World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:11, No:5, 2017

Data Projects for "Social Good": Challenges and Opportunities

Mikel Niño, Roberto V. Zicari, Todor Ivanov, Kim Hee, Naveed Mushtaq, Marten Rosselli, Concha Sánchez-Ocaña, Karsten Tolle, José Miguel Blanco, Arantza Illarramendi, Jörg Besier, Harry Underwood

***Abstract*—**One of the application fields for data analysis techniques and technologies gaining momentum is the area of social good or "common good", covering cases related to humanitarian crises, global health care, or ecology and environmental issues, among others. The promotion of data-driven projects in this field aims at increasing the efficacy and efficiency of social initiatives, improving the way these actions help humanity in general and people in need in particular. This application field, however, poses its own barriers and challenges when developing data-driven projects, lagging behind in comparison with other scenarios. These challenges derive from aspects such as the scope and scale of the social issue to solve, cultural and political barriers, the skills of main stakeholders and the technological resources available, the motivation to be engaged in such projects, or the ethical and legal issues related to sensitive data. This paper analyzes the application of data projects in the field of social good, reviewing its current state and noteworthy initiatives, and presenting a framework covering the key aspects to analyze in such projects. The goal is to provide guidelines to understand the main challenges and opportunities for this type of data project, as well as identifying the main differential issues compared to "classical" data projects in general. A case study is presented on the initial steps and stakeholder analysis of a data project for the inclusion of refugees in the city of Frankfurt, Germany, in order to empirically confront the framework with a real example.

***Keywords*—**Data-Driven projects, humanitarian operations, personal and sensitive data, social good, stakeholders analysis.

I. Introduction

THE increasing interest in the potential of data analytics, favored by the rise of Big Data technologies, has led to numerous applications in different industries and business sectors. The high expectations in this type of solution are also migrating from the corporate sector to public administrations and non-profit organizations in charge of projects aiming at social good or "common good" [1]. The interest in data analytics solutions in this area is not new, as since the late 1990s researchers have been analyzing administrative records from social service agencies, and applying the results for scientific and practical purposes [1]. These projects, however, were sporadic efforts, but during the last years several initiatives have been developed in the social field to promote and leverage the possibilities offered by this technological breakthrough around data-driven projects.

This paper reviews recent initiatives and relevant references on the area of data-driven projects for social good. This revision enables the identification of which opportunities are arising for the application of these projects for social purposes and which are the main challenges to face when undertaking such projects, in comparison to data projects in general.

Nowadays, there is a high demand of data-related services by humanitarian and social good-focused organizations, and also a considerable number of people volunteering to help in those projects [2]. This presents interesting opportunities to further develop the use of data-driven initiatives for social good. There are different motivations that boost this interest in data projects for social and humanitarian goals:

- The rising demand of traceability in the outcome of social action programs and evidence-based policy and practice [1]. Moreover, there is an increasing interest in making the outcome of below-the-radar social initiatives more visible (via data-enabled results, reports or visualizations) to the society in general and to the Public Administration in particular [3].
- The need for developing insights, thanks to the detection of hidden trends and patterns in data, as well as models that explain how the analyzed issue could evolve so that negative outcomes can be foreseen and prevented. In the case of humanitarian crises, for instance, these insights facilitate the response via an easier identification of the people in need and a more efficient allocation of resources [4].
- The need for an informed decision making process, with the possibility of a cyclical refresh of data that enables a feedback mechanism to this process [4]. Data-driven decision making is expected to enable more efficient organizations improving their productivity and results [5].

However, while the use and analysis of data is becoming more common among businesses to understand their customers and to develop new products and services, social organizations are not fully leveraging this potential yet [3] and the application of these solutions in this sector is lagging behind with respect to other business areas. This is mainly due to specific features of social projects that pose different challenges to overcome. As an example of this, the social sector has been slow to incorporate the practice of data

This work was supported by the Spanish Ministry of Economy and Competitiveness (MEC) [grant numbers FEDER/TIN2013-46238-C4-1-R, BES-2014-069367 and EEBB-I-16-11012].

M. Niño*, J. M. Blanco and A. Illarramendi are with the Department of Computer Languages and Systems, University of the Basque Country (UPV/EHU), San Sebastián, Spain (*corresponding author; phone +34 943018000; e-mail mikel.nino@ehu.es).

Roberto V. Zicari, Todor Ivanov, Kim Hee, Naveed Mushtaq, Marten Rosselli, Concha Sánchez-Ocaña and Karsten Tolle are with the Frankfurt Big Data Lab, Goethe University Frankfurt, Frankfurt am Main, Germany.

Jörg Besier and Harry Underwood are with Accenture GmbH, Kronberg im Taunus, Germany.

analytics into their policies [1]. This is partly because the pace of technological evolution and the new possibilities arising with data-related technologies make it difficult for policy makers to adapt the way they establish policies and for regulators to establish the legal environment for safer practices in this area. Therefore, the social sector has yet to develop a culture of transparent data exchange and aggregation [6], and governments and social organizations (which are some of the most relevant stakeholders in this field) have not developed clear data policies [4]. This increases the difficulty when arranging these projects and deciding how to address all the important processes related to data: security, privacy, confidentiality, etc.

Another important example is related to the use of sensitive information based on data from people in need. There is a lack of standardized good practices to leverage the power of data analysis for social purposes while avoiding unintended harm to the population providing the data [1]. For instance, in the case of humanitarian organizations, the vulnerability of the people subject to data analysis in the context of a humanitarian crisis constitutes an important obstacle to incorporate more data-driven projects among their activities [4]. Besides, the fear of potential security breaches makes organizations reluctant to share sensitive data that could be leveraged for analysis in this field [1].

Taking all these considerations into account, the goal of this paper is to provide a comprehensive framework with the key areas to be analyzed when designing a data project for social good, and to reinforce this framework with a case study on a data project for the inclusion of refugees in Frankfurt, Germany.

The paper is structured as follows: Section II reviews and analyzes relevant references on the topic of data projects for social good, establishing common terminology, listing noteworthy initiatives, and identifying key aspects to analyze; Section III presents a characterization framework for the key areas in data projects for social good, identifying the main challenges and opportunities, as well as the main differences with data projects in general; Section IV presents the case study on the aforementioned project for refugees' inclusion and contrasts this real case with the framework presented in the previous section; Section V presents the conclusions from this work.

II. Analysis of Related Work

The idea of data projects for social good is based on serving the people who are in need globally, improving the society we live in and people's conditions within it [7]. This is also linked to the concept of "common good", i.e. shared and beneficial for all or most members of a given community. We can consider diverse application areas comprised in this concept of social good: humanitarian crises [4], global health care and health disparities [8]-[10], ecology and global-scale environmental issues [6], rural development [11], human rights [12], crime prevention [13], child welfare, etc.

