OPEN DATA FOR DEVELOPING ECONOMIES CASE STUDIES

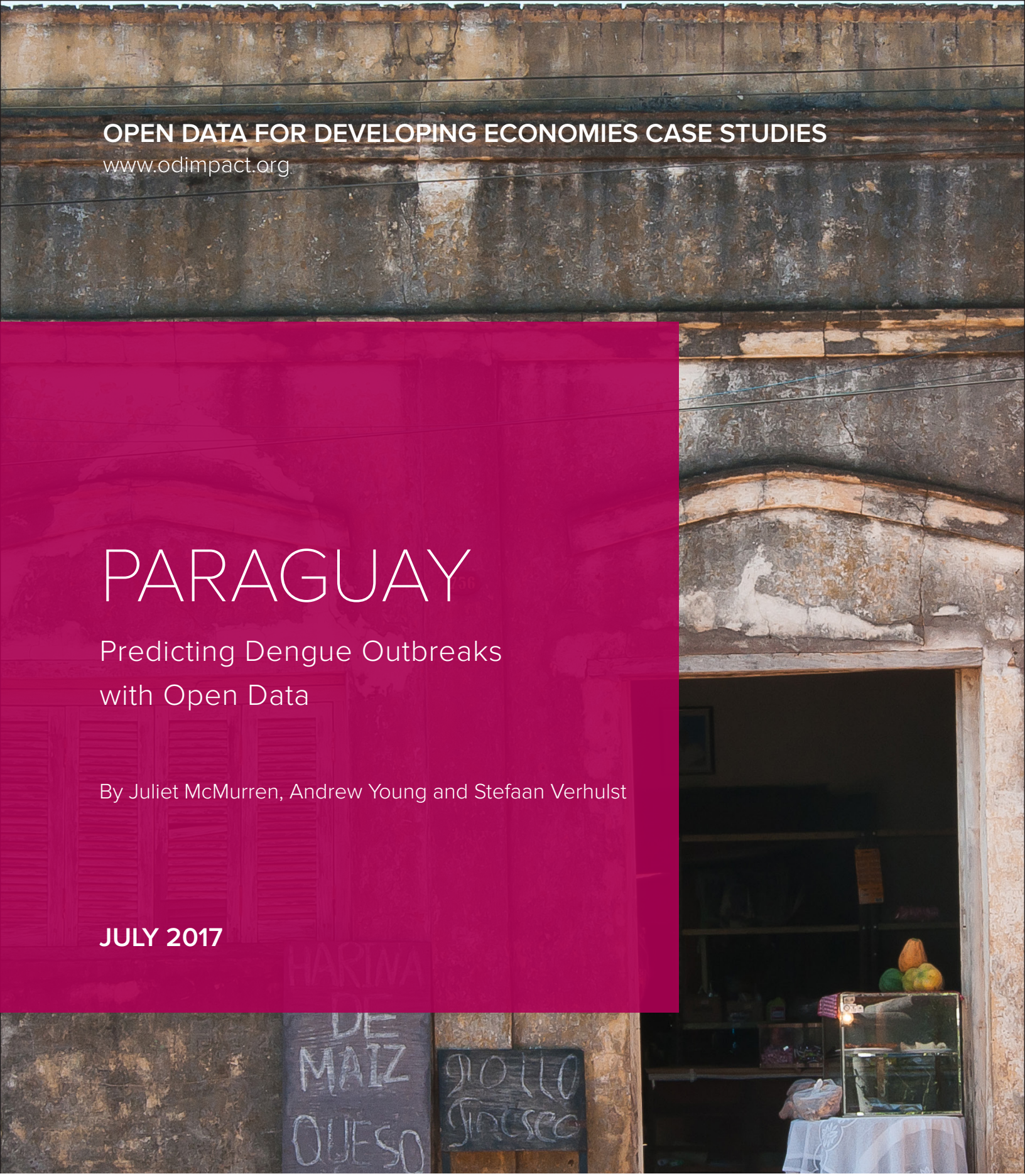
www.odimpact.org

PARAGUAY

Predicting Dengue Outbreaks with Open Data

By Juliet McMurren, Andrew Young and Stefaan Verhulst

JULY 2017



USAID FROM THE AMERICAN PEOPLE

