International Review of the Red Cross (2020), **102** (913), 75–94.
Digital technologies and war
doi:10.1017/S1816383121000084

IRRC

Humanitarian aid in the age of COVID-19: A review of big data crisis analytics and the General Data Protection Regulation

Theodora Gazi and Alexandros Gazis
Theodora Gazi is a lawyer and a PhD candidate in refugee law at the School of Law, University of Athens. She is a Data Protection Specialist for the Danish Refugee Council (DRC Greece) and has been working in humanitarian aid since 2017.

Alexandros Gazis is a PhD candidate in computer science at the School of Engineering, Democritus University of Thrace, where he works as a Teaching Assistant and a Lab Demonstrator. He is also a Software Engineer for Eurobank SA, specializing in core banking systems.

Abstract
The COVID-19 pandemic has served as a wake-up call for humanitarian aid actors to reconsider data collection methods, as old ways of doing business become increasingly obsolete. Although access to information on the affected population is critical now more than ever to support the pandemic response, the limitation of aid workers' presence in the field imposes hard constraints on relief projects. In this article, we consider how aid actors can use "big data" as a crisis response tool to support humanitarian projects, in cases when the General Data Protection Regulation is applicable. We also provide a framework for examining open-source platforms, and discuss the advantages and privacy challenges of big data.

Keywords: big data, humanitarian aid, COVID-19, GDPR, data collection, crisis response.

Introduction

"Big data" has emerged as one of the most used buzzwords in the digital world, promising unique insight into understanding the aftermath of a disaster and the areas of need. In a displacement context, both the lack of information and the overflow of data may be paralyzing. For traditional humanitarian organizations, the use of big data is still uncharted territory. The question that arises is how aid actors can make use of large amounts of data, the majority of which is unstructured. On the one hand, data analytics introduce new opportunities for aid actors to support affected populations. On the other hand, big data could have serious implications for vulnerable individuals and communities if applied without safeguards. Importantly, practitioners should ensure compliance with data protection rules and best practices prior to resorting to innovative data collection methods, as this goes hand in hand with the humanitarian principles of non-discrimination and "do no harm" in the digital environment.

The goal of this study is to address the relationship between data protection and big data. As a result, we do not delve deeper into the intrinsic complexities of either of these issues. To explore the application of big data in humanitarian aid projects, we organize this article into two sections. First, we discuss the different views on what constitutes big data and on its potential use by aid actors to tackle the issues presented by COVID-19, focusing our analysis on two open-source software case studies. Then, we lay out key data protection rules in the EU and present the particularities of applying the General Data Protection Regulation (GDPR) to the processing of data from vulnerable populations. While the GDPR is applicable only to a portion of aid actors, we believe that its careful consideration is important. Indeed, it constitutes a "last-generation" data protection law that is shaping global regulatory trends on how to protect personal data in an increasingly digital world, along with other global benchmarks such as the ICRC's *Handbook on Data Protection in Humanitarian Action*.[1] Our purpose is to summarize the literature on big data, offer insight into its contribution to humanitarian projects and highlight its potential use by aid actors during the pandemic.

Defining big data and its use during the COVID-19 pandemic

"Big data" is an umbrella term that originated in the mid-1990s and became popular from 2011 onwards.[2] Its definition varies depending on the sector, and will likely

1 Christopher Kuner and Massimo Marelli (eds), *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC, Geneva, May 2020, p. 93.

2 Amir Gandomi and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics", *International Journal of Information Management*, Vol. 35, No. 2, 2015, p. 138, available at: https://doi.org/10.1016/j.ijinfomgt.2014.10.007 (all internet references were accessed in January 2021).

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

evolve further, since what is defined as big data today may not be classified as such in a few years.[3] According to the independent European working party on the protection of privacy and personal data,[4] big data refers to "the gigantic amounts of digital data controlled by companies, authorities and other large organisations which are subjected to extensive analysis based on the use of algorithms. Big Data may be used to identify general trends and correlations".

In the data science industry, big data is defined by the "three Vs":[5] volume (large amounts of data), variety (data derived from different forms, including databases, images, documents and records) and velocity (the content of the data is constantly changing through complementary data from multiple sources). This list can be further enriched[6] to accommodate the intrinsic characteristics of aid projects by including veracity (credibility of the data for informed decision-making), values (respect of privacy and ethical use of crisis data), validity (mitigating biases and pitfalls), volunteers (motivation and coordination of volunteers) and visualization (presentation of big data in a coherent manner to support informed decisions). Throughout our work, we have adopted this enriched definition for aid projects in order to demonstrate the main data processing principles.

Moreover, big data refers to combining and analyzing information from diverse sources.[7] Depending on its source, data can be both structured (i.e., organized in fixed fields, such as spreadsheets and data sets) and unstructured (e.g., photos or words in documents and reports). In a crisis context, we identify the following sources for big data analysis:[8]

1. Data exhaust: information provided by individuals as by-products during the provision of humanitarian assistance, e.g. operational information, metadata records and web cookies. This refers to data which were not actively collected

3 See National Institute of Science and Technology, *NIST Big Data Interoperability Framework*, Vol. 1: *Definition*, US Department of Commerce, 6 September 2015, pp. 4–5, available at: http://dx.doi.org/10.6028/NIST.SP.1500-1; Council of Europe, *Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in a World of Big Data*, Strasbourg, 23 January 2017, p. 2.

4 Article 29 Data Protection Working Party, *Opinion 03/2013 on Purpose Limitation*, 2 April 2013, p. 45, available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

5 Jules J. Berman, *Principles and Practice of Big Data*, 2nd ed., Elsevier, London, 2018, p. 2.

6 Junaid Qadir, Anwaar Ali, Raihan ur Rasool, Andrej Zwitter, Arjuna Sathiaseelan and Jon Crowcroft, "Crisis Analytics: Big Data-Driven Crisis Response", *Journal of International Humanitarian Action*, Vol. 1, Article No. 12, 2016, p. 14, available at: https://doi.org/10.1186/s41018-016-0013-9.

7 Indicatively, see Alexandros Gazis and Eleftheria Katsiri, "Web Frameworks Metrics and Benchmarks for Data Handling and Visualization", in Yann Disser and Vassilios Verykios (eds), *Algorithmic Aspects of Cloud Computing*, Lecture Notes in Computer Science, Vol. 11409, Springer, Cham, 2018, available at: https://doi.org/10.1007/978-3-030-19759-9_9; Alexandros Gazis and Eleftheria Katsiri, "A Wireless Sensor Network for Underground Passages: Remote Sensing and Wildlife Monitoring", *Engineering Reports*, Vol. 6, No. 2, 2020, available at: https://doi.org/10.1002/eng2.12170.

8 UN Global Pulse, *Big Data for Development: Challenges and Opportunities*, May 2012, p. 16, available at: www.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf; Bapu Vaitla, *The Landscape of Big Data for Development: Key Actors and Major Research Themes*, Data2x, May 2014, available at: https://data2x.org/wp-content/uploads/2019/09/LandscapeOfBigDataForDevelopment_10_28-1.pdf.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

but rather "left behind" from other digital interactions. These data are used as a sensor of human behaviour.

