UNHCR staff answer questions and make appointments for refugees at a help desk in the Jordanian city of Zarqa. With approximately three quarters of Jordan's refugee population living in urban settings, help desks like this one provide critical information to refugees about protection, registration and health care.

7 Open Statistics: Improving Response Through Statistical Transparency

Introduction

THE 2012 STATISTICAL YEARBOOK was accompanied by the launch of the revised PopStats website, providing access to official historical statistics for public discovery and exploration.[95] UNHCR has made statistics available through the Internet since 2006. Structured data such as Microsoft Excel files are also accessible, providing additional access to these valuable statistics. A more recent development is the targeting of the exchange of this information between computers through an application programming interface (API), which provides programmers with a stream of dynamic data for their applications, visualizations, and data repositories.[96]

At first glance, these developments may seem obscure, trivial, or of little interest to the way the humanitarian field responds to the needs of the populations they aim to serve. In fact, this willingness and action to offer statistics to the public, and to other computers, is a fundamental shift. Indeed, this process has the potential to change the future of humanitarian work and is already impacting on operations. Imagine a future in which statistics are updated by governments, UN agencies, and NGOs and then are instantly available – both to those who need them to take action and to computers that can consume, 'visualize', and combine them in real time.

This chapter will explore this idea of 'open data',[97] focusing on how it specifically impacts on data and statistics in the humanitarian field. The concepts and principles behind open data are already impacting on governments, industries, and financial institutions, with academic and research entities developing specific curriculums to train the next generation of data scientists in this emerging field.

A paradigm shift

Over the past half-decade, there has been a fundamental shift in perspective on the old adage that 'information is power'. Information, statistics, and the data that underlay them are no less powerful than they have been in the past. But today, this power is being derived from an agency's or individual's ability to share quality data and statistics.

'Big data', a frequently used catchphrase, is not open data and 'open' here does not mean free of cost. This chapter does not go into the concepts around big data in detail, but the reader should be aware that big data can have varying degrees of openness. An example of big data would be the Human Genome Project and the resulting digital storage of DNA-related information, which has made very large amounts of data available to researchers around the world. In contrast, a survey of 100 households, while formidable from the standpoint of data volume, would not be considered big data. However, if the raw data from household surveys per country and per sector were to be made openly available across the globe, then these could be considered big data and could in turn be fully leveraged for decision-making.

The adoption of open data approaches has the potential to create big data from smaller, dispersed sources. The humanitarian field is not the only field that is struggling to harness big data. For instance, the Large Hadron Collider, in Europe, is only able to keep

95 http://popstats.unhcr.org/

96 http://data.unhcr.org/wiki/index.php/API_Documentation

97 For the purpose of this chapter, 'open data' and 'open statistics' will be used interchangeably. While statistics are often compiled from multiple data elements, if these data adhere to open data principles then the statistics can be considered 'open'.

UNHCR / J. KOHLER

0.0002 per cent of the data this project has collected.[98] In this way, the private sector, researchers, and humanitarians are all currently changing the ways they leverage the large amounts of data that are increasingly being opened up for use.

UNHCR provides support to some of the most vulnerable populations in the world. Dealing with personal, individual-level data requires that rigorous processes and confidentiality protocols are in place to protect this information and maintain the trust of those UNHCR serves. This is of paramount importance in UNHCR, and strict procedures and guidance exist to ensure that data are protected and managed appropriately.