The first documented uses of Big Data and data analytics in this field are related to mapping instances of violence after the elections in Kenya in 2007, and to tracing people's movements before and after the earthquake in Haiti in 2010 [4]. From there on, many different initiatives have been launched during recent years around the idea of data projects for social good, showing the increasing interest in applying Big Data and data analytics to solve social issues. In this section, we review a compendium of noteworthy examples.

Some of the most relevant activities in this area have been launched by United Nations, with the Global Pulse initiative as the most representative one. United Nations Global Pulse promotes awareness of the opportunities of Big Data for sustainable development and humanitarian action. As they explain on their web site [14], "Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action."

Framed in this global initiative, an independent expert advisory group on the "Data Revolution for Sustainable Development" was established in 2014 [15] to advise the United Nations Secretary-General on this matter. This group produced a report [16] with key recommendations for actions to mobilize the "data revolution for sustainable development". The group's web site also contains a list of on-going activities in this area [17] by the public sector, international organizations, the private sector and civil society.

Different companies, foundations and nonprofit organizations focus their activities on data projects for social good. One of the most active organizations in this field is DataKind [18], a nonprofit that connects socially minded data scientists with organizations working to address humanitarian issues. Another example is Simpa Networks [19], a technology company in India whose mission is to make sustainable energy affordable to people in need. They sell solar-as-a-service to energy-poor households and small business, and they collaborate with DataKind volunteers to use Simpa Networks' historical data on customer payment behavior to predict which new applicants are likely to be a good fit for this model [2]. Another case is the Flowminder Foundation [20], who collects and analyzes anonymous mobile operator data, satellite and household survey data, in order to characterize and map populations at risk in low- and middle-income countries. As an example of social actions led by big corporations, we can refer to the language translation systems built by Microsoft researchers for the aid relief workers in Haiti after the 2010 earthquake [21].

Some specific challenges related to data handling in this context have also been addressed by different research teams. For instance, the Data Ethics Research Initiative [22] launched by Accenture's Technology Vision team "brings together leading thinkers and researchers from Accenture Labs and over a dozen external organizations to explore the most pertinent issues of data ethics in the digital economy". They produced a report [23] analyzing the role of data ethics and how ethical controls could be integrated throughout data

supply chains.

Also focused on the use of data in this type of projects, the open letter "Data for Humanity" [24] was published in 2015 asking to use data for the common good and to serve humanity, and inviting everyone involved in working with data to sign the letter and support the five principles proposed in it. A related initiative, the "Health Data Bill of Rights" [25], was launched in 2009 (now discontinued) looking for adhesions on a list of principles about the use of personal health information.

The rising interest in data projects for social good has motivated diverse experts to publish different analyses reflecting on such projects and providing recommendations and guidelines to develop them. These recommendations vary on scope and depth depending on the case. Some are open reflections as a result of interviews [21], [26] or guest posts [27], [28]. Some others are published as white papers, with more detailed guidelines and providing information on use cases to exemplify the key reflections [1]-[4], [29].

There have also been efforts to launch education programs focused on "data science for social good". These types of programs was pioneered by the University of Chicago [30] in 2013 and later developed in other places such as Georgia Tech in Atlanta [31] and the University of Washington [32]. Furthermore, the profile of the "humanitarian data scientist" [33] has been coined in order to identify the skills required for a data scientist to effectively apply their knowledge during a humanitarian crisis.

Academic journals and conferences on data science and data mining have also put their focus on the emerging topic of data projects for social good. The Big Data journal devoted a special issue in 2015 to "Big Data for Social Good" [34]. The 2014 edition of the ACM's annual Conference on Knowledge Discovery and Data Mining (KDD 2014) designated "Data Science for Social Good" as the conference theme, and included in its program [35] a full-day workshop on this topic. The University of Chicago, who also co-organized this 2014 workshop (and pioneered education programs on this topic, as we mentioned before), organized in August 2016 the first edition of a two-day Conference on Data Science for Social Good [36]. This trend was followed by a workshop on Data Science for Social Good (SoGood 2016) [37] held in September 2016, associated to the 2016 edition of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML-PKDD 2016). Also related to this topic, the responsible and transparent use of data was addressed in the tutorial on "Data, Responsibly: Fairness, Neutrality and Transparency in Data Analysis" organized in the EDBT 2016 Conference [38].

III. Characterization of Key Areas in Data Projects for Social Good

The detailed revision of the references presented in the previous section provided crucial insights to identify the main challenges and opportunities present in these types of data projects. In order to review them in a structured way, we organized the analysis around some key areas to analyze in these projects (Fig. 1). Thus, the analysis can be leveraged as a reference framework when planning and designing the initial steps of these data projects. In this section, we present a thorough characterization of those key areas and the challenges and opportunities to be considered in each of them.

A. Social Issue to Address with Data Projects

All data projects must begin with a clear identification of the social issue to solve or mitigate [4]. However, in this context of social good, the process of problem discovery has its own difficulties [27]. Working with agents in this context will presumably derive in stakeholders not being very familiar with the latest technology, or at least much less technology-aware than stakeholders in other typical contexts for data projects (finance, online businesses, manufacturing). As a result, they lack a clear understanding of what can be achieved with data, which affects the discovery of relevant social problems (i.e. the articulation of needs) to be addressed with data projects. Furthermore, the differences in the domain languages of the different stakeholders need to be overcome, which is a time-consuming task.

When defining the problem to address, two important transversal questions must be analyzed:

- The integration of data projects into the issue and how they support its solution or mitigation. We can consider four main goals: advocating and facilitating; describing and predicting; facilitating information exchange; and promoting accountability and transparency [29].
- The key questions (to be answered with data) that are derived from the selected approach among the options above. Those key questions lead to determining the key indicators to measure, monitor or predict, which drives the identification of the relevant to capture and process.

Among the indicators to consider, we must include the required ones to measure the "success" of the project and to justify starting it. In other words, we need to measure the "social gain" obtained from the investment in designing, developing, deploying and maintaining a data-driven solution [7]. In order to justify a data project in an enterprise, in most cases there is a need to identify a quantitative "Return on Investment" (ROI). This ROI serves as a measure of the value that can be derived by analyzing these data. However, when data-driven solutions are used to improve our society and people's conditions within it, we cannot use the "standard ROI" as an impact measure (in these social projects, the profit is derived by Key Performance Indicators –KPIs- that are not contained in the income statement). Concepts such as social impact bonds [39] or "pay for success" [1] could provide a link between the investment required to finance a data project for social good and the social outcomes it provides. Nevertheless, the "social profitability" and potential of these actions might not look very promising when considering one-time initiatives in some specific location. The real potential arises when these solutions are designed to scale so that they can globally solve (or at least mitigate) a common social issue. The major challenge here is that a global approach to solve a common social issue in different locations around the globe

must also deal with the differences among geographical, political, cultural and socioeconomic contexts. Furthermore, even if we only focus on the technological aspects of the project, there might be a high heterogeneity in the type of data available, what language it is in, and whether it can be easily accessed [4]. Moreover, when we integrate all these different contexts into the design of the solution, there is an important challenge to be overcome from the point of view of Humanities and Social Sciences: they are based on the principle of ensuring that different perspectives are represented and, therefore, they fear that extracting trends from big data sets can complicate this goal by missing important variations and exceptions and misrepresenting certain groups and perspectives [40].