mobile Solutions Technical Assistance and Research

fhi360 THE SCIENCE OF IMPROVING LIVES

GOVLAB

OPEN DATA FOR DEVELOPING ECONOMIES CASE STUDIES*

www.odimpact.org

PARAGUAY

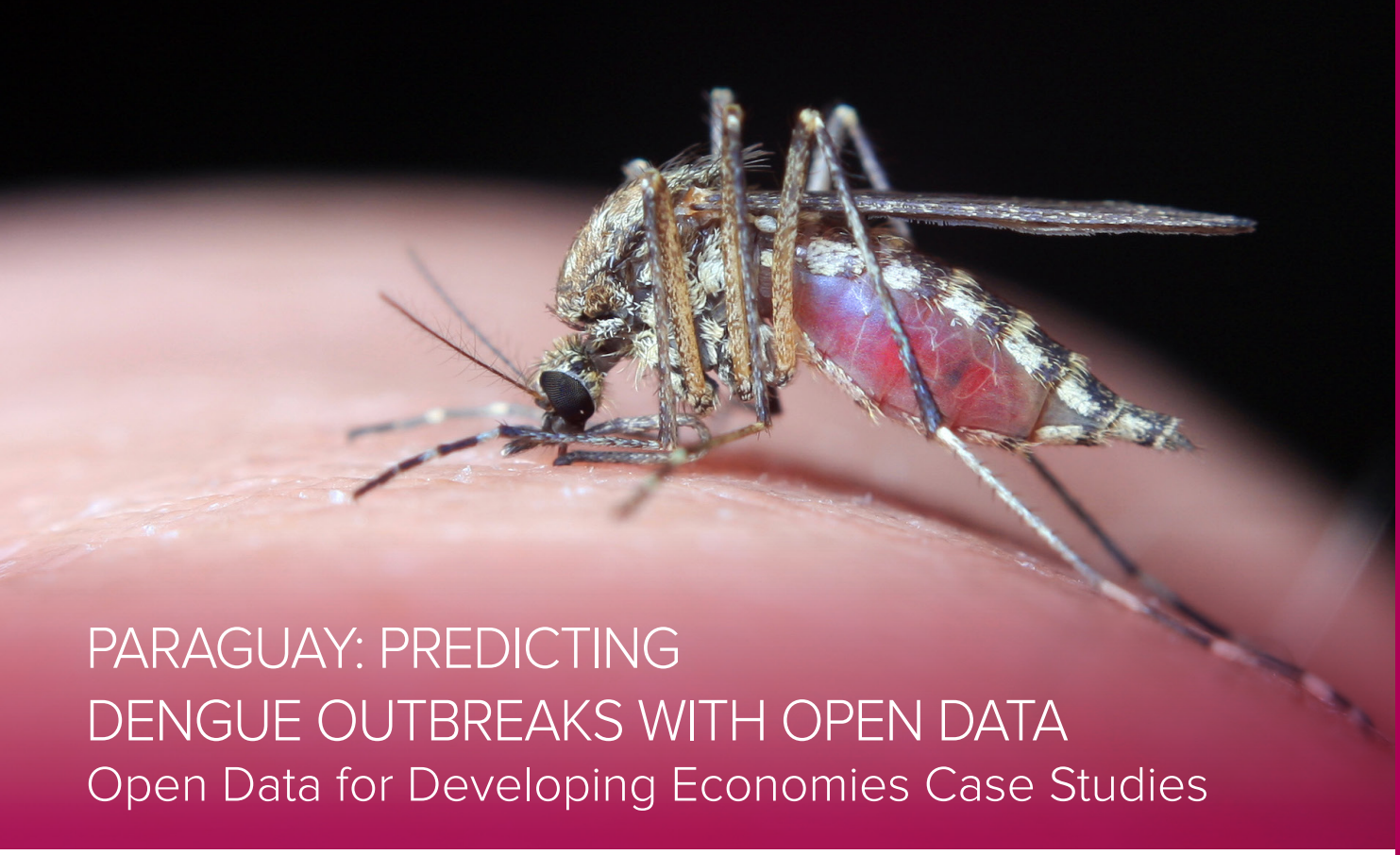
Predicting Dengue Outbreaks with Open Data