Other humanitarian agencies likewise recognize the importance of data protection. The Office for the Coordination of Humanitarian Affairs (OCHA) has advocated that 'Ensuring data security, developing robust guidelines for informed consent and tackling the ethical questions raised by open data' are key to fully embracing the power of open data. Additionally, OCHA calls for a charter or code of conduct to be in place by the end of 2015.[99] The relatively new International Committee of the Red Cross 2013 professional standards for protection work offer another strong example of clarifying the responsible use and governance of open data.[100]

Can data or statistics adhere to open data concepts and remain confidential? In part, these two concepts seem to be at odds with one another. How can we open our statistics and data but respect fundamental principles of confidentiality? For example, could confidential information about a vulnerable individual be shared in an open way among trusted and validated users who are approved to act on this information? The fact that statistics could be exchanged more predictably, processed more quickly, and acted upon by a trusted user group more efficiently could greatly improve the quality of service provided to individuals while adhering to confidentiality protocols. This is an area that needs to be considered carefully, but one that could have a profound impact.

So what are open statistics, anyway?

The term 'open data' has entered mainstream vocabularies, and by today is used widely when discussing anything related to data. Yet with this wide use has come differing understandings on what exactly is meant by open data, with most people of the view that the term has something to do with easy access to data or statistics. While this is true, not all open data are created equal – or, in this case, 'opened' equally.

Tim Berners-Lee, a driving force behind the creation of the World Wide Web, provides perhaps the clearest definition of what is meant by open data or statistics through his star (*) system.[101] This system, which also rates openness, works as follows:

- * Make your data or statistics available on the Web (in whatever format) under an open license
- ** Make these available as structured data (e.g. use Excel instead of an image scan of a table)
- *** Use non-proprietary formats (e.g. CSV instead of Excel)
- **** Use URIs to denote things, so that people can point at your content
- ***** Link your data to other data to provide context

For many readers, these instructions made a great deal of sense until the fourth and fifth stars, where the concepts may have exceeded their technical knowledge. Essentially, the concepts expanded on under the latter two stars embrace machine-readable, structured data with standards that guide their use. This is a simplification of these concepts, of course, and there is a great deal of work needed in this area to define the best technical solution to achieve the goal of fully open data.

Berners-Lee sums up the idea behind his star ratings system this way: 'These tools will let the Web – currently similar to a giant book – become a giant database.' Essentially, this is an evolution from data that people read to data that our computers can read so people can focus on the broader story (i.e. analysis, exploration, and meaning). While the technical solution proposed in the fourth and fifth stars is one of the potential options, others do exist.[102] The result of any solution would be structured machine-readable – meaning, computer-readable – data with licenses that allow for re-use.[103]

If this star ratings system were used as a metric to describe the degree of openness for datasets or statistics, many individuals would quickly discover that a favourite data source is not as open as initially envisioned. At first glance, the degree to which data or statistics are open may not seem to be that important. When interacting with these statistics, however, a user realizes quickly that relatively closed systems pose significant limitations that, at best, restrict use and, at worst, completely block a user from doing anything with the data. An example of this could be a PDF document that has all the data needed to do a great analysis in an annex at the back of the document. After struggling to get the data out of the annex and into a format that can be used, however, users quickly realize that the data were not as open as they had originally thought.

Finally, it is important to understand that statistics can include raw data and summary statistical information, which may combine one or more data elements. If a user has open access to the underlying data, then the statistics are considered to be open.

98 Open Data Initiative (ODI) Summit 2014, open data quiz, page 14.

99 *Humanitarianism in the Network Age*. See: http://ow.ly/EQS8M

100 https://www.icrc.org/eng/assets/files/other/icrc-002-0999.pdf

101 http://5stardata.info/

102 XML is a syntax and its data model is a tree. RDF is a data model based on a graph that uses URIs and has several different syntax, including an XML syntax. Nevertheless, both XML and RDF can be used to represent structured data on the web and move data around between applications. See: http://semanticweb.com/introduction-to-rdf-vs-xml_b31071

103 Creative Commons Attribution. See: http://creativecommons.org/licenses/by/4.0/

Why are open statistics important?

There are many benefits that can be derived from the adoption of open statistics. The list below is not exhaustive, but highlights some of the major points.