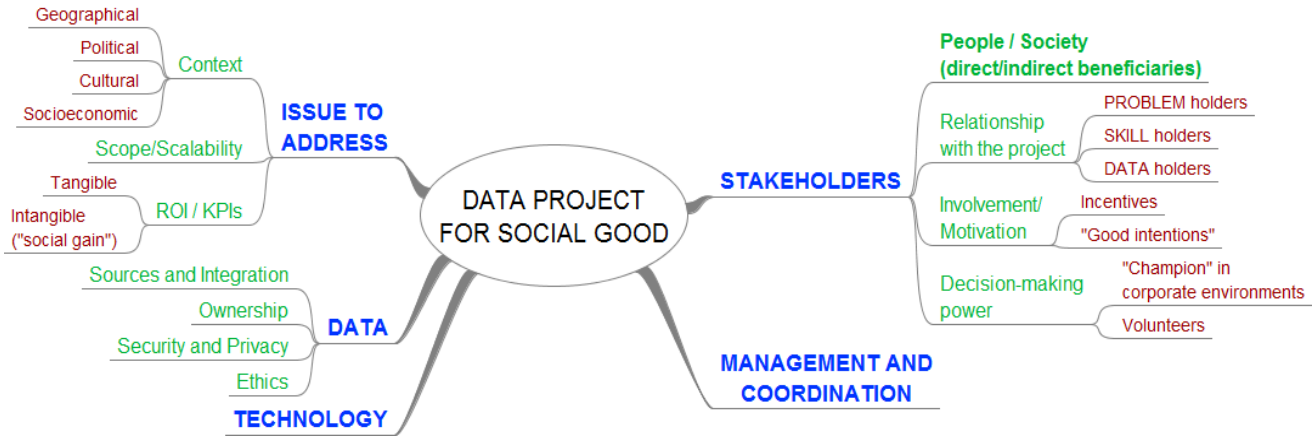

Fig. 1 Classification of key areas to analyze in a data project for social good

B. Characterization and Classification of Stakeholders

We classify the main stakeholders in these types of data projects in three different dimensions.

1. Relationship with the Project

Following the classification provided by [28] about the different types of stakeholders that have to collaborate in these data projects, we can identify *problem holders*, *data holders*, and *skill holders*.

Problem holders are the stakeholders closest to the social issue to be dealt with in the project. Given the type of social issues that are the focus of these projects, these stakeholders are usually representatives of public administration. More specifically, they are in charge of departments or organizations created by public administrations in order to take care of these issues and manage solutions to them. They provide key knowledge about the issue and important goals to be achieved within the project. In the absence of interaction with the people directly benefiting from the social gain achieved with the project (or as a complement to this interaction), the *problem holders* can act on behalf of this collective and express their needs. Their professional profile and background usually makes them more familiar with sociological issues but not so much with technology and its potential to be applied in these contexts. Therefore, special attention has to be paid in providing them with relevant information about the technological possibilities and combine it with their vision of the problem.

Data holders are the ones closest to the data to be captured, processed and analyzed in the project. Depending on the specific project, we can establish different relationships between *data holders* and the rest of stakeholders. For instance, in some cases they can be the same as the *problem holders*. This would be the case e.g. when the public organization in charge of this social issue has been gathering relevant data that can be analyzed. In some other cases third parties can play this role, e.g. external data providers. However, all relevant data may not already exist before the project is launched, and they might need to be generated *ex novo*. Sometimes the people or collective in need are indeed a data source in themselves. For instance, their interactions with different systems or services could generate relevant data to be processed and analyzed in the project.

Skill holders are necessary to counteract the lesser familiarity with these technological tools by other profiles of stakeholders and their potential lack of general data literacy [29]. In different case studies and reflections related to data projects for social good [1], [3], it has been highlighted the need for a combination of profiles, teaming the experts in the social issue to be addressed with data scientists who contribute with their skills and key ideas about what can be achieved with the data. Thanks to these data experts' providing tools and methods (taking into account that this type of training could be a challenge in itself), other non-technical stakeholders can also gain some important skills to interrogate and analyze data on their own terms [3]. This overcomes the problem of social work and social welfare degree programs not training students in these skills, beyond basic statistics [1].

2. Motivation and Incentives

The collaboration of different stakeholders in a consortium to develop a common project must be based in a shared commitment to working for the common good [41]. This is not always as easy as it may seem, because sometimes the collection and use of data may harm the self-interests of some of the organizations involved. Therefore, any attempt to

leverage data analytics will have to fight this resistance [42].

One important challenge to be tackled here is the potential conflict (and possible source of lack of motivation in some stakeholders) between the tangible and intangible goals by different stakeholders, linked to their different incentives for their participation in the project. We may consider the case of the *problem holder*, exemplified in an organization or department of public administration in charge of the social issue. Depending on their tactical goals (sometimes established by strategies beyond their direct control), their motivation and incentives might be focused on specific indicators that don't capture the whole picture of the social issue. Moreover, not all of their KPIs would be directly related to that specific social issue. This might generate a conflict with the expectations of other stakeholders and participants in the project when their motivation is more altruistic or with a more global view of the social need to be solved.

3. Decision-Making Power (Volunteering vs. Corporate "Champions")

Many initiatives to solve social problems are strongly based in volunteering work. Volunteers play an important role in these projects, as many of them have personal experiences related to the social issue, and this motivates them to collaborate in finding a solution. However, there is always a need for determining specific project goals and deploying coordination mechanisms, so that these good intentions and efforts are correctly managed to avoid dispersion and to produce a valuable outcome (as in any other project). This poses an important challenge: how to translate those project management skills used in "business projects" into an environment where people may prefer to contribute "freely" and may react against being "managed". Moreover, most projects require considerable resources and full-time commitment, and there are limits to what you can accomplish by volunteering [2], which is indeed necessary to sustain these efforts but it is not enough. In the words of Jake Porway [28], "using data science in the service of humanity requires much more than free software, free labor and good intentions". This is where the involvement of corporate partners comes to play. As stated in [27], "the public sector […] cannot fully exploit big data without leadership and partnership from the private sector".