2. Crowdsourcing: information actively produced or submitted by individuals for big data analysis, via online surveys, SMS, hotlines etc. This method has been described as "the act of taking a job traditionally performed by a designated agent and outsourcing it to an undefined, generally large group of people in the form of an open call".[9] This information is valuable for verification and feedback.
3. Open data: publicly available data sets, web content from blogs and news media etc. Web content is used as a sensor of human intent and perceptions.
4. Sensing technology: satellite imagery of landscapes, mobile traffic and urban development. This information monitors changes in human activity.

The use of big data analysis peaked during the COVID-19 pandemic, which progressed from a worldwide public health emergency to a social and economic crisis. Scholars have claimed that at the time of writing (late 2020), all countries are using big data analytics to visualize COVID indicators in real time (such as case data, epidemic distribution and situation trends), inform the public about the epidemic situation and support scientific decision-making.[10]

Big data is especially relevant for aid actors in the context of disaster management, e.g. during migration crises, epidemics, natural disasters or armed conflicts.[11] During the COVID-19 pandemic, aid agencies switched to remote methodologies for data collection, such as phone surveys, remote key informant interviews and secondary data analysis.[12] Remote data collection relies heavily on the use of telecommunications and digital tools, such as phone calls, online surveys, SMS and messaging apps (such as WhatsApp and Signal). Big data analysis can also support aid actors in epidemic surveillance and response. However, the application of big data analysis to medical data is not widespread, due to the sensitive nature of medical records and the lack of a common technical infrastructure that can facilitate such analysis. The use of big data for

9 Jeff Howe, *Crowdsourcing: How the Power of the Crowd Is Driving the Future of Business*, Random House, New York, 2008.

10 See Qiong Jia, Yue Guo, Guanlin Wang and Stuart J. Barnes, "Big Data Analytics in the Fight against Major Public Health Incidents (Including COVID-19): A Conceptual Framework", *International Journal of Environmental Research and Public Health*, Vol.17, No. 17, 2020, available at: https://doi.org/10.3390/ijerph17176161. Their argument is strengthened by the data published through dashboards created by Johns Hopkins University (available at: https://coronavirus.jhu.edu/map.html) and the World Health Organisation (available at: https://covid19.who.int/) regarding active and past positive cases.

11 J. Qadir *et al.*, above note 6.

12 See International Organization for Migration, "Adapting to Change: IOM Faces COVID-19 Pandemic by Strengthening Outreach Tools", 6 February 2020, available at: www.iom.int/news/adapting-change-iom-faces-covid-19-pandemic-strengthening-outreach-tools; Save the Children, "Tipsheet: Remote and Digital Collection & COVID-19", 29 March 2020, available at: www.ready-initiative.org/wp-content/uploads/2020/06/COVID-19-and-MEAL-Remote-Data-Collection_v1.0-Save-the-Children1.pdf; Office of the UN High Commissioner for Refugees (UNHCR), Global Data Service, Innovation Service and Global Tri-Cluster Group, "Data Collection in Times of Physical Distancing", 13 August 2020, available at: www.unhcr.org/blogs/data-collection-in-times-of-physical-distancing/.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

epidemic surveillance mainly involves the processing of crowdsourced data from volunteers who report protection needs.

Big data platforms include both commercial and free open-source products – i.e., software whose source code is open and publicly available for organizations to access, adjust or further enhance for any purpose.[13] Crisis management tools[14] may either be built from scratch or be revamped to fulfil existing needs. To better understand the advantages and disadvantages of big data analysis, we draw on a number of previous projects. First, we will review two recent projects led by a government agency and the private sector, which each developed an algorithm to predict migration trends. Then, we will focus on the Ushahidi and Sahana projects, which we believe are the most suitable open-source platforms to support humanitarian operations, and discuss their use for COVID-19 monitoring, depending on the size of the operation and the intended use.

Prediction of migration trends

Predictions of migration flows enable actors to better plan their resources in order to respond in a timely manner to humanitarian needs. The Swedish Migration Agency, the government body responsible for evaluating applications for asylum and citizenship in Sweden, has initiated a relevant big data project.[15] The Agency is using big data analysis to predict migration trends via annual comparisons of stored data. In this way, it gains insight into the expected needs and can plan for up to six months ahead in order to deploy resources to alleviate bottlenecks.[16] For instance, in October 2015, the Agency accurately predicted the number of refugees expected to arrive in Sweden by the end of the year.[17] However, while it predicted a high influx for 2016,[18] the number of submitted asylum applications declined significantly in that year.[19] The lower number of asylum-seekers in

13 Karim Lakhani and Eric Hippel, "How Open Source Software Works: 'Free' User-to-User Assistance", in Cornelius Herstatt and Jan G. Sander (eds), *Produktentwicklung mit virtuellen Communities*, Springer Gabler, 2004, available at: https://doi.org/10.1007/978-3-322-84540-5_13.

14 These include open/interactive mapping platforms (e.g. OpenStreetMap, Ushahidi, Sahana); health-care modules (e.g. OpenMRS); applications for volunteer management (e.g. Collabbit), budget and finance monitoring (e.g. Mifos), relief response (e.g. Relief Response Database), communication (e.g. FrontLineSMS) and food supply (e.g. LibreFoodPantry); and geographical information system tools (e.g. QGIS).

15 "Migrant Crisis: Sweden Doubles Asylum Seeker Forecast", *BBC News*, 22 October 2015, available at: www.bbc.com/news/world-europe-34603796.

16 Organisation for Economic Cooperation and Development and European Asylum Support Office, *Can We Anticipate Future Migration Flows?*, Migration Policy Debates No. 16, May 2018, p. 6, available at: www.oecd.org/els/mig/migration-policy-debate-16.pdf.

17 More specifically, the Swedish Migration Agency announced that 140,000–190,000 refugees were expected to arrive in Sweden by the end of the year, including 29,000–40,000 unaccompanied children. Indeed, 163,000 applications for international protection were registered in 2015, out of which 35,400 were from unaccompanied children. See Swedish Migration Agency and European Migration Network Sweden, *EMN Annual Report on Migration and Asylum 2016: Sweden*, 2017, available at: https://ec.europa.eu/home-affairs/sites/homeaffairs/files/27a_sweden_apr2016_part2_final_en.pdf.

18 An influx of 100,000–170,000 people was predicted for 2016, including up to 33,000 unaccompanied children.

19 In 2016, 28,939 asylum-seekers were registered, out of which 2,199 were unaccompanied children.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

Sweden during 2016 was linked to the EU–Turkey Statement[20] signed in March 2016 to stop crossings to the Greek islands[21] and the border closure of the "Balkan route" to Europe. This has led scholars to argue that long-term decision-making based on migration forecasts is prone to error from unforeseen future events, while short-term predictions are far more useful.[22]

Another example of using big data for predictions of migration is the Danish Refugee Council's (DRC) partnership with IBM to develop a foresight model[23] (called Mixed Migration Foresight) in 2018. The project[24] focused on migration patterns from Ethiopia to six other countries. Anonymous data of thousands of migrants interviewed by the DRC revealed the main reasons for migration: lack of rights and/or access to social services, economic necessity, or conflict. Subsequently, these factors were mapped as quantitative indicators. Then, statistics about Ethiopia were processed, including its labour economy, education system, demographics and governance. Using these indicators, forecasts were produced for mixed migration flows to other countries. On average, the model was 75% accurate for 2018 figures.[25]

According to the forecasting software, the COVID-19 pandemic would lead to the displacement of more than 1 million people during 2020 across the Sahel region of Africa.[26] This prediction indeed captured the high increase of displacement that occurred in the area. However, already in November 2020, 1.5 million people had been displaced in the Central Sahel region, due to "unprecedented levels of armed violence and rights violations".[27] Moreover, based on the DRC's analysis, 6 million people residing in Mali, Niger and Burkina Faso were pushed into extreme poverty due to the pandemic. Again, this example shows that, while predictions based on big data may not fully factor the

20 See the full statement in European Council, "EU–Turkey Statement", 18 March 2016, available at: www.consilium.europa.eu/en/press/press-releases/2016/03/18/eu-turkey-statement/.

