inf. Q. 35 (2), 259–272. https://doi.org/10.1016/j.giq.2017.09.004.

Comes, T., 2016. Cognitive biases in humanitarian sensemaking and decision-making lessons from field cognitive biases in humanitarian sensemaking and decision-making lessons from field research. In: 2016 IEEE Int. Multi-disciplinary Conf. Cogn. Methods Situat. Aware. Decis. Support. CogSIMA 2016, pp. 56–62.

Comes, T., Chan, J., van de Walle, B., Meesters, K., van den Homberg, M., Bruggemans, B., 2014. A Journey into the Information Typhoon Haiyan Disaster: Resilience Lab Field Report Findings and Research Insights: Part I-into the Fields.

De Vries, W., 2002. Dimensions of Statistical Quality A discussion note about the quality initiatives of some international organisations. https://unstats.un.org/unsd/accsub/2002docs/sa-02-6add1.pdf, Accessed date: 14 June 2018.

Fayyad, U.M., 1996. Data mining and knowledge discovery: making sense out of data. IEEE Expert 11, 20–25. https://doi.org/10.1109/64.539013.

Geodash, https://geodash.gov.bd/,Accessed 26-January-2016 2016.

Giunchiglia, F., Yatskevich, M., Shvaiko, P., 2007. Semantic Matching : algorithms and implementation. J. Data Semant 1, 1–38. https://doi.org/10.1007/978-3-540-74987-5_1.

Gralla, E., Goentzel, J., Van de Walle, B., 2015. Understanding the information needs of field-based decision-makers in humanitarian response to sudden onset disasters. In: Proceedings of the 12th International Conference on Information Systems for Crisis Response and Management, pp. 1–7.

Guha-Sapir, D., Hoyois, Ph, Wallemacq, P., Below, R., 2016. 2016 Annual Disaster Statistical Review 2016: the Numbers and Trends. CRED, Brussels.

Haak, E., Ubacht, J., van den Homberg, M., Cunningham, S., van der Walle, B., 2018. A framework for strengthening data ecosystems to serve humanitarian purposes. In: Zuiderwijk, Anneke, Hinnant, Charles C. (Eds.), Proceedings of 19th Annual in-Ternational Conference on Digital Government Research (dg.o'18). ACM, NewYork, NY, USA.

Horita, F.E.A., de Albuquerque, J.P., Marchezini, V., Mendiondo, E.M., 2017. Bridging the gap between decision-making and emerging big data sources: an application of a model-based framework to disaster management in Brazil. Decis. Support Syst. 97, 12–22. https://doi.org/10.1016/J.DSS.2017.03.001.

Hristidis, V., Chen, S.C., Li, T., Luis, S., Deng, Y., 2010. Survey of data management and analysis in disaster situations. J. Syst. Software 83 (10), 1701–1714. https://doi.org/10.1016/j.jss.2010.04.065.

IASC Guidelines, 2010. Common Operational Datasets (CODs) in Disaster Preparedness and Response. http://fscluster.org/sites/default/files/documents/IASC_Guidelines_on_Common_Operational_Datasets_in_Disaster_Preparedness_and_Response_1_Nov._2010%5B1%5D.pdf, Accessed date: 15 May 2018.

IFRC Emergency Items Catalogue, 2016. http://procurement.ifrc.org/catalogue/#1_113, Accessed 27-February-2016.

Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S., 2014. AIDR: artificial intelligence for disaster response. In: Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, pp. 159–162. https://doi.org/10.1145/2567948.2577034.

ITU Statistics, http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx,Accessed 24-January-20162015.

Keßler, C., Hendrix, C., 2015. The Humanitarian eXchange Language: coordinating disaster response with semantic web technologies. Semantic Web 6, 5–21. https://doi.org/10.3233/SW-130130.

Kitchin, R., 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage Publications.

Kimball, R., Reeves, L., Ross, M., Thornthwaite, W., 1998. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons.