In order to effectively involve and engage corporate partners in this kind of projects, the profile of "champions" in these organizations plays a crucial role. These champions combine three important features that contribute to the success of the project: they have the personal motivation and interest to participate in initiatives with social purposes, they have the decision-making power in order to engage staff and resources in the project, and they can establish a link between the social outcome of the project and its contribution to the corporate goals.

C. Managing the Involvement of Stakeholders

The task of managing and coordinating a data project in this area entails important challenges in the way stakeholders are 'engaged' into the project and how their contributions and expertise are combined into the desired outcome.

If we see a data project as a macro-process, the input is composed of several key elements provided by each stakeholder: skills and know-how (on the issue to solve, on data, on technology), resources, motivation and incentives, etc. Furthermore, if we consider the different organizations acting as stakeholders in the project, there is also a need to identify a champion in each organization. Each champion (as explained in the previous subsection), plays a crucial role in the engagement of each organization in the projects, as they have the decision-making power to allocate resources, budget and personnel to the project. One important issue here is to achieve the right alignment of champion's motivation and incentives with those from their organization, as in these contexts there may be a combination between business goals and some more 'altruistic' ones that should not be in conflict.

Given this 'input' to the macro-process, an effective management of these data projects can be seen as a cycle (Fig. 2) that has to deal with the following challenges:

- *Management and Coordination*. The project will integrate different profiles of an organization as stakeholders, each providing different degrees of background and expertise on the key aspects of the project (social issue to solve, data, technology), and with different interests in being part of the project. This implies an effective integration of all these backgrounds (from 'silos' to cooperating organizations) and a need for saving the gaps between these (potentially big) differences on skills, know-how and motivation. It is needed to understand staff capacity to incorporate them into these projects, and to find the right balance between staff and volunteering force [4].
- *Social gain/outcome*. All the input has to be channeled into the right output, i.e. setting the right scope and tangible/intangible goals for the project so that the social issue becomes the focus and all resources provided by stakeholders are used coordinately for that purpose.
- *Periodic visible milestones*. If the project does not incorporate the right short-term milestones to make visible the outcome of the volunteer's effort, it becomes complicated to sustain their motivation and engagement during the whole project. In order to keep this motivation high, one should also avoid long phases of not being productive or idle times. Sometimes it is necessary to periodically renew the volunteering efforts with new people, in order to sustain the project's momentum [3].
- *Ensure sustainability*. A sustainable outcome of the project must be ensured, so that the achieved benefits are not a one-off, isolated effort and they can be extended in time via a continuous involvement of stakeholders and resources. There is also a need for establishing feedback loops so that users remain involved in the work beyond an initial contribution [29].

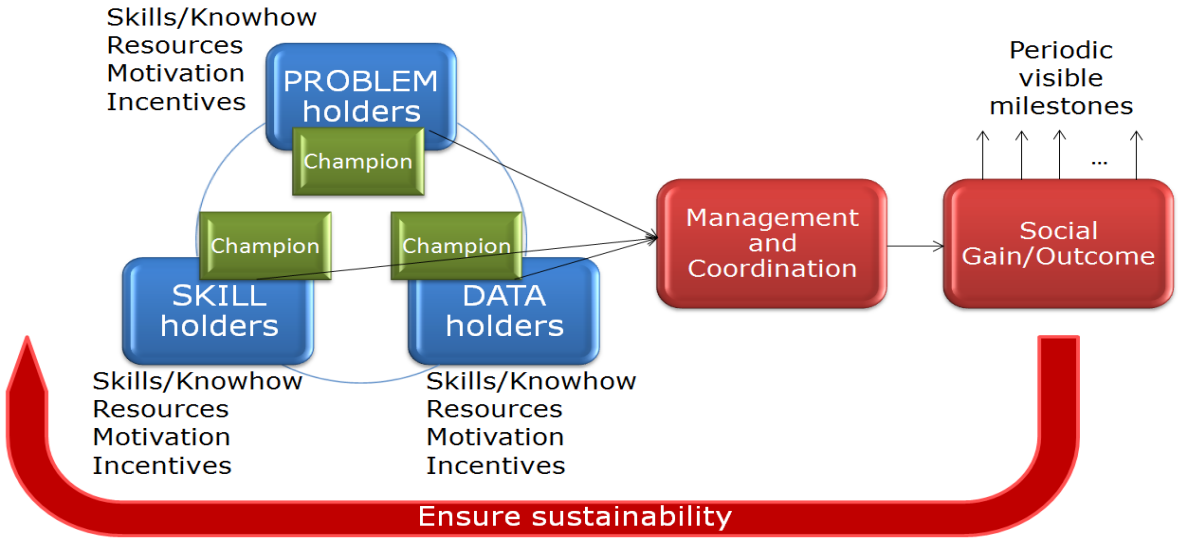

Fig. 2 The "cycle" of stakeholder engagement and sustainability assurance in data projects for social good

D. Considerations on Data

The challenges and differential issues on the use of data in these scenarios can be analyzed according to different dimensions, which we present next.

1. Data Sources and Integration

Most often the data used for analysis in these scenarios is not expressly created for this use. Instead, it is generated as a byproduct of other processes [10]. This derives into two main types of data sources that can be combined in these projects: (a) accessible data repositories that were not designed for these goals but that can be leveraged for this purpose [21], and (b) new data generated and captured with mechanisms that are deployed for this specific purpose. In these cases, where data were not originally generated for the social good-related purpose, it is necessary to understand the process or scenario that generated them and to carefully plan the data preprocessing, cleaning and validating phases, so that those data can be effectively used in the new scenario [1]. Moreover, in these cases sometimes it stays unclear whether users understand how their data can be used and who may access them [10].

As several organizations (Public Administration at different locations and levels of responsibility, non-profit organizations, etc.) are usually concerned with the same type of social issue, there is an effect of "data silos" that have to be integrated. This integration may presumably involve a considerable effort, given the need for dealing with legal and structural barriers to data sharing and the different data governance practices across institutions [1]. The effort for integration is increased when those data have to be merged with data from external providers such as social network platforms, e.g. for real-time monitoring and sentiment analysis. Resources such as the Humanitarian Exchange Language (HXL) [43] might ease the burden of this integration.

2. Data Ownership

A key question to address in these projects is who owns the data that have to be captured, stored and processed [26]. Parting from the idea that Big Data is increasingly about real behavior, we can think of those data as a "model" of a person's actions and their outcomes. Therefore, on one hand, we have the people from whom we want to obtain data. On the other hand, most of these social problems are managed by organisms from Public Administration (although there may be differences depending on the area in the world). There is a need for a clear model on who (the people, the public agencies, some external data provider…) is the owner of which data in these scenarios. Furthermore, we have to distinguish between volunteered data and observed data [29]. The former indicates data whose owner –e.g. the person generating those data with their activities– has volunteered to offer those data, and the latter is related to the use of data that either their owner did not explicitly volunteered of that were originally captured for a different purpose. As stated in [29], there is a need for developing new and comprehensive ethical systems to protect data subjects where data is observed rather than volunteered.

