Documents without Borders

Civil Society and the Open Data Movement

Jim Church

The GODORT International Documents Task Force held a pre-conference at the 2013 ALA Annual Conference in Chicago titled "International Statistics: Helping Library Users Understand the Global Community,"[1] with which I was only marginally involved. But I was asked by the committee if it was worth presenting on Nongovernmental Organization (NGO) data, to which I replied it was not: most civil society organizations were not yet in the business of repurposing publicly available data or publishing their own.

Much has changed since then, to say the least. I now have an entire website devoted to NGO data, and numeric, geospatial and textual data have become the coin of the realm in the social sciences. Civil society organizations routinely publish their own research data, while others ingest and repurpose publicly available data to make it more accessible to users or hold national and international organizations to account. Civil society activists are also largely responsible for the open data movement, and the web is now populated by growing numbers of open data catalogs and API lists. This is a tremendously significant development, and the topic of this column.

NGOs and OGD

There are several definitions of NGOs so some clarification of the term may be needed. An NGO is simply an organization not established by governmental agreement, including volunteer organizations, grassroots organizations, and transnational social movements. NGOs are "durable, bounded, voluntary, relationships among individuals to produce a particular product."[2] An NGO can be anything from a local gardening club to Green Peace International: it does not need to be a registered nonprofit. For the purposes of this article I concentrate on social change organizations with research agendas. While it is true that some organizations (such as the Pew Research Center) have published research data for years, the willingness to make such data open has increased as the demand for data has grown and the open data movement has gained traction.

NGOs are also closely tied to Open Government Data (OGD). Safarov, Meijir, and Grimmelikhuijsen provide an exhaustive analytical review of the OGD literature (many articles appear in *Government Information Quarterly*),[3] and the number of OGD articles shows a steadily upward trend: from four in 2011 to 39 in 2015.[4] Uses of OGD are categorized by the authors into five broad themes: transparency and accountability, economic development, citizen participation, public services development, social value, and citizen trust. The chief uses include innovation, data analytics, decision making, anticorruption, smart cities, new services, research, and hackathons.[5] The categories presented below are examples the author has encountered via research consultations and personal investigation. They hopefully shed some light on the subject domains where civil society organizations have been particularly successful working with research data.

Weapons, Crime, and Corruption

Governments routinely publish statistics on economic growth, employment, demographics, education, consumer prices, and so forth. But there are some figures governments are not so keen to reveal, and even if the data is available it may be difficult to find in official sources. Military spending and arms sales are certainly among these: good luck finding which country sold what weapons to whom and for how much over time. Fortunately, the Stockholm International Peace Research Institute (SIPRI) provides easy access to data on arms transfers, military expenditures, and arms-producing and military services companies. The original data is typically found in government reports, budgets, white papers, newspapers, and commercial military periodicals such as *Jane's Defense Weekly.*[6] While these sources are available to the public, the data is much easier to access on SIPRI. The organization still publishes a *Yearbook of World Armaments and Disarmament,* and for a while restricted access to its data via subscriber user id and password. Like many data providers they have since changed this policy, and now the data is free online.

The *Offshore Leaks Database* is an example of a different kind: this contains information on the offshore entities, companies, and individuals revealed in the infamous Panama Papers, Offshore Leaks, and Bahamas Leaks files. The database has information on more than 500,000 offshore corporations and trusts, as well as the names of over 370,000 people and companies linked to these entities, from more than two hundred countries and territories. The Panama Papers were originally leaked to the German Newspaper *Süddeutsche Zeitung*, but due to the size of the data (2.6 terabytes in 11.5 million documents) the newspaper sought help from the International Consortium of Investigative Journalists (ICIJ), a nonprofit originally which focuses on cross-border crime, offshore secrecy, corruption, and the misuse of power. The ICIJ worked for months reverse

engineering the data to find relations between unlinked tables for thousands of companies and persons. The resulting public database is searchable and allows users to download data by originating country (people, companies and addresses with ties to offshore entities) and the jurisdiction of the offshore corporations and trusts. It was published under an Open Database License.

Another example with widespread academic use is Transparency International (TI) a Berlin-based NGO working to combat corruption. I first presented on this source at a conference years ago and showed a visualization of the group's *Corruption Perceptions Index,* which color-codes nations worldwide according to public perceptions of corruption. This, again, is not the type of thing governments typically investigate: how many civil servants would survey citizens about how corrupt people think the country is? Transparency International also publishes tabular data and visualizations in its *Bribe Payer's Index*, which captures perceptions about the likelihood of companies from large economies to pay bribes abroad. It also publishes research reports which can be searched and filtered by country and topic, on subjects related to transparency and corruption, from information access to youth and sport.