Kitchin, R., Lauriault, T.P., 2015. Small data in the era of big data. Geojournal 80, 463–475. https://doi.org/10.1007/s10708-014-9601-7.

Kurgan, L.A., Musilek, P., 2006. A survey of knowledge discovery and data mining process models. Knowl. Eng. Rev. 21, 1–24. https://doi.org/10.1017/S0269888906000737.

Leidig, M., Teeuw, R.M., 2015. Quantifying and mapping global data poverty. PLoS One 10. https://doi.org/10.1371/journal.pone.0142076.

Liu, W., Dugar, S., McCallum, I., Thapa, G., See, L., Khadka, P., Budhathoki, N., Brown, S., Mechler, R., Fritz, S., Shakya, P., 2018. Integrated participatory and collaborative Risk mapping for enhancing disaster resilience. ISPRS Int. J. Geo-Inf. 7 (68). https://doi.org/10.3390/ijgi7020068.

Lu, X., Wrathall, D.J., Sundsøy, P.R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Tatem, A., Canright, G., Engø-Monsen, K., Bengtsson, L., 2016. Unveiling hidden migration and mobility patterns in climate stressed regions: a longitudinal study of six million anonymous mobile phone users in Bangladesh. Global Environ. Change 38, 1–7. https://doi.org/10.1016/j.gloenvcha.2016.02.002.

Maaß, D., Spruit, M., Waal, P. de, 2014. Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques. Decision Analytics 1 (1). https://doi.org/10.1186/2193-8636-1-4.

Madianou, M., 2015. Digital inequality and second-order disasters: social media in the Typhoon haiyan recovery. Soc. Media Soc. 1. https://doi.org/10.1177/2056305115603386.

McNeill, F., Gkaniatsou, A., Bundy, A., 2014. Dynamic data sharing for facilitating communication during emergency responses. In: ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management, pp. 369–373.

Meier, P., 2015. Digital Humanitarians, How Big Data Is Changing the Face of Humanitarian Response. Taylor & Francis Press.

MIRA Multi-sector Initial Rapid Assessment Guidance - Revision July 2015. https://www.humanitarianresponse.info/en/programme-cycle/space/document/multi-sector-initial-rapid-assessment-guidance-revision-july-2015, Accessed date: 24 January 2016.

Monné, R., 2016. Determining Relevant Disparate Disaster Data and Selecting an Integration Method to Create Actionable Information. MSc thesis. Utrecht University.

Montibeller, G., von Winterfeldt, D., 2015. Cognitive and motivational biases in decision and Risk analysis. Risk Anal. 35, 1230–1251. https://doi.org/10.1111/risa.12360.

Mozzammel Hoque, M., 2014. Development of Flood Hazard and Risk Maps with Effect of Climate Change Scenario. http://www.buet.ac.bd/iwfm/climate/report/Component_1.pdf, Accessed date: 26 January 2016.

Nissen, Lars-Peter, 2015. Keynote III Wag the Dog – information management and decision making in the humanitarian sector. In: Büscher, Palen (Ed.), Introduction Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24–27. Comes & Hughes.

Pachidi, S., Spruit, M., Van De Weerd, I., 2014. Understanding users' behavior with software operation data mining. Comput. Hum. Behav. 30, 583–594. https://doi.org/10.1016/j.chb.2013.07.049.

Parsons, M.A., Godøy, Ø., Ledrew, E., De Bruin, T.F., Danis, B., Tomlinson, S., Carlson, D., 2011. A conceptual framework for managing very diverse data for complex, interdisciplinary science. J. Inf. Sci. 37, 555–569. https://doi.org/10.1177/0165551511412705.

Payne, K., Florance, P., Shain, S., 2012. The role of data repositories in humanitarian information management and crisis mapping. J. Map Geogr. Libr. 8, 118–133. https://doi.org/10.1080/15420353.2012.662931.