By Juliet McMurren, Andrew Young and Stefaan Verhulst**

JULY 2017

* Project conducted in collaboration with the Web Foundation, United States Agency for International Development (USAID), and the Mobile Solutions, Technical Assistance and Research (mSTAR) program at FHI 360.

** "Special thanks to Akash Kapur who provided crucial editorial support for this case study, and to the peer reviewers [odimpact.org/about] who provided input on a pre-published draft."



PARAGUAY: PREDICTING DENGUE OUTBREAKS WITH OPEN DATA

Open Data for Developing Economies Case Studies

SUMMARY

Dengue Fever has been endemic in Paraguay since 2009. Recognizing that the problem was being compounded by the lack of a strong system for communicating dengue-related dangers to the public, the National Health Surveillance Department of Paraguay opens data related to dengue morbidity. Leveraging this data, researchers created an early warning system that can detect outbreaks of dengue fever a week in advance. The data-driven model can predict dengue outbreaks at the city-level in every city or region in Paraguay—as long as data on morbidity, climate and water are available.

CONTEXT AND BACKGROUND

PROBLEM FOCUS / COUNTRY CONTEXT

Paraguay is a tropical to subtropical country of 6.7 million inhabitants, of whom almost a third live in the capital, Asunción.[1] Following several decades of rapid economic growth, the 2015 UN Human Development Index classifies it as a country of medium human development,[2] and the World Bank now considers it an upper middle income nation.[3] The percentage of the Paraguayan population living below the poverty line has declined sharply over the last two decades, from 49 percent in 2002 to 22.2 percent in 2015.[4]

While most of Paraguay's urban population has access to clean drinking water, rural and/or indigenous communities are frequently reliant on surface or rainwater, raising the risk of water- and mosquito-borne disease.[5] In 2013, the Millennium Development Goals Fund reported that only 6 percent of Paraguay's indigenous households had access to drinking water, and only 3 percent had adequate sanitation. Furthermore, only 10 percent of Paraguay's sewage was treated.[6]

1 Wikipedia, "Paraguay," https://en.wikipedia.org/wiki/Paraguay.
2 United Nations Development Program, "Human Development Index," Human Development Reports, http://hdr.undp.org/en/content/human-development-index-hdi.
3 World Bank, "World Bank Country and Lending Groups," https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.
4 World Bank, "Data: Paraguay," http://data.worldbank.org/country/paraguay.
5 Natalia Ruiz Diaz, "Paraguay: Clean Water Out of Reach for Native Peoples," Inter Press Service,June 29, 2010, http://www.ipsnews.net/2010/06/paraguay-clean-water-out-of-reach-for-native-peoples/.
6 Millenium Development Fund Achievement Goals, "Paraguay," http://www.mdgfund.org/country/paraguay.

Dengue is a mosquito-borne tropical infection caused by four viruses (DENV-1, DENV-2, DENV-3, and DENV-4) in the *Flavivirdae* family. These viruses are transmitted by infected female *Aedes aegypti* and *Aedes albopictus* mosquitoes that feed diurnally both indoors and outdoors, and breed in settings with standing water (including in puddles, water tanks, containers and old tires), poor sanitation, and a lack of garbage collection. The mosquitoes that transmit dengue are endemic in parts of Central and South America, Africa, Asia, and Oceania, with most cases occurring during the rainy season or warmer months in urban and suburban areas.[7] Up to 100 million people worldwide contract dengue each year, with 500,000 developing severe illness and 22,000 dying. Some 2.5 billion people live in dengue-endemic areas. Worldwide, cases of dengue have increased thirtyfold since 1960, driven by urbanization, population growth, increased international travel, and climate change.[8]