21 According to the European Commission, the EU–Turkey Statement was a "game changer". Irregular arrivals of migrants to the EU dropped by 97% from 2016 onwards compared to 2015. See European Commission, "EU–Turkey Statement: One Year On", 17 March 2017, available at: https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-migration/background-information/eu_turkey_statement_17032017_en.pdf.

22 George Disney, Arkadiusz Wiśniowski, Jonathan J. Forster, Peter W. F. Smith and Jakub Bijak, *Evaluation of Existing Migration Forecasting Methods and Models*, Economic and Social Research Council, Centre for Population Change, Southampton, 10 October 2015, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/467405/Migration_Forecasting_report.pdf.

23 Rahul Nair, Bo Madsen, Helena Lassen, Serge Baduk, Srividya Nagarajan, Lars Mogensen, Rana Novack, Rebecca Curzon, Jurij Paraszczak and Sanne Urbak, "A Machine Learning Approach to Scenario Analysis and Forecasting of Mixed Migration", *IBM Journal of Research and Development*, Vol. 64, No. 1/2, 23 October 2019, available at: https://doi.org/10.1147/JRD.2019.2948824.

24 Rahul Nair, "Machine Learning in Action for the Humanitarian Sector", *IBM Research Blog*, 21 January 2019, available at: www.ibm.com/blogs/research/2019/01/machine-learning-humanitarian-sector/.

25 Karen Faarbæk de Andrade Lima, "Ethiopia Prototype", *DRC INSITE*, 21 January 2020.

26 Kate Hodal, "Covid to Displace More than a Million across the Sahel, New Tool Predicts", *The Guardian*, 11 August 2020, available at: www.theguardian.com/global-development/2020/aug/11/covid-to-displace-more-than-a-million-across-the-sahel-new-tool-predicts.

27 DRC, "Central Sahel is Rapidly Becoming One of the World's Worst Humanitarian Crises", 11 November 2020, available at: https://drc.ngo/about-us/for-the-media/press-releases/2020/11/central-sahel-crisis/.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

highly politicized migration context, they can recognize migration trends and imminent humanitarian crises.

Consequently, both examples showcase that big data analysis is indeed useful as a prediction tool for recognizing migration patterns and informing decisions on expected needs. However, this analysis becomes "old news" quite soon, since external factors, such as climate change,[28] political decisions and the pandemic, can severely impact migration flows. Recognizing these limitations, however, big data can still serve as an indicator for preparedness, advocacy and programme planning.

The Ushahidi project

The Humanitarian Free and Open-Source Software community developed Ushahidi (meaning "testimony" in Swahili) in 2008 using PHP programming language.[29] Ushahidi is considered a "micro-framework" application, meaning that it adopts a minimalist approach, providing organizations with basic functions to fulfil three specific tasks: data collection, data management and visualization. Its main outputs are the visualization of data on a map, after applying data mining techniques.[30] Despite its fairly simple design, Ushahidi is included in the big data ecosystem for crisis management[31] because it is capable of analyzing both small and large data sets from diverse sources (per the defining "three Vs" of big data: volume, variety and velocity).

Initially, the application analyzed crowdsourced data solely via SMS messages[32] that reported incidents. Text messages were chosen as the most reliable method for data collection, given the limited network coverage at the time in the affected areas.[33]

For instance, during the 2010 Haiti earthquake, Ushahidi was used as a crowdsourcing platform to produce a crisis map based on information shared by volunteers who generated around 50,000 incident reports.[34] At the time, the US Federal Emergency Management Authority proclaimed the Ushahidi map as the "the most comprehensive and up-to-date source of information on Haiti for the

28 See International Organization for Migration, *Climate Change and Migration: Improving Methodologies to Estimate Flows*, IOM Research Series, No. 33, 2008, https://publications.iom.int/system/files/pdf/mrs-33.pdf.

29 Okolloh Ory, "Ushahidi, or 'Testimony': Web 2.0 Tools for Crowdsourcing Crisis Information", *Participatory Learning and Action*, Vol. 59, No. 1, 2009, available at: www.researchgate.net/publication/233563796_Ushahidi_or_%27testimony%27_Web_20_tools_for_crowdsourcing_crisis_information.

30 Stephen Kovats, *The Future of Open Systems Solutions, Now*, UNESCO World Summit on the Information Society, Berlin, 6 May 2013, available at: www.academia.edu/8746057/The_Future_of_Open_Systems_Solutions_Now.

31 J. Qadir *et al.*, above note 6.

32 The use of SMS for crowdsourcing had already been successfully adopted in previous NGO projects by UNICEF: Rapid SMS in 2010 (code available at: https://github.com/rapidsms/rapidsms) and U-Report in 2011 (code available at: https://github.com/unicefuganda/ureport).

33 See the code for using Ushahidi for crowdsourcing by SMS at the project's SMSSync GitHub public repository, available at: https://github.com/ushahidi/SMSSync).

34 Femke Mulder, Julie Ferguson, Peter Groenewegen, Kees Boersma and Jeroen Wolbers, "Questioning Big Data: Crowdsourcing Crisis Data Towards an Inclusive Humanitarian Response", *Big Data & Society*, 10 August 2016, available at: https://doi.org/10.1177/2053951716662054.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

humanitarian community".[35] A four-digit telephone number was published and Haitians were encouraged to share urgent needs via text messages or emails, to be made public after they were translated. Three distinct groups contributed to this process: the digital humanitarians who ran the platform, Haitians affected by the earthquake, and global volunteer translators. "Implied" consent was used as a legal basis to make the incident reports public, based on the broad information-sharing about the project's purpose via radio and TV messaging.[36] However, we should note that this approach is problematic according to global data protection standards, since tolerance of a practice should not equal its acceptance, and the vulnerability of data subjects should also be taken into consideration.[37] We will further explore consent as a legal basis in the section below on "Applying Data Protection in Big Data", in light of the GDPR, which introduced strict requirements for the validity of consent. We should clarify that if the GDPR had been in force during that time, it would have applied if the digital humanitarians who ran the platform were EU-based actors.

Moreover, Ushahidi has been used retroactively by researchers to ameliorate aid response. For instance, researchers analyzed the geographic mobile phone records of nearly 15 million individuals between June 2008 and June 2009 in order to measure human mobility in low-income settings in Kenya and understand the spread of malaria and infectious diseases.[38] The Kenyan phone company Safaricom provided de-identified information to researchers, who then modelled users' travel patterns.[39] Researchers estimated the probability of residents and visitors being infected in each area by cross-checking their journeys with the malaria prevalence map provided by the government. This case would raise privacy concerns if the mobile phone data were publicly available, due to the re-identification risks based on persons' unique activity patterns. For this reason, when de-identified personal data are used for analysis purposes, anonymization procedures typically alter the original data slightly (causing a loss of data utility) in order to protect individuals' identities.[40] As we will analyze in the section below on "Applying Data Protection in Big Data", however, true anonymization of personal data is not always possible.