3. Data Security and Privacy

Despite the growing concern about access to and sharing of personal information [44], this field lacks a clear and effective framework to address legal, ethical and privacy issues related to the use of personal and sensitive data [1]. Moreover, there is a need for clear policies determining how long data should be kept and what to do with it after their use [4], and for establishing guidelines on how cooperating organizations can exchange and link useful data incorporating the right measures for data confidentiality and anonymization. In this regard, UN Global Pulse propose that the "private sector and public sector must come together" to develop "mechanisms to protect privacy based on shared guidelines, regulations and technology". This will "help build frameworks in which data

can be safely and ethically analyzed for insights that can be used to help protect populations" [27]. Advances in the application of privacy-by-design methodologies [45] in data projects for social good might contribute to achieve this goal.

Privacy has to be interpreted in the right context of the initiative leveraging the data. For instance, it is clearly different to reveal personal health details in case of an urgent medical intervention or to share health details with an insurance company [5]. A context-sensitive interpretation of ethical obligations may help understand which data uses are acceptable or not, and for which purposes [10].

It could be debated whether privacy considered in an isolated way as an individual right outranks greater societal interests in any case or not, as sometimes the common good may demand a balanced response between privacy rights and other values [5]. This debate is being held e.g. in health care applications, reflecting on whether requirements for individual consent for data sharing are creating barriers to life-saving innovation [9]. Paul Miller reflects on this same issue in [21]:

> *"Much of this information is about people. Can we extract enough information to help people without extracting so much as to compromise their privacy? Partly, this calls for effective industrial practices. Partly, it calls for effective oversight by Government. Partly – perhaps mostly – it requires a realistic reconsideration of what privacy really means… and an informed grown up debate about the real trade-off between aspects of privacy 'lost' and benefits gained. Rather than offering blanket privacy policies, perhaps customers, regulators and software companies should be moving closer to some form of explicit data agreement; if you give me access to X, Y, and Z about yourself, I will use it for purposes A, B, and C… and you will gain benefits/services D, E, and F. The first two parts are increasingly in place, albeit informally. The final part – the benefits – is far less well expressed".*

The right methods for data anonymization and the use of data-sharing agreements and non-disclosure protocols are mandatory tools in these scenarios. There is, however, an intrinsic difficulty in incorporating to these agreements the potential future uses of data than cannot be foreseen when settling the grounds for handling these data.

Sometimes the confidentiality is a key trust-building aspect of an organization's claim towards the people they gather sensitive personal data from. This hampers the possibility of leveraging those data by "opening" them or taking advantage of them for different purposes than the one they were originally meant for. An organization facing this dilemma can feel more comfortable with sharing data once the right protocols for sensitive data handling are established and put into practice [3]. Furthermore, not only regulations on data collection have to be considered in these scenarios. We can differentiate between regulations focused on data use or on data collection. In [46] it is suggested that use-based protections are more appropriate to deal with "latent" knowledge than can be inferred from data by machine learning techniques. Besides, these use-based approaches are preferred by industry, as they are generally opposed to constraining data collection [46]. However, one main difficulty here is that determining whether future uses of data might be beneficial or harmful may be a too complex question for most people to make an informed consent and to self-manage their privacy [5].

4. Data Ethics

When dealing with this type of sensitive data, especially in social initiatives related to humanitarian crises, we have to consider two different thresholds on the privacy and the feasible uses of data: one is established by the legal regulations on personal data where the project is being developed, whereas the other one is related to the emotional sensitivity in the data and the "ethically allowable" uses for them [3]. Indeed, depending on the type of legal boundaries limiting the use of some type of data, the inferences obtained by machine learning techniques could eventually circumvent these limitations [46]. The use of machine learning and inference-based techniques, as well as the possibility to aggregate massive data from different sources, make it more difficult to link the original purposes of shared data and the information that eventually can be derived from them [46]. This presents an opportunity to design technological solutions incorporating the mechanisms to prevent ethical issues from coming into being [5]. However, the perception of ethical uses of data and the preservation of rights might be different depending on the geographical and cultural background. Therefore, ethical designs will depend on cultural expectations [5].

There are initiatives focusing on ethics and trust as a crucial aspect when handling data in these projects, alongside security and privacy. For instance, in [23] it is proposed that the ethics of data collection, manipulation and use requires attention at each stage of the data supply chain and collaboration with every stakeholder. A new set of best practices should be created to build "trusted by design" solutions and to guide practitioners through the process of embedding ethical considerations at every stage of product development, service delivery and the data supply chain.

E. Considerations on Technology

It is necessary to analyze limitations accessing the technology, which can be due to lack of resources or skills. This has to be analyzed for different stakeholders: subjects of analysis (to provide data), beneficiaries of solution (to access results), agents managing the data project (to understand the possibilities and to analyze data accordingly), and policy makers (to monitor the initiatives and to promote the right actions). For instance, humanitarian crises often strike populations who do not possess the latest technologies and communication methods. Therefore, not taking into account the technological limitations of the potential beneficiaries when basing the analysis on data gathered electronically could lead to a misrepresentation of collectives in need [40]. For instance, important issues might go unnoticed because the only ones affected are people without easy access to

technology and, therefore, without a significant digital footprint [1]. As well as this, there is the technical challenge of developing a platform reliably available to low-income users [29], which are the usual direct beneficiaries of these social actions. Moreover, given that in most cases these social issues are dealt with by different public agencies managing sensitive personal data, it is important to consider the right technological resources for these scenarios. For example, it is not always possible to leverage cloud computing platforms from external third parties because, depending on their location and data policies, there might be a conflict over regulations on data ownership, privacy and acceptable uses.

IV. Case Study: InteGreat, A Project for the Inclusion of Refugees in Frankfurt am Main

The goal of this section is to provide an example of how a data project for social good can be analyzed using the reference framework presented in the previous section. For that purpose, we analyze an ongoing data project for the inclusion of refugees in Frankfurt am Main (Germany). The project (in its initial steps of planning and design) is reviewed through the lens of the key areas and challenges identified before, focusing on the analysis of stakeholders driving the project. This section presents some background information on the social issue to address and the scope and goals of the project, details the characterization of main stakeholders in the dimensions introduced in the framework, and finally includes some preliminary considerations on data and technology identified in these initial steps.