Dengue fever is asymptomatic in as many as 50 percent of those infected, while a further minority, particularly among the young and those contracting dengue for the first time, experience an undifferentiated fever only.[9] Symptoms of dengue, which appear four to seven days after a bite, include a sudden high fever lasting two to seven days; headache and pain behind the eyes; muscle, joint, and bone pain; and skin rash and bruising. Treatment consists of supportive care, and no antiviral treatment is available.[10] In severe cases, patients may progress to Dengue Hemorrhagic Fever (DHF), with severe abdominal pain, vomiting, diarrhea, convulsions, bruising, and uncontrolled bleeding. Complications can lead to potentially fatal circulatory system failure and shock, also known as Dengue Shock Syndrome (DSS). Dengue infection confers immunity to future infections with the same virus serotype, and a transient immunity to other serotypes. Once that transient immunity passes, however, patients contracting other dengue serotypes are at increased risk of developing DHF.[11]

Dengue has been declared endemic in Paraguay since 2009.[12] The Pan American Health Organization reported that there were over 173,000 probable cases of dengue for the year 2016, with 48 cases of severe dengue and 16 deaths.[13] The Direccion General de Vigilancia de la Salud (DGVS) (National Health Surveillance Department of Paraguay) heads up the country's prevention and response efforts.

7 International Association for Medical Assistance to Travellers, "Country Health Advice: Paraguay," https://www.iamat.org/country/paraguay/risk/dengue.
8 Wikipedia, "Dengue Fever Outbreaks," https://en.wikipedia.org/wiki/Dengue_fever_outbreaks
9 Centers for Disease Control and Prevention, "Clinical Guidance: Dengue Virus," Updated September 6, 2014, http://www.cdc.gov/dengue/clinicallab/clinical.html.
10 International Association for Medical Assistance to Travellers, "Country Health Advice: Paraguay," https://www.iamat.org/country/paraguay/risk/dengue.
11 Ibid.
12 Juan Pane, Julio Paciello, Verena Ojeda, Natalia Valdez, "Enabling dengue outbreak predictions based on open data," Open Data Research Symposium Draft Paper, October 5, 2016, https://drive.google.com/file/d/0B4TpC6ecmrM7Q1lpQ0xoNlJnZlU/view.
13 "Number of Reported Cases of Dengue and Severe Dengue (SD) in the Americas, by Country: Figures for 2016," Pan American Health Organization, World Health Organization, February 6, 2017, http://www.paho.org/hq/index.php?option=com_docman&task=doc_download&Itemid=270&gid=37782&lang=en.

OPEN DATA IN PARAGUAY

Paraguay ranked 62nd in the 3rd Open Data Barometer, ahead of Venezuela but behind the majority of Latin American countries, including Argentina (52nd), Peru and Costa Rica (44th), and Colombia (28th). Paraguay's ranking is largely the result of low scores regarding government policies and government action related to open data.[14] The Open Knowledge Foundation's Open Data Index ranked it 50th worldwide in 2015, moving down from its previous ranking of 41 in the 2014 Index. Its open data on procurement tenders and government budget information received high marks, but many other datasets from sectors like the environment and company registers were non-existent or low quality.[15]

KEY ACTORS

KEY DATA PROVIDERS

Direccion General de Vigilancia de la Salud (DGVS) (National Health Surveillance Department of Paraguay): DGVS is the agency responsible for the prevention and control of epidemic disease in Paraguay. It collects and publishes data on disease outbreaks and morbidity.[16]