Furthermore, to reduce its cost for participants, the Ushahidi application integrated additional features for data collection and text processing in the

35 Jessica Heinzelman and Carol Waters, *Crowdsourcing Crisis Information in Disaster-Affected Haiti*, United States Institute of Peace, Washington, DC, 29 September 2019.

36 "Crisis Mapping Haiti: Some Final Reflections", *Ushahidi Blog*, 14 April 2020, available at: www.ushahidi.com/blog/2010/04/14/crisis-mapping-haiti-some-final-reflections.

37 See C. Kuner and M. Marelli (eds), above note 1, pp. 61–63.

38 For the full study, see Amy Wesolowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow and Caroline O. Buckee, "Quantifying the Impact of Human Mobility on Malaria", *Science*, Vol. 338, No. 6104, 12 October 2012, available at: https://doi.org/10.1126/science.1223467.

39 Harvard School of Public Health, "Using Cell Phone Data to Curb the Spread of Malaria", press release, 11 October 2012, available at: www.hsph.harvard.edu/news/press-releases/cell-phone-data-malaria/.

40 Ling Yin, Qian Wang, Shih-Lung Shaw, Zhixiang Fang, Jinxing Hu, Ye Tao and Wei Wang, "Re-identification Risk versus Data Utility for Aggregated Mobility Research Using Mobile Phone Location Data", *PLOS One*, Vol. 10, No. 10, 2015, available at: https://doi.org/10.1371/journal.pone.0140589.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

following years. Nowadays, more data streams may be processed, including emails, web forms and tweets based on hashtags. Since 2017, the application has also adopted artificial intelligence processing to automate data gathering via the use of chatbots.[41] More specifically, automation bots can interact with users via the Facebook Messenger application. Following a short "dialogue" between the user and the bot, immediate suggestions are offered based on algorithms or the request is catalogued for further processing.[42]

During the COVID-19 pandemic, the Ushahidi platform has also been used to map the availability of public services, volunteer initiatives and requests for help. For instance, the Italian organization ANPAS[43] visualized services offered by volunteers across Italy in order to respond to recurrent needs for food, medicine and necessary goods.[44] Similarly, the FrenaLaCurva project[45] allowed Spanish language-speakers to share needs and available resources in Spain and the Canary Islands.[46] The Redaktor project[47] focused on empowering institutions and journalists across the globe, in addition to promoting community-oriented dissemination of information by mapping their needs for support. These examples demonstrate that big data can be and has been used in various ways to support the provision of help and various services to those affected by COVID and related restrictions.

Ushahidi is fairly easy to set up and serves as a crowdsourcing platform which may be accessed from multiple devices in remote areas, even if network connectivity is low. Its main disadvantage is its dependence on unstructured data (i.e., words in different languages and metadata), which frequently results in missing or inaccurate information.[48] Additionally, aid actors should take into consideration that big data analysis may be inherently biased, since it can exclude marginalized and under-represented groups, such as children, illiterate persons,

41 "'Hi This Is the Ushahidi Facebook Messenger Chatbot' – Meeting People Where They Already Are", *Ushahidi Blog*, 25 August 2017, available at: www.ushahidi.com/blog/2017/08/25/hi-this-is-the-ushahidi-facebook-messenger-chatbot-meeting-people-where-they-already-are-1.

42 Joanna Misiura and Andrej Verity, *Chatbots in the Humanitarian Field: Concepts, Uses and Shortfalls*, Digital Humanitarian Network, May 2019, available at: www.academia.edu/40918719/Chatbots_in_the_humanitarian_field_concepts_uses_and_shortfalls.

43 See the interactive map developed by Ushahidi for ANPAS, available at: https://anpas.ushahidi.io.

44 June Mwangi, "ANPAS: Supporting Vulnerable Communities in Italy during Covid-19 Lockdowns", *Ushahidi Blog*, 20 March 2020, available at: www.ushahidi.com/blog/2020/03/20/anpas-supporting-vulnerable-communities-in-italy-during-covid-19-lockdowns.

45 See the interactive map developed by Ushahidi for FrenaLaCurva, available at: https://es.mapa.frenalacurva.net/.

46 Angela Oduor Lungati, "Frena La Curva: Connecting Spanish Speakers with Critical Resources Around Them", *Ushahidi Blog*, 25 March 2020, available at: www.ushahidi.com/blog/2020/03/25/frena-la-curva-connecting-spanish-speakers-with-critical-resources-around-them.

47 See the interactive map developed by Ushahidi for the Redaktor project, available at: https://redaktor.ushahidi.io/.

48 Unstructured data do not have a predefined structure, so big data analysis requires additional processing power (i.e., CPU and RAM usage), a higher execution time or the purchase of costly computer systems to mine the data. Indicatively, see A. Gazis and E. Katsiri, "Web Frameworks Metrics", above note 7; Kiran Adnan and Rehan Akbar, "An Analytical Study of Information Extraction from Unstructured and Multidimensional Big Data", *Journal of Big Data*, Vol. 6, Article No. 91, 17 October 2019, available at: https://doi.org/10.1186/s40537-019-0254-8.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

the elderly, indigenous communities and people with disabilities.[49] Furthermore, it does not always provide aid actors with sufficient information on the incidents reported, e.g. location, description and number of affected individuals.[50]

Moreover, the applicable data protection law needs to be taken into consideration when aid organizations invite users to post public reports through the platform. For instance, while Ushahidi has updated its policies and practices to comply with the GDPR,[51] actors which are either EU-based or which target individuals residing in the EU (irrespective of the organization's place of establishment) still need to acquire users' consent as defined by GDPR's strict criteria and inform them accordingly about any processing activities. This is because compliance with data protection is not just about the use of appropriate software tools; it extends to all aspects of data life-cycle management and to respecting data subjects' rights.

Sahana project

In 2009, the Humanitarian Free and Open-Source Software community developed the Sahana project (meaning "relief" in Sinhalese). Sahana consists of two applications, Agasti[52] and Eden.[53] In contrast to Ushahidi, this project includes framework applications providing organizations with versatile options during big data analysis, instead of only core functions. We will focus our analysis on Eden, since its numerous modules serve multiple purposes during humanitarian projects, from support services to programmatic and field needs. Eden, which stands for Emergency Development Environment, is a more sophisticated application than Ushahidi, using the Python programming language. By processing structured data (mainly in CSV format[54]), it supports organizations in managing people, assets and inventory.

The modules integrated in Eden may be utilized[55] for both supporting (e.g. for inventory and human resources management) and programming purposes (e.g.

49 Shweta Bansal, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani and Cécile Viboud, "Big Data for Infectious Disease Surveillance and Modeling", *Journal of Infectious Diseases*, Vol. 214, No. 4, 2016, available at: https://doi.org/10.1093/infdis/jiw400.

50 Patrick Meier, *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*, Routledge, New York, 2015, available at: https://doi.org/10.1201/b18023.

51 Charlie Harding, "Ushahidi has Updated Its Privacy Policy and Is GDPR Compliant", *Ushahidi Blog*, 24 May 2018, available at: www.ushahidi.com/blog/2018/05/24/ushahidi-has-updated-its-privacy-policy-and-is-gdpr-compliant.