Raymond, N., Al Achkar, Z., 2016. Data preparedness: connecting data, decision-making and humanitarian response. In: Harvard Humanitarian Initiative: Signal Program on Human Security and Technology - Standards and Ethics Series 01, Retrieved from. http://hhi.harvard.edu/sites/default/.

Sangameswar, M.V., Nagabhushana Rao, M., Satyanarayana, S., 2017. An algorithm for identification of natural disaster affected area. J. Big Data 4, 1–11. https://doi.org/10.1186/s40537-017-0096-1.

Spruit, M., Lytras, M., 2018. Applied data science in patient-centric healthcare: adaptive analytic systems for empowering physicians and patients. Telematics Inf. 35 (4), 643–653.

Spruit, M., Vlug, B., 2015. Effective and efficient classification of topically-enriched domain-specific text snippets. Int. J. Strat. Decis. Sci. 6, 1–17. https://doi.org/10.4018/IJSDS.2015070101.

Susha, I., Janssen, M., Verhulst, S., 2017. Data collaboratives as a new frontier of cross-sector partnerships in the age of open data: taxonomy development. In: Proc. 50th Hawaii Int. Conf. Syst. Sci, pp. 2691–2700. https://doi.org/http://hdl.handle.net/10125/41480.

Syed, S., Borit, M., Spruit, M., 2018. Narrow lenses for capturing the complexity of fisheries: a topic analysis of fisheries science from 1990 to 2016. Fish Fish. 00, 1–19. https://doi.org/10.1111/faf.12280.

Tapia, A.H., Bajpai, K., Jansen, J., Yen, J., Giles, L., 2011. Seeking the Trustworthy Tweet: Can Microblogged Data Fit the Information Needs of Disaster Response and Humanitarian Relief Organizations.

Tong, A., Sainsbury, P., Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int. J. Qual. Health Care 19, 349–357. https://doi.org/10.1093/intqhc/mzm042.

US Intelligence best practices:http://wikivisually.com/wiki/Intelligence_collection_management#Collection_department_ratings, Accessed 31-May-20172017.

van den Homberg, M.J.C., Monné, R., Spruit, M.R., 2018. Bridging the information gap: mapping data sets on information needs in the preparedness and response phase. In: Hostettler, S., Najih Besson, S., Bolay, J.C. (Eds.), Technologies for Development. UNESCO 2016. Springer, Cham. https://doi.org/10.1007/978-3-319-91068-0_18.

Van Den Homberg, M., Meesters, K., Van De Walle, B., 2014. Coordination and information management in the Haiyan response: observations from the field. Procedia Eng 78, 49–51. https://doi.org/10.1016/j.proeng.2014.07.037.

van den Homberg, M., Visser, J., van der Veen, M., 2017. Unpacking Data Preparedness from a humanitarian prioritization perspective: towards an assessment framework at subnational level. In: Proceedings of the 14th ISCRAM Conference 2017.

Verschuren, P., Doorewaard, H., Mellion, M.J., 2010. 2010, Designing a Research Project, vol. 2 Eleven International publishing house, The Hague.

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events. In: Proc. 28th Int. Conf. Hum. Factors Comput. Syst. - CHI '10 1079, . https://doi.org/10.1145/1753326.1753486.

Wahed, A., Rahman, M., Hoque, A., Costello, L., Burley, J., Walton-Ellery, S., 2014. Flooding in North-Western Bangladesh HCTT Joint Needs Assessment. http://reliefweb.int/sites/reliefweb.int/files/resources/0809_NW_Flooding_JNA_FinalFINAL.pdf, Accessed date: 7 November 2015.

Whipkey, K., Verity, A., 2015. Guidance for Incorporating Big Data into Humanitarian Operations, Digital Humanitarians Network. http://digitalhumanitarians.com/sites/default/files/resource-field_media/IncorporatingBigDataintoHumanitarianOps-2015.pdf, Accessed date: 15 May 2018.