KEY DATA USERS AND INTERMEDIARIES

Juan Pane: A researcher at the Facultad Politecnica-Universidad de Asunción with an interest in open data and open government, Juan Pane leads a team seeking to develop data models to provide early warning of dengue outbreaks in Paraguay. He also works for a democracy initiative funded by USAID assisting the Paraguayan government with transparency portals. Paraguayan by birth, Pane completed a doctorate in computer science at the University of Trento, Italy in 2012, followed by a postdoctoral fellowship. He returned to Paraguay with his family in 2013, just as the country was experiencing a dengue epidemic, with 150,000 reported cases and 233 deaths.[17] Pane reports that the probability of acquiring dengue in some Asunción neighborhoods that year was as high as one in four, a rate that filled him with alarm for his family, but also motivated him to find ways to address the problem of dengue.[18]

Iniciativa Latinoamericana por los Datos Abiertos ILDA): ILDA, a network of NGOs and research organizations focused on Latin America, played a key enabling and funding role for the initiative studied here. ILDA's "overarching objective" is to "strengthen the accountability and legitimacy of public institutions, improve public services, and fuel economic growth in Latin

14 World Wide Web Foundation, Open Data Barometer, Third Edition, WWWF, April 2016, http://opendatabarometer.org/3rdedition/regional-report/latin-america/.
15 "Paraguay," Global Open Data Index 2015, http://index.okfn.org/place/paraguay/.
16 Juan Pane, Julio Paciello, Verena Ojeda, Natalia Valdez, "Enabling dengue outbreak predictions based on open data," Open Data Research Symposium Draft Paper, October 5, 2016, https://drive.google.com/file/d/0B4TpC6ecmrM7Q1lpQ0xoNlJnZlU/view.
17 World Bank, "The Dengue Mosquito Bites and Makes Latin America Sick," World Bank News, April 7, 2014, http://www.worldbank.org/en/news/feature/2014/04/07/dengue-en-latinoamerica.
18 GovLab interview with Juan Pane, September 9, 2016.

America and the Caribbean through research and innovation on open data initiatives."[19]

KEY BENEFICIARIES

The direct key beneficiary was DGVS itself, since the data model provided an early warning system of future demands on the healthcare system. Beyond that, Pane intended to help the people of Paraguay: "Dengue doesn't distinguish between a government minister and my child. Mosquitoes don't care who they bite. I don't want *anyone* to get dengue."[20]

PROJECT DESCRIPTION

INITIATION OF THE OPEN DATA ACTIVITY

DGVS collects and publishes incidence and morbidity data on dengue outbreaks in Paraguay. Despite the presence of this data, DGVS lacks an automated predictive tool to enable it to predict dengue outbreaks. In 2013, shortly after returning to Paraguay from his doctoral studies in Italy, researcher Juan Pane and his colleagues at Facultad Politecnica-Universidad de Asunción noted that there was no open source tool available that could be adapted for this purpose by DVGS, nor had any work been done to examine the correlation between incidence of dengue in Paraguay and variables such as climate, cartography, and population.[21]

Pane's initial hope was to build dynamic maps using the published data to show the origin and spread of outbreaks. He quickly found, however, that the available data would not support this type of granular geo-spatial tracking.[22] Looking to other dengue-affected countries in Latin America for examples of disease modeling, he found that the few other countries where data was collected, such as Brazil, had similar problems with inadequate granularity and comparability of data, creating major obstacles to longitudinal analysis that could inform predictive modeling. He successfully applied to Iniciativa Latinoamericana por los Datos Abiertos (ILDA), a Latin American open data research, funding and advocacy network, for a research grant to study data modeling of dengue. He and his colleagues then defined the required epidemiological variables and co-variables such as climate, geographic and demographic information, and surveyed 30 dengue-affected countries to assess the availability and format of published dengue data, as well as relevant government agencies responsible for publishing such data.[23] Pane and his team surveyed the reporting forms used throughout Latin America, identifying 285 variables collected across the 30 countries. Finally, Pane's team reviewed literature to identify those variables necessary to model dengue incidence.[24]

19 "About ILDA," Iniciativa Latinoamericana por los Datos Abiertos, http://idatosabiertos.org/about-ilda/.
20 Ibid.
21 Juan Pane, Julio Paciello, Verena Ojeda, Natalia Valdez, "Enabling dengue outbreak predictions based on open data," Open Data Research Symposium Draft Paper, October 5, 2016, https://drive.google.com/file/d/0B4TpC6ecmrM7Q1lpQ0xoNlJnZlU/view.
22 GovLab interview with Juan Pane, September 9, 2016.
23 GovLab interview with Juan Pane, September 9, 2016.
24 Ibid.