A. Background Information on the Case Study

1. Social Issue to Address

The Directorate-General for European Civil Protection and Humanitarian Aid Operations, according to figures provided by the UN Refugee Agency, states that more than 1 million refugees, displaced persons and other migrants moved to the UE in 2015. This includes people escaping conflict in their countries of origin and also those searching for better economic prospects [47].

Focusing on the city of Frankfurt, 4,300 refugees had settled there by March 2016. It is anticipated that by the end of 2019 a total of 16,000 refugees will have to be integrated in the city.

Inclusion also requires that the receiving population is open to accept the refugees. However, some citizens do have fears. Therefore, it is important to establish a transparent way of communication with and between all the actors (refugees, citizens, government, etc.). Most citizens do not know what exactly happens and how trustworthy the numbers found in various sources are. Refugees have different backgrounds, beliefs and value systems. There are differences in culture and life style. A dialog is needed between many parties to understand and to deal with these differences.

2. Scope and Goals of the Project

The project aims at addressing the following main challenges, in this case in the city of Frankfurt:

- An initial fast inclusion into the labor market for refugees, in order to provide a perspective of the society and feel part of it. This is vital for the civil society to see the positive effects for the community and the will of the refugees to contribute.
- A mid- and long-term effective inclusion process for refugees into the city fabric. This includes getting the motivation upright of the refugees and other stakeholders to participate.
- A measurable and sustainable inclusion process.

To address these challenges, the project will provide the foundation and the implementation for services and opportunities to bring an inclusive experience to refugees in the city of Frankfurt.

The project will help refugees at first to more easily acquire the necessary skills to find a job, and in the long run to be an active participant in the host society.

There is potentially no limit to the scalability of the proposed solution, provided that the incentive/benefit mechanisms in place work effectively. The infrastructure will be designed to be scalable and adaptable to changes.

B. Characterization of Main Stakeholders

According to the dimensions identified in the framework presented in section III, next we characterize the beneficiaries of the social project, the main stakeholders as *problem* or *skill holders*, and relevant considerations on the motivation and incentives for the engagement of these stakeholders.

1. Beneficiaries of the Social Project

As explained before, the project is focused on refugees (single persons, families, children) living already in or arriving at the city of Frankfurt, and their inclusion into the labor market. The *problem holder* (see next subsection) provided the project with a detailed characterization of this collective.

From the refugees coming to the city of Frankfurt, 80% are male and about 70% are below the age of 30. In global figures, 30% of refugees are coming from Afghanistan, 25% from Syria, 20% from Eritrea, 15% from Iran and Iraq. Some 4,300 refugees arrived in Frankfurt since 2015 until March 2016, and each week 200 new refugees are arriving in the city. It was expected that 8,000 to 9,000 refugees were to be integrated in Frankfurt until the end of 2016.

Although the amount of refugees coming into the country is slightly decreasing, the need for an integration project remains high. The global needs of this collective are related to anything that can help in the integration process and, in particular for this project, to the necessary learning in order to receive the basic school diploma required to work in Germany.

Although most of this collective have a low level of education in general and a limited communication ability in English, they have a high motivation level for learning. At the same time, they are familiar with smartphones and the use of social media networks to organize reunions and exchanges.

The main issue is about gaining trust from the side of the refugees (they likely faced a lot of difficulties along their route, there is a lot of mistrust and sometimes rejection to

being helped) and also regarding the perception of society about the collective in need, which hampers society's contribution to the social issue. This crucial aspect is closely related to the need for short-term and periodic visible outcome (Fig. 2) that the project has to provide in order to break potential barriers of mistrust with clear examples of the benefits and positive effects of the initiative for all involved.

2. The Problem Holder: FRAP Agency

The FRAP Agency (FRAP Agentur) [48] is a company dependant on the Municipal Council of the city of Frankfurt. The object of the FRAP Agency is the coordination and development of local employment promotion. They operate an advisory center for the Frankfurt labor market to offer individual counseling related to work, vocational and continuing education. In particular, they run the project "Frankfurt Helps" [49], which aims at coordinating the activities towards the inclusion of refugees in the city of Frankfurt.

They interact with potential employers and with training-related service providers, on one hand, and on the other hand with the refugees via a consultation process where they do an assessment of their skills. This assessment is based on the time spent in school and on the knowledge of the school system in the country of origin. This implies the need for a previous analysis of the socioeconomic and cultural context in those countries. The interaction with refugees is done thanks to the help of volunteers working as translators for the languages spoken by refugees (Tigrinya, Farsi, Arabic). This exemplifies the combination of corporate staff and volunteering effort to deal with the same social issue.

Apart from obviously considering the direct uses by the beneficiaries of the social initiative (in this case, the refugees in the city of Frankfurt), the design of the data-driven solution must be based on a wider vision of the stakeholders and their informational needs. For instance, the *problem holders* can leverage these technologies to achieve a better coordination and communication between different organizations and staff involved (in this case, for example, to facilitate the work between service providers and volunteers). This will improve the way they address the social issue.

3. The Skill Holders: Accenture Foundation and the Frankfurt Big Data Lab

Two other main stakeholders take part in the project as *skill holders*:

- Accenture Foundation is experienced in partnering with non-governmental organizations to deliver projects and value for beneficiaries. Their volunteers are highly experienced in setting up new projects and organizing and implementing refugee support activities. Volunteers come from different nations and have organized activities related to language and cultural training for refugees.
- Frankfurt Big Data Lab (from Goethe Frankfurt University) has highly qualified competence in interdisciplinary expertise on digital technologies, economy and social areas.

The problem of the distance to data-driven technological possibilities by stakeholders close to the social issue (i.e. the *problem holders*) is also present the other way around. People working intensively with technology and data analytics may be more accustomed to business and corporate-related uses of those solutions, and therefore they need more effort in this area to fully understand and characterize the social problem at hand.

4. Motivation and Incentives for the Engagement of Stakeholders

The previous description of stakeholders exemplifies the required combination of profiles (related to social problem vs. data technology savvy, corporate staff vs. volunteers) to effectively build a solution for the social issue at hand. Another relevant consideration is the combination of goals and the engagement of champions from all organizations involved (in this case, their directors) who have the decision-making power required to engage staff and resources in the project.

An important lesson learned from this stakeholder analysis is that there may be a potential conflict between the tangible and intangible goals in a project for social good. A very common scenario in these projects is that the stakeholder with the profile of *problem holder* will be a public agency (a specific branch of some Public Administration) focusing on the social issue at hand. This will generally involve some dependency on key performance indicators (directly related to the social issue) they are accountable for. Moreover, those indicators might only be part of the goals that the agency must put their attention on. In this scenario, it is probable that the focus of this public agency's work only covers part of the social issue (the one with more tangible and quantifiable goals). This poses an important challenge in regard to the motivation of the other participants in the project (mainly volunteers), who may be moved to action by a more altruistic and global perspective of the problem at hand and the social needs to be solved. The potential conflict between those two visions may lead to a progressive lack of motivation by the volunteering stakeholders. Therefore, a clear vision of the expectations by all stakeholders involved, shared by all of them since the start of the project, is required in order to achieve the right balance in expectations and to avoid frustration in later stages of the project.