52 The Agasti application uses PHP programming language and has two sub-applications: Mayon, for emergency personnel and resource management, and Vesuvius, for disaster preparedness and response. See the code for both projects, available at: https://launchpad.net/sahana-agasti/+series.

53 Mifan Careem, Chamindra De Silva, Ravindra De Silva, Louiqa Raschid and Sanjiva Weerawarana, "Sahana: Overview of a Disaster Management System", in Institute of Electrical and Electronic Engineers, *Proceedings of the International Conference on Information and Automation*, 15–17 December 2016, available at: https://doi.org/10.1109/ICINFA.2006.374152.

54 Khanh Ngo Duc, Tuong-Thuy Vu and Yifang Ban, "Ushahidi and Sahana Eden Open-Source Platforms to Assist Disaster Relief: Geospatial Components and Capabilities", in Alias Abdul Rahman, Pawel Boguslawski, François Anton, Mohamad Nor Said and Kamaludin Mohd Omar (eds), *Geoinformation for Informed Decisions*, Lecture Notes in Geoinformation and Cartography, Vol. 102, Springer, Cham, 2014, available at: https://doi.org/10.1007/978-3-319-03644-1_12.

55 Sahana Software Foundation, "Sahana Eden: Open Source Disaster Management Software Platform", 13 December 2011, available at: www.slideshare.net/SahanaFOSS/sahana-eden-brochure-10577413.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

registry of disaster survivors and messaging system for the reception of and automated response to emails, SMS and social networks). Moreover, Eden can visualize inputs in maps and produce automated scenario templates for managing crisis, based on predetermined resources and past experience (e.g. number of resources and employees needed, time frames). Additionally, Sahana modules are particularly relevant to COVID-19 response. They cover shelter and inventory management, which can be used to track the availability of hospital beds, quarantine centres, childcare facilities (e.g. for medical staff or patients) and medical supplies (e.g. surgical masks and COVID-19 tests). Sahana also allows incident reporting and mapping of requests for food and supplies.

Indeed, Sahana has been utilized to respond to the COVID-19 pandemic, improve data collection and coordinate volunteers globally. In the northwest of England, the county council of Cumbria[56] used Sahana to track vulnerable individuals, coordinate the distribution of protection equipment and supplies to families, and manage volunteers. Additionally, Pakistan's government used the relevant applications for supply chain and mapping to plan its logistic needs and perform case tracking.[57]

Sahana has integrated Ushahidi's functions, so it can process crowdsourced data and visualize them on a map. However, due to its ease of use, Ushahidi better fits the missions of smaller aid actors to coordinate rapid responses to disaster situations. Eden allows organizations to transfer data generated from Ushahidi[58] when they need to scale up their operation, but the opposite is not feasible automatically. In sum, Sahana is suitable for long-term projects and larger organizations, offering a vast range of options for designing, monitoring and executing disaster relief interventions. While both platforms have their benefits in the appropriate operational contexts, both come with privacy concerns depending on the data processed and the outputs produced. These concerns will be examined next.

Applying data protection in big data

The right to privacy

The information processed for big data analysis is not always personal data – i.e., data relating to an identified or identifiable natural person.[59] However, in the field of humanitarian assistance, personal data are typically processed to facilitate

56 Devin Balkind, "Sahana EDEN Used for COVID-19 Responses", *Sahana Foundation Blog*, 23 April 2020, available at: https://sahanafoundation.org/sahana-eden-used-for-covid-19-responses/.

57 Sahana Software Foundation, "Sahana Applicability for COVID-19", 20 March 2020, available at: https://tinyurl.com/rb2fpbgw.

58 Mark Prutsalis, "Developing a Service Industry to Support the Sahana Disaster Management System", *Open Source Business Resource Journal*, December 2010, available at: https://timreview.ca/article/400.

59 This definition of personal data is stated in the General Data Protection Regulation (Regulation on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC), (EU) 2016/679, 27 April 2016 (GDPR), Art. 4(1).

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

the identification of individuals in need and the recognition of patterns.[60] When aid actors perform data analysis, they usually promote a participatory model through the combination of open and crowdsourced data, especially when outputs are used to inform decision-making.[61] Despite this, the potential for abuse is high, since data analytics could lead to infringement of privacy and discrimination if proper safeguards are not adopted. While big data analysis promises insight, there is a risk of establishing a "dictatorship of data" in which communities are judged not by their actions, but by the data available about them.[62] Thus, issues of privacy must be tackled before applying big data analysis in crisis contexts.

The right to privacy is a fundamental human right recognized in international law by numerous international instruments, such as the United Nations (UN) Declaration of Human Rights, the International Covenant on Civil and Political Rights and the European Convention on Human Rights. Moreover, the UN Special Rapporteur on the Right to Privacy, whose mandate is to monitor, advise and publicly report on human rights violations and issues, plays an important role in highlighting privacy concerns and challenges arising from new technologies.[63] Additionally, an important binding legal instrument on data protection (a notion which originates from the right to privacy[64]) is the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) adopted by the Council of Europe.

In the EU context, the Charter of Fundamental Rights of the EU states that everyone has the right to respect for their private and family life (Article 7), in addition to the protection of personal data concerning themselves (Article 8). Moreover, the GDPR sets out common rules both for EU-based actors who process personal data of individuals located within or outside the EU and for actors who target their services to EU residents, irrespective of the actors' place of establishment. The focus of the remainder of this article is the GDPR, which came into force in May 2018. The reasons why we decided to analyze this legislation are threefold. Firstly, following the 2015 refugee migration crisis, multiple EU-based actors are currently implementing aid projects both in countries outside the EU and within member States, and these require the continuous processing of beneficiary communities' data. Secondly, the existing literature on the application of the GDPR to the humanitarian aid sector is limited. Thirdly, while the GDPR applies to a portion of aid actors, it is a "last-

60 European Data Protection Supervisor (EDPS), *Meeting the Challenges of Big Data: A Call for Transparency, User Control, Data Protection by Design and Accountability*, Opinion 7/2015, 19 November 2015, p. 7, available at: https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf.

61 F. Mulder *et al.*, above note 34.

62 Datatilsynet (Norwegian Data Protection Authority), *Big Data: Privacy Principles under Pressure*, September 2013, p. 7, available at: www.datatilsynet.no/globalassets/global/english/big-data-engelsk-web.pdf.

63 See Joseph A. Cannataci, *Recommendation on the Protection and Use of Health-Related Data*, UN Doc. A/74/277, December 2019; Joseph A. Cannataci, *Preliminary Evaluation of the Privacy Dimensions of the Coronavirus Disease (COVID-19) Pandemic*, UN Doc. A/75/147, 27 July 2020.

64 EDPS, "Data Protection", 2017, available at: https://edps.europa.eu/data-protection/data-protection_en.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

generation" EU law which has incorporated international data protection principles and is highly likely to set the standard and affect global regulatory trends.

When is the GDPR applicable to big data analysis?