1. Data and derived statistics produced to benefit individuals belong to humanity. Typical examples are genomes, data on organisms, medical science, and environmental data following the Aarhus Convention.[104]
2. Article 19 of the UN Universal Declaration of Human Rights states that 'Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.'
3. Facts cannot legally be copyrighted and thus should be shared openly.
4. Data are required for the smooth running of communal human activities and are an important enabler of socio-economic development (health care, education, economic productivity, etc.).[105]
5. In research, the rate of discovery is accelerated by better access to data.[106] While this has traditionally been applied to hard science, the social sciences derive similar benefits.

Some practical examples of open statistics impact

Transparent Financial Information

THE INTERNATIONAL AID TRANSPARENCY INITIATIVE (IATI)

Countries impacted by a humanitarian crisis face huge challenges in accessing up-to-date information on financial aid and impact. Individuals in both affected countries and donor countries lack the information they need to hold their governments to account for the use of those resources. IATI aims to address these challenges by making information about aid spending easier to access, use, and understand.[107] IATI brings together donor and recipient countries, civil society organizations, and other experts in aid information who share the aspirations of the original IATI Accra Statement and are committed to working together to increase the transparency of aid.[108]

JORDAN: IATI COMPLIANCE AND RESPONSE

With well over half a million Syrian refugees in Jordan, there have been appeals and subsequent funding to support operations.[109] The IATI work exposes statistics and funding information related to these operations, and in turn proposes standards for the collection and sharing of this information. While first envisioned to provide unfettered access to these statistics, this process has actually highlighted gaps in this information, largely due to issues of compliance in reporting. This is an excellent illustration of how the adoption of open data can allow operations to lead by example. While additional steps are needed to meet the goal of full transparency or open data for humanitarian financing, the hard work of putting in place standards and principles is moving forward.

Open Data Initiatives Around Humanitarian Statistics

While opening statistics is revolutionizing access to and re-use of these data, there are still some obstacles to the free flow of information across institutions, sectors, and emergency responses. This is related to documentation of the data – i.e. the methods, caveats, accuracy etc. – collectively referred to as metadata. This is a common language that defines each data element clearly so that comparisons, calculations, and collations are accurate. While the International Organization for Standardization (ISO) is a global body working in this area, the humanitarian sector also has a few initiatives that specifically target humanitarian and development datasets.[110]

104 http://www.unece.org/fileadmin/DAM/env/pp/documents/cep43e.pdf
105 http://sciencecommons.org/about/towards
106 http://drexel-coas-elearning.blogspot.ch/2006/09/open-notebook-science.html
107 http://www.aidtransparency.net/
108 *Ibid.*
109 http://d-portal.org/ctrack.html?country=JO#view=main
110 http://www.iso.org/iso/home/about.htm

STATISTICAL DATA AND METADATA EXCHANGE (SDMX)

Under-Secretary General Sha Zukang stated: 'The United Nations system has accumulated over the past 60 years an impressive amount of information. UNdata, developed by the UN Statistics Division,[111] is a powerful tool which will bring this unique and authoritative set of data not only to the desks of decision-makers and analysts, but also to journalists, to students and to all citizens of the world.'[112]

This work has indeed opened an incredible amount of development and humanitarian data for public use, and has been a major step forward in data transparency in the United Nations. To envision the scale of this work, the associated databases, tables, and glossaries are estimated to contain over 60 million data points, covering a wide range of themes.[113]

The Statistical Data and Metadata Exchange (SDMX) takes the work achieved by UNdata one step further by providing standard definitions, structure, metadata, and exchange protocols related to these statistics. Such standards are of paramount importance for the use of these statistics. Now that the data have been opened up through UNdata, SDMX ensures that these data can be appropriately collated, combined, and defined, with metadata that clearly describe the methods, confidence, and overall quality.