C. Preliminary Considerations on Data and Technology

In terms of the role of *data holders*, there is little pre-existent data that could be leveraged in the project. The City Council provides FRAP Agency with some basic identification data, but most of this information is captured in the consultation process described in the previous subsection.

Information is stored on their own server. In order to manage data privacy, the server is operated by an agency that abides by the data security standards specified by the City Council.

Data analytics will be focused on the interaction of refugees with what the service providers offer to complete refugees' training. The identification of relevant data is based on what is

required to assess use and create visualizations and segmentations: location, submitted offers, used offers, time spent reading, and search results, etc.

V. Conclusions and Future Work

The development of data-driven projects is gathering increasing attention in the field of social good, covering diverse types of applications. Global and local institutions, big and small corporations, experts and volunteers, as well as academia, have been launching different initiatives during recent years to promote and facilitate the use of these data projects for different social problems, ranging from humanitarian crises to global health care and environmental issues, to name a few.

A joint analysis of relevant references from all these sources shows that there are important challenges, specific to this application field, that have to be addressed when designing these sorts of data projects. The framework of key areas to analyze presented in this paper aims at contributing with a structured compilation of specific features of data projects for social good. This can help guide new projects in order to identify from the start the main challenges that will have to be overcome, as well as the opportunities for valuable contributions to strengthen this field under development.

In this sense, the analysis of the initial steps of the case study presented here provides an important validation of this identification of key areas and challenges and confirms its utility to guide the initial steps of these types of data projects. More specifically, the analysis of stakeholders in the case study through the lens of the characterization presented in the framework has contributed with relevant insights and reinforces the key elements included in the framework that can be leveraged in other projects.

This work has yet to be extended further, as more projects are developed and more reflection is made on their development and outcome. Given that an increasing number of academia activities are focusing on this area, it is expected that new research can contribute to consolidate the framework in the short term. Besides, as the case study discussed in the paper is further developed and the requirements analysis is extended, it is expected to gain valuable lessons learned to integrate in this vision of data projects for social good.

Acknowledgment

We would like to thank Mr. Conrad Skerutsch, Managing Director of the FRAP Agentur (City of Frankfurt), for his valuable collaboration in the InteGREAT project.

References

[1] C. J. Coulton, R. Goerge, E. Putnam-Hornstein, B. de Haan, *Harnessing Big Data for social good: A grand challenge for social work*. Working Paper No. 11, American Academy of Social Work and Social Welfare, 2015. Retrieved in Aug. 2016, from: http://aaswsw.org/publications/

[2] M. Barlow, *Data and Social Good: Using data science to improve lives, fight injustice, and support democracy*. Technical Report, O'Reilly Media, 2015. Retrieved in Aug. 2016, from: http://www.oreilly.com/data/free/data-and-social-good.csp

[3] P. Baeck, *Data for Good: How big and open data can be used for the common good*. Nesta, 2015. Retrieved in Aug. 2016, from: http://www.nesta.org.uk/publications/data-good

[4] K. Whipkey and A. Verity, *Guidance for incorporating Big Data into humanitarian operations*. Technical Report, Decision Makers Needs & Digital Humanitarian Network, 2015. Retrieved in Aug. 2016, from: http://digitalhumanitarians.com/resource/incorporating-big-data-humanitarian-operations

[5] S. Dobner and C. Voigt, *Social, ethical and legal aspects of Big Data and urban decision*. Project UrbanData2Decide, deliverable D2.4, 2015. Retrieved in Aug. 2016, from: http://www.urbandata2decide.eu/media-centre/

[6] S. E. Hampton et al., "Big data and the future of ecology", in *Front. Ecol. Environ.*, vol. 11, no. 3, pp. 156-162, Mar. 2013. doi:10.1890/120103

[7] R. V. Zicari et al., "Setting up a Big Data project: Challenges, opportunities, technologies and optimization", in *Big Data optimization: Recent developments and challenges*, A. Emrouznejad, Ed., Studies in Big Data vol. 18, pp. 17-47. Springer International Publishing Switzerland, 2016. doi:10.1007/978-3-319-30265-2_2

[8] S. I. Hay, D. B. George, C. L. Moyes, and J. S. Brownstein, "Big Data opportunities for global infectious disease surveillance", in *PLoS Med.*, vol. 10, no. 4, e1001413, pp. 1-4, Apr. 2013. doi:10.1371/journal.pmed.1001413

[9] E. B. Larson, "Building trust in the power of "Big Data" research to serve the public good", in *JAMA The Journal of the American Medical Association*, vol. 309, no. 23, pp. 2443-2444, Jun. 2013. doi:10.1001/jama.2013.5914

[10] E. Vayena, M. Salathé, L. C. Madoff, and J. S. Brownstein, "Ethical challenges of Big Data in public health", in *PLoS Comput. Biol.*, vol. 11, no. 2, e1003904, pp. 1-7, Feb. 2015. doi:10.1371/journal.pcbi.1003904

[11] K. R. Varshney et al., "Targeting villages for rural development using satellite image analysis", in *Big Data*, vol. 3, no. 1, pp. 41-53, Mar. 2015. doi:10.1089/big.2014.0061

[12] F. Chen and D. B. Neill, "Human rights event detection from heterogeneous social media graphs", in *Big Data*, vol. 3, no. 1, pp. 34-40, Mar. 2015. doi:10.1089/big.2014.0072

[13] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Finding Patterns with a Rotten Core: Data Mining for Crime Series with Cores", in *Big Data*, vol. 3, no. 1, pp. 3-21, Mar. 2015. doi:10.1089/big.2014.0021

[14] UN Global Pulse, *United Nations Global Pulse: Harnessing big data for development and humanitarian action (main website)*. Retrieved in Aug. 2016, from: http://unglobalpulse.org/

[15] UN Global Pulse, *Independent expert advisory group formed to advise UN Secretary-General on 'data revolution'*, 2014. Retrieved in Aug. 2016, from: http://unglobalpulse.org/IEAG-Data-Revolution

[16] L. González Morales, Y.-C. Hsu, J. Poole, B. Rae, and I. Rutherford, *A world that counts: Mobilizing the data revolution for sustainable development*. Technical Report, UN Data Revolution Group, 2014. Retrieved in Aug. 2016, from: http://www.undatarevolution.org/report/

[17] UN Data Revolution Group, *Initiatives on the data revolution for post-2015*. Retrieved in Aug. 2016, from: http://www.undatarevolution.org/catalog/