An important question is that of when big data analysis falls within the scope of the GDPR. The Regulation applies in principle to every kind of operation and activity performed by EU-based public authorities, companies and private organizations that process personal data, regardless of the location of the individuals whose data are processed (within or outside the EU). Additionally, the GDPR applies to non-EU actors when they target their services to individuals residing in the EU.[65]

It is important to note that the GDPR does not apply to anonymous information or to personal data that is rendered anonymous, in that the data subject is no longer identifiable.[66] This includes personal data collected for humanitarian aid, provided that they can be truly anonymized. However, this is not always possible, as one cannot exclude the possibility of re-identification of individuals from other data, even when anonymization techniques have been applied. This is because anonymization is not achieved by just deleting direct identifiers, since the accumulation of different pieces of data increases the probability of re-identification. This is especially true when the target population is small and/or the subjects have a combination of rare and intrinsic characteristics. The UN Special Rapporteur on the Right to Privacy has also highlighted this risk when combining closed and open data sets.[67] Because a person's identity could be revealed by combining anonymous data with publicly available information and other data sets, de-identified data may be considered personal even after anonymization techniques have been employed. Subsequently, when NGOs attempt to anonymize personal data, they should examine whether there is a risk of re-identification. In any case, anonymization is not a one-off activity; anonymization techniques and software are constantly updated with new modules and more complex algorithms to prevent re-identification and to preserve the anonymity of data sets.

As for pseudonymized data, they fall inside the scope of the GDPR[68] because they may still be attributed to an identifiable person by the use of additional information. In a big data context, pseudonymized data may be the preferred approach given that identifiability is sometimes necessary for validating the outputs.[69] Consequently, EU-based organizations remain subject to data

65 GDPR, above note 59, Art. 3.
66 *Ibid.*, Recital 26.
67 Joseph A. Cannataci, *Report of the Special Rapporteur on the Right to Privacy*, UN Doc. A/72/540, 19 October 2017, available at: https://undocs.org/A/72/540.
68 GDPR, above note 59, Recital 26.
69 For further reading, see ICRC, *The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era*, Geneva, 2018, available at: www.icrc.org/en/download/file/85089/the_humanitarian_metadata_problem_-_icrc_and_privacy_international.pdf.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

protection rules when they analyze big data that has been pseudonymized or may be re-identified through reverse engineering.[70]

Applicable legal bases for big data analysis

According to the GDPR, a legal basis must be identified for any data processing activity. The majority of data handled by aid actors are sensitive, especially the information required for COVID-19 monitoring, which includes the processing of health data. Based on Article 9(2) of the GDPR, the applicable legal bases for aid organizations to process sensitive data are: (i) the data subject's explicit consent; (ii) protection of the data subject's vital interests and those of others who are incapable of providing consent; and (iii) public interest in the area of public health.

As mentioned, crowdsourced data – i.e., data retrieved from individuals based on their consent and on a voluntary basis – is a key data source for big data analysis. Based on Recital 32 of the GDPR, consent should be specific, freely given and informed. This means that individuals must have a clear understanding of what they are agreeing to. Consent may be expressed in writing, electronically or orally; however, silence does not imply consent.[71] The definition of "explicit" is not provided by the Regulation; in practice, it means that consent should be confirmed by a clear statement for a specific purpose, separately from other processing activities.[72] Moreover, for consent to be meaningful, data subjects need to have efficient control over their data.[73] Consent is valid until it is withdrawn and for as long as the processing activity remains the same.[74] Interestingly, we notice that while lawfulness of processing is a separate requirement to the rights of data subjects,[75] both of these GDPR requirements are interlinked for consent to be valid.

To be more specific regarding humanitarian assistance, valid consent is not just about ensuring that individuals "tick a box" to indicate their informed decision. Data subjects need to be informed about the use of their data, in a language and format they understand. Moreover, the request for consent must be direct and explicit, and an equivalent process must be available to withdraw consent. Indeed, valid consent presents many difficulties during a crisis context, due to language barriers and the complexity of data processing activities for the provision of

70 C. Kuner and M. Marelli (eds), above note 1, p. 93.
71 European Data Protection Board, *Guidelines 05/2020 on Consent under Regulation 2016/679*, 4 May 2020, available at: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf.
72 Information Commissioner's Office (ICO), *Consultation: GDPR Consent Guidance*, March 2017, available at: https://ico.org.uk/media/about-the-ico/consultations/2013551/draft-gdpr-consent-guidance-for-consultation-201703.pdf.
73 EDPS, above note 60, p. 11.
74 UNESCO, *Report of the International Bioethics Committee of UNESCO on Consent*, 2008, p. 17, available at: https://unesdoc.unesco.org/ark:/48223/pf0000178124.locale=en.
75 Data subjects' rights are analyzed further in Theodora Gazi, "Data to the Rescue: How Humanitarian Aid NGOs Should Collect Information Based on the GDPR", *International Journal of Humanitarian Action*, Vol. 5, Article No. 9, July 2020, available at: https://doi.org/10.1186/s41018-020-00078-0.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

humanitarian aid. Since aid organizations target specific communities, information about the intended big data analysis must be provided in the local language, in an understandable manner, regardless of the reader's educational level.[76] Therefore, big data analysis based on crowdsourced data may rely on explicit consent as long as individuals are properly informed about the processing purpose in a user-friendly way, such as a pop-up window or a text message with the relevant information and consent request. Consequently, aid actors' mandates to assist conflict-affected populations do not give them carte blanche to perform data processing.[77]

The debate on whether beneficiaries' consent to the use of their data is valid is not new. Indeed, when data processing is necessary for the provision of life-saving services, consent is not the appropriate legal basis. Recital 46 states that the "vital interests" legal basis may apply when actors are processing data on humanitarian grounds, such as to monitor epidemics and their spread or in situations where there is a natural or man-made disaster causing a humanitarian emergency. Indeed, data protection must never hinder the provision of assistance to vulnerable people at risk, who would be excluded from data collection when incapable of providing consent.[78] The "vital interests" basis applies when personal data must be processed in order to protect an individual's life, while the person is incapable of consenting and "the processing cannot be manifestly based on another legal basis". In these cases, big data analysis which facilitates the rapid assessment of patients' needs and their access to life-saving aid services can be based on vital interests. However, the condition of vital interest is not met when big data analysis is undertaken in non-urgent situations. Thus, processing of personal data focused on research or donor compliance cannot rely on vital interest. When data processing could be performed in a less intrusive manner, the conditions for applying this legal basis are not met.[79]

Lastly, based on Recital 54 of the GDPR, processing of sensitive data may be necessary for reasons of public interest, without acquiring the data subjects' consent. Moreover, according to Article 9, sensitive data may be processed "for reasons of public interest in the area of public health, such as protecting against serious cross-border threats". Based on the above, the "public interest" legal basis can be invoked by aid actors, for instance when they collaborate with the public authorities to support medical aid. Indeed, data processing for reasons of public health is an outcome of the State's duty vis-à-vis its citizens to safeguard and promote their health and safety. Given that public interest is determined by the States themselves, this regulatory leeway allows for States to acquire sensitive

76 Article 29 Data Protection Working Party, *Guidelines on Transparency under Regulation 2016/679*, 11 April 2018, p. 11, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.
77 Nicole Behnam and Kristy Crabtree, "Big Data, Little Ethics: Confidentiality and Consent", *Forced Migration Review*, No. 61, June 2019, p. 6, available at: www.fmreview.org/sites/fmr/files/FMRdownloads/en/ethics/ethics.pdf.
78 Lisa Cornish, "Is Data Consent in Humanitarian Contexts Too Much to Ask?", 6 August 2018, available at: www.devex.com/news/is-data-consent-in-humanitarian-contexts-too-much-to-ask-93133.
79 C. Kuner and M. Marelli (eds), above note 1.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

personal data in the context of a global pandemic. However, this basis enables data processing even when the purpose is not compatible with the data subjects' best interests. The risk of "aiding surveillance" should be highlighted as a significant concern when applying this legal basis, since big data analysis by aid actors could potentially be weaponized to achieve national security objectives.[80] As a result, actors should use this legal basis with caution, taking proportionality into consideration when asked to collect or share data with public authorities. Data-sharing agreements to regulate this data exchange and strict application of the basic data protection principles analyzed in the next section (especially data minimization and purpose limitation) are crucial to avoiding excessive data collection.