THE HUMANITARIAN EXCHANGE LANGUAGE (HXL)

HXL is a project led by OCHA and supported by other UN agencies, with UNHCR a significant contributor. The overarching goal is to refine the management and exchange of data for emergency response. Data exchange in this field, which often has to deal with chaotic environments due to natural disaster or armed conflict, still takes place through labour-intensive sharing, with often poorly defined processes and standards. The goal of HXL is to automate many of these processes, saving valuable time for staff in the field and improving the flow of information for decision-makers who need to allocate resources for response activities.[114] The HXL 'common language' is leading the effort to collect data that are predictable, with the final goal being better interoperability.

Open Mapping Data

OPENSTREETMAP (OCM)

OpenStreetMap is built by a community of mappers that contribute and maintain data about basic infrastructure, locations of interest, and much more across the globe. OpenStreetMap emphasizes local knowledge, empowering communities, refugees, and others to map the world.

Essentially, OCM is similar to a 'wiki' used not for text-based information but for geographic data.[115] Contributors can use aerial imagery, the Global Positioning System (GPS) now a part of many mobile devices, and low-tech field maps to collect and verify geographic information. 'OpenStreetMap is *open data*: you are free to use it for any purpose as long as you credit OpenStreetMap and its contributors,' the project explains. 'If you alter or build upon the data in certain ways, you may distribute the result only under the same license.'[116]

ZA'ATARI REFUGEE CAMP: OPEN GEOGRAPHIC DATA AND STATISTICS

With the ongoing crisis in the Syrian Arab Republic, bordering countries have received large influxes of refugees over the past three years. In Za'atari, located in northern Jordan, refugee populations have been as high as approximately 200,000 in mid-2013, despite having had only a very sparse population at the beginning of 2012.

This radical growth over such a short period has required radical methods to collect statistics and geographic information, to provide context to this information, and to plan the growth of this camp at a pace that served the needs of this vulnerable population. Adherence to open statistic and geographic data collection and sharing has supported this pace of development and added value over time by empowering communities, partners, and individuals to maintain and update information.

111 Department of Economic and Social Affairs, United Nations, New York.

112 http://data.un.org/

113 http://data.un.org/Host.aspx?Content=About

114 OCHA: *Developing Humanitarian Data Standards: An introduction and plan for 2014* (http://docs.hdx.rwlabs.org/wp-content/uploads/HXL_Paper-forsite.pdf). See: http://docs.hdx.rwlabs.org/standards/

115 http://en.wikipedia.org/wiki/Wiki

116 http://www.openstreetmap.org/about

Fig. 7.1 **Za'atari refugee camp, from above and OpenStreetMap view**

Conclusion

Open data and statistics concepts and practices applied to the humanitarian field have the potential to fundamentally change response and to impact on the most vulnerable populations in the world. Economically, open data are of great importance. Several studies have estimated the economic value of open data at several tens of billions of euros annually in the European Union alone. New products and companies are re-using open data.[117]

Transparency in activities, actions, and financing for humanitarian response will improve informed decision-making and accountability of actions. By definition, open data are social, networked, and collaborative, all of which are key attributes that must be inherent in a humanitarian response for this idea to be efficient, well informed, and timely.

Open statistics often get pushed to technologists, statisticians, and data scientists. Yet the main limitations to adoption are related to institutional change and shifts in mindsets to different business models and governance of data management. In particular, these need to be based on a fundamental understanding of the benefits that can be derived from adopting open data principles.

Users, field practitioners, statisticians, and drivers of governance and technology in this realm are convinced of the benefits, as they have a direct and immediate impact on their work. To reach the full potential of open statistics, individuals and agencies need to understand the concepts and principles outlined in this chapter and invest and advocate for their adoption. The humanitarian world is currently in the early adoption phase of this work, with many agencies investing in exposing data. While this is a necessary first step, there is enormous potential for growth when a critical mass of data begins to be better leveraged for informed decision-making and action. ■

117 *Open Data Handbook* © 2010–2012, Open Knowledge Foundation. Licensed under Creative Commons Attribution (Unported) v3.0 License.