Pane's team then correlated the dengue incidence data with open climatic, geographic, demographic, and sanitation data, and produced a prototype model which was shared with DGVS. The open source web application allowed DGVS to incorporate collected data on a weekly basis and produce early warning maps of predicted dengue incidence for the following week.[25]

DEMAND AND SUPPLY OF DATA TYPE(S) AND SOURCES

Pane's team used existing DGVS data on dengue incidence. The data, which was being collected on forms to report confirmed or probable cases of notifiable diseases to DGVS for subsequent reporting to the World Health Organization, provided information on number of cases, incidence of the four dengue serotypes, and demographics and location of patients. Some of this data was published in PDF format on a weekly basis, but was spread across multiple documents and tables, and did not follow a standard format in each publication. In order to access the raw data, Pane made an agreement with DGVS to supply them with the data model and training in data collection in exchange for granting his team access to the data itself.[26] This arrangement demonstrates how a clear problem definition and understanding of specific datasets that could help address the problem can enable progress even while government open data efforts lag behind standards and expectations.

FUNDING

As noted, the project was partially funded through a research grant from Iniciativa Latinoamericana por los Datos Abiertos (ILDA). Aside from this funding, however, the project has been conducted entirely by the university research team.

OPEN DATA USE

Data on dengue morbidity that feeds into the prototype application was already opened by DGVS. Additional data accessed by the research and development team was also opened as part of the process of developing the data model. Additionally, all source code used to build the predictive tool is open. As described above, however, much of the data was provided to the researchers in a reciprocal arrangement, rather than broadly opened to the public by the government itself.

25 Ibid.
26 Ibid.

IMPACT

The dengue prevention tool exists as a prototype and proof of concept on how open data can be used to inform the fight against dengue in Paraguay. As such, the principal success indicator to date is successful prediction of future outbreaks, with a secondary indicator of adoption of the data model by the intended key user, DGVS.

ACCURATE FORECASTING

The research and development team's preliminary results indicated that the open data-driven model was able to predict dengue outbreaks a week ahead with an accuracy of 94.78 percent.[27] The prototype data model was given to DGVS after the first round of research to enable their uptake of the tool and its continued development. The follow on impacts of providing this type of predictive capacity to the government entity responsible for managing dengue prevention and response remains to be seen. As of early 2017, there is little indication that this new predictive capacity has fundamentally shifted the intervention strategy at DGVS, but with this newly developed and demonstrably accurate tool in their dengue-prevention toolkit, there is significant potential for impact going forward. Any such impact, however, will be largely dependent on DGVS's responsiveness, especially in the form of a commitment to act on insights generated through the tool; readiness for change and commitment to ensuring sustainability for the effort through consistent resource allocation and data provision.

27 Juan Pane, Julio Paciello, Verena Ojeda, Natalia Valdez, "Enabling dengue outbreak predictions based on open data," Open Data Research Symposium Draft Paper, October 5, 2016, https://drive.google.com/file/d/0B4TpC6ecmrM7Q1lpQ0xoNlJnZlU/view.

RISKS

The potential for privacy harms is likely the central risk of the use of open data to predict dengue outbreaks in Paraguay. As is the case with any data-driven efforts focused on public health concerns, the possibility exists for personally identifiable information to made accessible, open information to be mashed up with other accessible datasets to create new privacy concerns and disease history to inform future decisions (e.g., insurance, housing or hiring) in an unacceptable way.