Data protection principles and big data analysis

The basic principles of data protection, set out in Article 5 of the GDPR, constitute the backbone of the legal framework when engaging in data analytics. In this section we will analyze these principles in the context of humanitarian assistance and big data analysis.

First, the Regulation requires "lawfulness, fairness and transparency". This means that apart from identifying the relevant legal basis, actors must ensure that data processing is fair and transparent. When performing big data analysis, its purpose constitutes an important factor for assessing the "fairness" and "transparency" principles – i.e., that individuals are informed about the envisioned use of their data in simple, clear language.[81] Fairness is linked to whether data will be handled in an expected way while not causing unjustified adverse effects on data subjects, individually or as a group. Equally, vulnerabilities must be considered when assessing the data subject's likely level of understanding.[82] Lack of transparency could entail that individuals have no idea or control over how their data are used.

Moreover, the GDPR refers to the principles of data minimization, storage limitation and purpose limitation. These principles were already well established in the humanitarian aid sector, long before the GDPR.[83] They specify that aid actors should limit the collection and retention of personal data to the extent that is necessary to accomplish a specific purpose. It is true that data minimization and storage limitation could clash with the key prerequisite for big data – i.e., "volume". Indeed, stockpiling personal data "just in case" they become useful clearly breaches the GDPR. However, a "save everything" approach does not

80 Ben Hayes, "Migration and Data Protection: Doing No Harm in an Age of Mass Displacement, Mass Surveillance and Big Data", *International Review of the Red Cross*, Vol. 99, No. 904, 2018, p. 193, available at: https://doi.org/10.1017/S1816383117000637.

81 ICO, *Big Data, Artificial Intelligence, Machine Learning and Data Protection*, Version 2.2, 2017, p. 20.

82 Article 29 Data Protection Working Party, above note 76, p. 11.

83 See UN Office for the Coordination of Humanitarian Affairs (OCHA), *Building Data Responsibility into Humanitarian Action*, May 2016, available at: www.unocha.org/fr/publication/policy-briefs-studies/building-data-responsibility-humanitarian-action.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

necessarily benefit big data analysis. Scholars have argued that storage of data for big data analysis is considered a thing of the past, during the present era of real-time data.[84] Additionally, appropriate data classification and clear policies on data processing improve data quality and the outputs of data science.[85] In any case, data protection "by design" solutions could involve anonymization, where possible, when personal data storage is not justifiable.

As for the purpose limitation principle, big data projects for COVID-19 have a specific aim – namely, to limit the spread of the virus and to protect public health. However, the reuse of personal data collected during humanitarian assistance may place this principle under pressure. Based on Article 5 of the GDPR, personal data "shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". This principle allows data subjects to make an informed choice about entrusting their data to the aid actor, and to be certain that their data will not be processed for irrelevant purposes, without their consent or knowledge. In some cases, determining whether big data analysis is compatible with the initial purpose might not be straightforward. In any case, the purpose must be expressly stated and legitimate – i.e., there should be an objective link between the purpose of the processing and the data controller's activities.

When big data analysis is used as a tool for policy and decision-making, an important data protection principle that must be respected is that of accuracy. This principle applies when individuals are affected by the outcome of the analysis. Big data typically processes information from diverse sources, without always verifying their relevance or accuracy. This presents several challenges. Firstly, analysis of personal data initially processed in different contexts and for other purposes may not portray the actual situation. Similarly, while working with anonymized data is less intrusive, it increases the risk of inaccuracy. As such, open data may not constitute an appropriate factual basis for decision-making, since this information may not be verified to the same degree as personal data. Predictive analysis can result in discrimination, promotion of stereotypes and social exclusion;[86] this is why big data has been accused of presenting misleading and inaccurate results that fail to consider specific particularities of the community or individuals. In any case, predictive models are inherently biased, regardless of what data they draw on. Data quality can increase the accuracy of predictive models but is not a remedy for their methodological bias.

Therefore, open and anonymous data must be selected with diligence so as to guarantee that the data processed are of the right quality and produce credible outputs. In contexts where actors largely rely on open data, best practices include giving prominence to updated and relevant data sets and enhancing cooperation between aid actors in order to encourage regular information-sharing. Furthermore,

84 Quentin Hardy, "Jeff Hawkins Develops a Brainy Big Data Company", *New York Times*, 28 November 2012, available at: https://bits.blogs.nytimes.com/2012/11/28/jeff-hawkins-develops-a-brainy-big-data-company/.
85 ICO, above note 81, p. 42.
86 EDPS, above note 60, p. 8.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

open data need to be validated via beneficiaries' inputs.[87] The combination of open data and big data analysis, through crowdsourcing, enables actors to cross-reference and triangulate the data of specific groups, understand their needs and increase the effectiveness of the operation.

Another key data protection principle is that of confidentiality, meaning that data must be sufficiently protected from unauthorized disclosure.[88] Security measures for big data are linked to the outputs of the analysis, especially when it produces more sensitive data than those included in the initial data sets. Data security is also achieved by applying the data minimization and storage limitation principles, since decreasing the collection of data reduces the risk of data breaches. Additionally, aid actors should use safe data analysis tools and train employees on their proper use. While aid actors' data sets usually uphold security standards, users have been identified as "the weakest link" for data breaches, e.g. due to loss of IT equipment and phishing scams. Encryption and pseudonymization of data sets – i.e., replacing personal identities with codes – are also promoted by the GDPR as a security measure[89] that can prevent the misuse of data.

Finally, a separate requirement introduced in Article 25 involves applying the above principles "by design and by default", along with other legal obligations. Since any data processing activity includes inherent protection risks, aid actors must always assess these risks and adopt appropriate safeguards. In any case, prior to launching big data analysis, aid actors are advised to conduct a data protection impact assessment (DPIA), as described in Article 35 of the GDPR. DPIAs[90] are a key requirement prior to activities that include the processing of sensitive data on a large scale.[91]

87 European Data Portal, *Open Data Best Practices in Europe: Learning from Cyprus, France, and Ireland*, May 2020, available at: www.europeandataportal.eu/sites/default/files/report/20200518_AR16_ODM%20Top%20Performing%20Countries_V1.1_FINAL.pdf.

88 For an in-depth analysis of the principle of confidentiality, see Kurt Schmidlin, Kerri Clough-Gorr and Adrian Spoerri, "Privacy Preserving Probabilistic Record Linkage (P3RL): A Novel Method for Linking Existing Health-Related Data and Maintaining Participant Confidentiality", *BMC Medical Research Methodology*, Vol. 15, Article No. 46, 30 May 2015, available at: https://doi.org/10.1186/s12874-015-0038-6; Loredana Caruccio, Domenico Desiato, Giuseppe Polese and Genoveffa Tortora, "GDPR Compliant Information Confidentiality Preservation in Big Data Processing", *IEEE Access*, Vol. 8, 9 November 2020, available at: https://doi.org/10.1109/ACCESS.2020.3036916.