[18] DataKind, *DataKind: Harnessing the power of data science in the service of humanity (main website)*. Retrieved in Aug. 2016, from: http://www.datakind.org/

[19] Simpa Networks, *Simpa Networks: We make clean energy simple, affordable and accessible to everyone (main website*). Retrieved in Aug. 2016, from: http://simpanetworks.com/

[20] Flowminder Foundation, *Flowminder.org: Providing priceless information for free for the benefit of those who need it the most (main website)*. Retrieved in Aug. 2016, from: http://www.flowminder.org/

[21] R. V. Zicari, *Big Data for good: A distinguished panel of experts discuss how Big Data can be used to create social capital*. ODBMS Industry Watch, 2012. Retrieved in Aug. 2016, from: http://www.odbms.org/blog/2012/06/big-data-for-good/

[22] Accenture, *Data Ethics Research Initiative*, 2016. Retrieved in Aug. 2016, from: https://www.accenture.com/us-en/insight-data-ethics

[23] H. Lynch et al., *Building digital trust: The role of data ethics in the digital age*. Technical Report, Accenture, 2016. Retrieved in Aug. 2016, from: https://www.accenture.com/us-en/insight-data-ethics

[24] R. V. Zicari and A. Zwitter, *Data for Humanity: An Open Letter*, 2016. Retrieved in Aug. 2016, from: http://www.bigdata.uni-frankfurt.de/dataforhumanity/

[25] T. O'Reilly, *A manifesto on health data rights*, 2009. Retrieved in Aug. 2016, from: http://radar.oreilly.com/2009/06/manifesto-health-data-rights.html
[26] A. Pentland, *Reinventing society in the wake of Big Data*. Edge.org, 2012. Retrieved in Aug. 2016, from: https://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data
[27] R. Kirkpatrick, *The importance of Big Data partnerships for sustainable development*. United Nations Global Pulse, 2016. Retrieved in Aug. 2016, from: http://www.unglobalpulse.org/big-data-partnerships-for-sustainable-development
[28] J. Porway, *Five principles for applying data science for social good*. O'Reilly Media, 2015. Retrieved in Aug. 2016, from: https://www.oreilly.com/ideas/five-principles-for-applying-data-science-for-social-good
[29] Bellagio Big Data Workshop Participants, *Big data and positive social change in the developing world: A white paper for practitioners and researchers*. Oxford Internet Institute, 2014. Retrieved in Aug. 2016, from: https://www.rockefellerfoundation.org/report/big-data-and-positive-social-change-in-the-developing-world/ doi:10.13140/2.1.3931.7761
[30] University of Chicago, *Eric & Wendy Schmidt data science for social good summer fellowship (main website)*. Retrieved in Aug. 2016, from: https://dssg.uchicago.edu/
[31] University of Atlanta, *Data science for social good program (main website)*. Retrieved in Aug. 2016, from: http://dssg-atl.io/
[32] A. Rokem et al., "Building an urban data science summer program at the University of Washington eScience Institute", in *Bloomberg Data for Good Exchange Conference*, New York City, NY, USA, Sept. 2015
[33] A. Verity, *Humanitarian data scientist - who and how?*, 2014. Retrieved in Aug. 2016, from: http://blog.veritythink.com/post/105715607274/humanitarian-data-scientist-who-and-how
[34] C. Catlett and R. Ghani, "Big Data for social good", in *Big Data*, vol. 3, no. 1, pp. 1-2, Mar. 2015. doi:10.1089/big.2015.1530
[35] KDD2014, *Program of the 20th ACM SIGKDD conference on knowledge discovery and data mining*, 2014. Retrieved in Aug. 2016, from: http://www.kdd.org/kdd2014/program.html
[36] University of Chicago, *Data science for social good conference*, 2016 Retrieved in Aug. 2016, from: http://dssg.uchicago.edu/data-science-for-social-good-conference/
[37] SoGood 2016, *Workshop on data science for social good*, 2016. Retrieved in Aug. 2016, from: https://sites.google.com/site/ecmlpkdd2016sogood/home
[38] J. Stoyanovich, S. Abiteboul, and G. Miklau, "Data, responsibly: Fairness, neutrality and transparency in data analysis", in *Proc. 19th Int. Conf. on Extending Database Technology (EDBT'16)*, Bordeaux, France, Mar. 2016, pp. 718-719. doi:10.5441/002/edbt.2016.103
[39] UK Government's Cabinet Office, *Social impact bonds*, 2013. Retrieved in Aug. 2016, from: https://www.gov.uk/guidance/social-impact-bonds
[40] P. Fletcher and C. Hall, "Digital humanities and the common good", in *Maine Policy Review*, vol. 24, no. 1, pp. 124-131, Jun. 2015. doi: http://digitalcommons.library.umaine.edu/mpr/vol24/iss1/35
[41] E. Birney, "The making of ENCODE: Lessons for big-data projects", in *Nature*, vol. 489, pp. 49-51, Sept. 2012. doi:10.1038/489049a
[42] D. Bollier, *The promise and peril of Big Data*. Technical Report, The Aspen Institute, 2010. Retrieved in Aug. 2016, from: https://www.aspeninstitute.org/publications/promise-peril-big-data/
[43] C. Keßler and C. Hendrix, "The Humanitarian eXchange Language: Coordinating disaster response with semantic web technologies", in *Semantic Web*, vol. 6, no. 1, pp. 5-21, 2015. doi:10.3233/SW-130130
[44] S. Chaudhuri, "What next? A half-dozen data management research goals for Big Data and the Cloud", in *Proc. 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems (PODS '12)*, Scottsdale, AZ, USA, May 2012, pp. 1-4. doi:10.1145/2213556.2213558
[45] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi, "Privacy-by-design in big data analytics and social mining", in *EPJ Data Science*, vol. 3, no. 10, pp. 1-26, Dec. 2014, doi:10.1140/epjds/s13688-014-0010-4
[46] E. Horvitz and D. Mulligan, "Data, privacy, and the greater good", in *Science*, vol. 349, no. 6245, pp. 253-255, Jul. 2015. doi:10.1126/science.aac4520
[47] European Commission, *Refugee crisis in Europe*, 2016. Retrieved in Aug. 2016, from: http://ec.europa.eu/echo/refugee-crisis_en
[48] FRAP Agentur, *FRAP Agentur: Gemeinnützige Gesellschaft für das Frankfurter Arbeitsmarktprogramm (main website)*. Retrieved in Aug. 2016, from: http://frap-agentur.de/
[49] Frankfurt City Council, *Frankfurt hilft: Engagement für Flüchtlinge (main website)*. Retrieved in Aug. 2016, from: http://frankfurt-hilft.de/