Additionally, countries affected by dengue are tropical and subtropical, often with a substantial economic dependence on tourism. As a result, they stand to see their economies suffer as a result of full disclosure of the true incidence of dengue and other mosquito-borne viruses. Many of the data-driven efforts to fight dengue and mosquito-borne illnesses focus on mapping high-risk areas and encouraging additional vigilance.[28] Although important for minimizing the spread of such diseases, such interventions could lead to a downtick in tourism and greater reluctance to inform this type of openness from government.[29]

Finally, the initiative is being driven by a small team and championed by a single individual. While this structure helped enable agility in the project development, the project's large dependence on one individual introduces risks to its longer-term sustainability.

LESSONS LEARNED

Several important lessons with wider applicability emerge from this particular case study. These can broadly be categorized by considering the key enablers of the project, as well as the most important barriers or challenges to its success.

ENABLERS

LEVERAGING EXISTING RELATIONSHIPS

The research team behind the effort found success not only thanks to data science capabilities, but also the ability of Pane to leverage contacts from his various professional roles as a researcher and transparency consultant to the Paraguayan government to advance the project. For example, Pane's ability to broker an agreement with DGVS to access their unpublished data, despite their initial concerns about the privacy status of the data, was critical to the tool's launch; he was only able to reach such an agreement because of his preexisting relationship of trust. Dedicated data champions outside government (the demand side of open data) can play a central role, especially if they are able to leverage pre-existing relationships, networks and associations within government.

28 Andrew Young, David Sangokoya and Stefaan Verhulst, "Singapore's Dengue Cluster Map: Open data for public health," GovLab, http://odimpact.org/case-singapores-dengue-cluster-map.html.
29 GovLab interview with Juan Pane, September 9, 2016.

CLEAR PROBLEM DEFINITION AND UNDERSTANDING OF DATA NEEDS

As described above, important data that feeds into the prototype dengue prediction tool was only made available to the researchers as a result of a reciprocal data-sharing arrangement. While this arrangement would likely not be possible were it not for the existing relationships just discussed, the clear problem definition and granular understanding of the specific datasets that could be brought to bear to help solve the problem also played a key enabling role. Rather than being driven exclusively by the data already available, the university research team developed a clear understanding of the objective of their data use (i.e., a longitudinal understanding of incidences of dengue in Paraguay toward the development of a predictive tool for DGVS), which led to a clear understanding of which datasets needed to be accessed and the development of a strategy to loosen the government's grip on them.

BARRIERS

RELUCTANCE TO SHARE

Pane identifies an unwillingness to share data—manifested both as data hugging and exaggerated fears about personal privacy violations—as the single greatest barrier to the project's success.[30] Before he built his tool, the data published by DGVS was in static rather than machine readable format, and was of limited usability for automatic data processing.[31] Better, more complete and more usable data existed, but was being withheld. "The biggest issue is not the technology: it's convincing people to do transparency based on open data," says Pane. [32] Pane also adds that the World Health Organization and Pan American Health Organization could play a more pro-active role, arguing that they too sometimes withhold or otherwise restrict the free flow of data.[33] "We need good data," he says. "The more people publish the data, the better we all collectively will be."[34]

OTHER MOSQUITO-BORNE PRIORITIES

The dengue data model benefitted in part from growing awareness of and concern about not just dengue, but a host of related mosquito-borne illnesses, such as Zika and Chikungunya. On the other hand, the rapid emergence of these multiple illnesses, with often overlapping symptoms, has also created challenges for the team. For example, Pane reports that DGVS is currently withholding data updates while it struggles to come to terms with the impact of Zika on its dengue data. He adds that the new viruses make identifying and modeling dengue much more complex, in large part because the symptoms being reported that previously indicated probable cases of dengue are the same as those for Zika and Chikungunya.[35]

30 GovLab interview with Juan Pane, September 9, 2016.
31 Juan Pane, Julio Paciello, Verena Ojeda, Natalia Valdez, "Enabling dengue outbreak predictions based on open data," Open Data Research Symposium Draft Paper, October 5, 2016, https://drive.google.com/file/d/0B4TpC6ecmrM7Q1lpQ0xoNlJnZlU/view.
32 GovLab interview with Juan Pane, September 9, 2016.
33 Ibid.
34 Ibid.
35 Ibid.