89 GDPR, above note 59, Art. 32(1)(a).

90 Useful DPIA templates have been developed by the ICRC (see C. Kuner and M. Marelli (eds), above note 1, pp. 300–302) and the French supervisory data protection authority (Commission Nationale Informatique & Libertés, *Privacy Impact Assessment Template*, February 2018, available at: www.cnil.fr/sites/default/files/atoms/files/cnil-pia-2-en-templates.pdf).

91 A DPIA determines the relevant data sources (e.g. open data, pre-existing data sets and crowdsourcing) and the safeguards implemented to achieve fair results and compliance with the GDPR. It also contains a description of the processing activities involved, a risk analysis of the rights of data subjects, and an examination of whether anonymization is applicable; the latter is achieved by running tests to assess the probability of re-identification, especially when the group of data subjects is not sufficiently large. Dariusz Kloza, Niels Van Dijk, Simone Casiraghi, Sergi Vazquez Maymir, Sara Roda, Alessia Tanas and Ioulia Konstantinou, *Towards a Method for Data Protection Impact Assessment: Making Sense of GDPR Requirements*, Vrije Universiteit Brussel, Policy Brief No. 1, 2019, available at: https://cris.vub.be/files/48091346/dpialab_pb2019_1_final.pdf.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

Conclusions

The COVID-19 pandemic intensified existing inequities, increasing the financial insecurity of vulnerable people. This meant that the number of households in need of humanitarian support multiplied, while direct access to them became harder. Specifically, the health risks and government measures caused by COVID-19 have severely limited traditional methods for primary data collection, such as conducting household visits, field assessments and focus group discussions.[92] Aid actors can address these major challenges by applying big data analysis to continue their operations and monitor their response to the pandemic. Big data has been defined as a technological phenomenon which relies on the interplay of technology (the use of computation power and algorithmic accuracy to link and compare large volumes of data) and analysis (the recognition of patterns in order to predict behaviours, inform decisions and produce economic or social indicators).[93]

Indeed, the continuation of humanitarian assistance and the monitoring of epidemic responses can be facilitated by technological innovations. As with any technological tool, big data may support disaster management responses, provided that its use does not derail humanitarian efforts or harm beneficiaries' rights. The UN Office for the Coordination of Humanitarian Affairs (OCHA)[94] has underlined that using big data for humanitarian purposes is one of the greatest challenges and opportunities of the network age. Big data is addressed in this context with the possibility of predicting, mapping and monitoring COVID-19 responses.[95]

The belief that big data is a "panacea for all issues" is the main cause of concern expressed by scholars.[96] Big data analysis entails privacy risks and may produce biased results, leading aid actors to misguided decisions and inequity in the provision of humanitarian assistance.[97] Aid actors should be mindful of the shortcomings of both big data and open data. First, both categories often lack demographic information that is crucial for epidemiological research, such as age and sex. Second, this data represents only a limited portion of the population – i.e.,

92 UNHCR, above note 12.

93 Danah Boyd and Kate Crawford, "Critical Questions for Big Data", *Information, Communication and Society*, Vol. 15, No. 5, 2012, pp. 662–663.

94 See OCHA, *Humanitarianism in the Network Age*, OCHA Policy and Study Series, Geneva, 2013; OCHA, above note 83.

95 Alana Corsi, Fabiane Florencio de Souza, Regina Negri Pagani and João Luiz Kovaleski, "Big Data Analytics as a Tool for Fighting Pandemics: A Systematic Review of Literature", *Journal of Ambient Intelligence and Humanized Computing*, 29 October 2020, available at: https://doi.org/10.1007/s12652-020-02617-4; Pravin Kumar and Rajesh Kr Singh, "Application of Industry 4.0 Technologies for Effective Coordination in Humanitarian Supply Chains: A Strategic Approach", *Annals of Operations Research*, 3 January 2021, available at: https://doi.org/10.1007/s10479-020-03898-w.

96 See UN Global Pulse, above note 8, pp. 24–34; Miguel Luengo-Oroz, "10 Big Data Science Challenges Facing Humanitarian Organizations", UNHCR Innovation Service, 22 November 2016, available at: www.unhcr.org/innovation/10-big-data-science-challenges-facing-humanitarian-organizations; Iffat Idris, *Benefits and Risks of Big Data Analytics in Fragile and Conflict Affected States*, 17 May 2019, available at: https://tinyurl.com/5xmftqcy.

97 C. Kuner and M. Marelli (eds), above note 1, p. 93.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press

excluding marginalized and under-represented groups such as infants, illiterate persons, the elderly, indigenous communities and people with disabilities – while potentially under-representing some developing countries where digital access is not widespread.[98] Third, specifically for the COVID pandemic, short-term funding of big data projects does not acknowledge the long timelines required to measure health impact. During emerging outbreaks, aid agencies may lack accurate data about case counts, making it challenging to adapt decision-making models.[99] Finally, capacity-building of aid workers in information management is a prerequisite for them to develop the necessary know-how in applying data analysis.

As a matter of law, aid actors must adopt a privacy-first approach with any data collection methods implemented. For crowdsourced data, they must provide adequate information to data subjects, so that their consent is meaningful, instead of an illusory choice. When they use personal data collected for different purposes, they must check that further data processing is compatible and whether anonymization can apply. Failure to address these issues may compromise compliance with core data protection principles.

Though there are many challenges and risks involved, aid actors should adopt technological innovations such as big data in order to address the impact of the pandemic. Past big data projects could serve as case studies for identifying best practices and lessons learned. In any case, humanitarians must ensure that they "do no harm" – i.e., that big data does not cause or exacerbate power inequities. The European Data Protection Board has stressed the importance of protecting personal data during the COVID-19 pandemic, but it has also noted that "[d]ata protection rules … do not hinder measures taken in the fight against the coronavirus pandemic".[100] Even in the context of a pandemic, there is no real dilemma between an effective or a GDPR-compliant use of data. The GDPR does introduce exceptions (e.g. vital interest basis) so as not to hinder access to aid services, while ensuring that privacy principles are respected. Robust data protection policies and practices should help aid actors to mitigate the challenges of big data. Finally, any measure to address the COVID-19 pandemic should be consistent with the aid actor's mandate, balancing all relevant rights, including the rights to privacy and health.

98 Shweta Bansal, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani and Cécile Viboud, "Big Data for Infectious Disease Surveillance and Modeling", *Journal of Infectious Diseases*, Vol. 214, No. 4, 14 November 2016, available at: https://doi.org/10.1093/infdis/jiw400.

99 Caroline Buckee, "Improving Epidemic Surveillance and Response: Big Data Is Dead, Long Live Big Data", *The Lancet Digital Health*, Vol. 2, No. 5, 17 March 2020, available at: https://doi.org/10.1016/S2589-7500(20)30059-5.

100 European Data Protection Board, "Statement by the EDPB Chair on the Processing of Personal Data in the Context of the COVID-19 Outbreak", 16 March 2020, available at: https://edpb.europa.eu/news/news/2020/statement-edpb-chair-processing-personal-data-context-covid-19-outbreak_en.

https://doi.org/10.1017/S1816383121000084 Published online by Cambridge University Press