LOOKING FORWARD

CURRENT STATUS

In 2016, Pane's team released preliminary results and a prototype open source web application that makes use of their data model as proof of concept. In collaboration with another group of researchers, Pane is currently modifying the existing model to enable it to predict the number of dengue cases. The current model merely predicts whether there will be an outbreak or not, but Pane is dissatisfied with the subjective nature of the prediction, since there is no accepted definition of what constitutes an outbreak other than disease incidence beyond what would normally be expected.[36]

Pane notes that the rules of engagement have changed dramatically since the emergence of two new mosquito-borne viruses, Zika and Chikungunya. "The world changed. We don't have just dengue now," he says. "Here we have two more diseases that we don't understand."[37] For example, he says that in the past, if a region had 10 confirmed and 40 probable cases of dengue, it was reasonable to assume that the probable cases were also dengue. That assumption can no longer safely be made. Pane and his team are now trying to determine whether to continue to attempt to model dengue, or to attempt to model the suite of symptoms common to all three viruses.[38]

At the same time, Pane acknowledges that the crisis of Zika may catalyze change, forcing the Paraguayan government and other affected countries to embrace greater openness in order to contend with the threat the disease poses. "We should use this momentum to boost the conversation about openness, " he says.[39]

SUSTAINABILITY

The project's results are preliminary, but the fact that an apparently successful open source model has already been developed suggests that it is sustainable. Future use would depend on the development of an immediately replicable open source model.

Pane identifies a number of potential risks to the project's longevity. Like other open data projects, the Paraguay data model is being driven by the passion and conviction of a single individual, and could therefore fall victim to changes in his time and circumstances. Pane also acknowledges the possibility that his model could fail to attract international attention, languishing in obscurity while other researchers attempt to produce similar models. In an attempt to prevent this, he has spoken about the project at several international open data conferences, and all the source codes are open, so that other researchers can benefit from the work already done.[40]

36 Ibid.
37 Ibid.
38 Ibid.
39 Ibid.
40 Ibid.

REPLICABILITY

Although it is not yet ready for immediate adoption elsewhere, Pane's intent is to produce an open source model that can be readily adapted for use in other countries and with other diseases. Within Paraguay, he hopes to extend its use beyond dengue to include other mosquito-borne viruses such as Zika and Chikungunya.[41]

Potential barriers to replicability outside Paraguay foreseen by Pane include national data privacy legislation; varying definitions of dengue infection; lack of technical infrastructure and national data collection and management; and political reluctance to jeopardize tourism revenue by exposing the true incidence of dengue. [42]

CONCLUSION

While it remains a work in progress, Pane and his team have demonstrated that it is possible to use open health data to build a highly accurate early warning system for dengue. Although its continuance has been cast into doubt by the confounding variables of Zika and Chikungunya, Pane remains optimistic that these challenges can be overcome, and that his predictive model could be useful both within Paraguay and abroad.[43]

Pane has sometimes been exasperated by the reluctance of Paraguayan authorities to share data with his team of researchers. He emphasizes the need for governments to consider the usefulness of the data they publish—and withhold: "If there's a message I could send to disease authorities around the world, it is that you are not on your own. There are people around who are smart, who could help you understand what is going on. But for that to happen, you need to publish your data in a way that is actually useful for researchers."[44]

41 Ibid.
42 Juan Pane, Julio Paciello, Verena Ojeda, Natalia Valdez, "Enabling dengue outbreak predictions based on open data," Open Data Research Symposium Draft Paper, October 5, 2016, https://drive.google.com/file/d/0B4TpC6ecmrM7Q1lpQ0xoNlJnZlU/view.
43 GovLab interview with Juan Pane, September 9, 2016.
44 Ibid.