Official and Unofficial Aid Data

Few economic topics have generated as much controversy as foreign aid. For years the main sources for international aid data or "Official Development Assistance" (ODA) were the annual *Overseas Loans and Grants* or "Green Book" and the OECD's *Geographical Distribution of Financial Flows to Developing Countries*. The Green Book is a freely available US government publication while versions of OECD development data have also been free for years. But this data is primarily from OECD Development Assistance Committee (DAC) countries, so data from India, China, Brazil and other economies is unavailable on the platform. In the past this may have been justified, but with emerging economies now comprising significant shares of the aid sector, this data can no longer be ignored.

I well remember encountering the AID Data site when it was first released—it was a revelation. AID Data was originally a partnership between the College of William & Mary, the Development Gateway, and Brigham Young University. Prior to its release most development data was aggregated at the national level and published annually: with Aid Data the user can access funding at the project level with the selected data sets geocoded. Aid Data also reposts academic journal articles accompanied by replication aid datasets that address and at times debunk the conventional wisdom about development aid and its consequences. On the main site dashboard a search for India reveals 422 development projects funded from 2006 to 2010, worth $4.1 billion: project details include the name of the funding organization, the sector, the amount, and title of the project. But the data also includes other financial flows besides projects, such as Foreign Direct Investment and Remittances: for India these totaled $167.8 billion for India from 1976 to 2012. The addition of this data provides another perspective on the nature of development assistance. There is also a new Aid Data spinoff called China.aiddata.org—a platform for Chinese development finance to Africa. For years access to this has been problematic so this is a very welcome development. There are also new datasets not incorporated into the main dashboard, including one on Brazil's South-South Cooperation, Aid Locations During Civil Wars South of the Sahara, and historical data web scraped from the US Agency for International Development (USAID). All told the data on this site is incredibly rich.

Aid data visualization is also a big deal. One article about this posted by the UN Refugee Agency is entitled "The road to hell is paved with brightly coloured bubble maps."[7] This is only funny if you have been to such sites and spent hours clicking only to discover positively nothing of any import: the author speculates that some data sites are just show pieces that agencies point to when asked about open data policies. But *d-portal* (http://d-portal.org/) is not one of these. The site is a user-friendly visualization tool using data from the International Aid Transparency Initiative (IATI)—an undertaking working to improve the availability of development and humanitarian data across multiple sectors. The IATI Standard is an international framework for publishing data used by governments, the private sector and national and international NGOs. Unfortunately, the IATI website is not easy to use, which is where the *d-portal* comes in. The user can select a donor, recipient, time frame, sector, and publisher, and the site returns a list of current and completed projects, some of which are sponsored by private charities. Browsing the donors and publishers displays major aid agencies such as the United Nations Development Programme, as well as smaller ones like "Lively Minds" and "Send a Cow Uganda." When downloading the d-portal also offers the option of taking the user the original IATI registry file, where other options are presented.

Fake Fish

If you ever wondered if some of the data governments report was potentially misleading you were probably right: many governments overcount the good and undercount the bad. An interesting and sobering example of this is fishery statistics. The major source of world fishery data is the Food and Agriculture Organization of the United Nations which reports, among other things,

on national reported fish catches. But the official data is most likely a significant undercount of the actual fish caught due to unreported landings and discards from commercial fishing vessels. The Sea Around Us, a nonprofit based in British Columbia (whose name derived was from Rachel Carson's bestselling book) has data models that illustrate these discrepancies, suggesting that from 1950 to 2010, global fish catches were almost 50 percent higher than reported.[8] In 2012 the nonprofit Oceana collected seafood samples from hundreds of retail outlets nationwide to determine if they were accurately labeled. They found that one-third of the samples were bogus: particularly popular fish such as red snapper and tuna, which were mislabeled 87 and 59 percent of the time, respectively.[9] But the news is not all about fakery. Other environmental research organizations such as the Global Footprint Network repurpose data from international organizations such as the Food and Agriculture Organization of the United Nations, the UN Statistics Division, the International Energy Agency, and academic sources, to produce a National Footprint Account: a measure combining thousands of data points per country over time to calculate the ecological resource use and capacity of nations.

Liberating the 990

In 2010, there were 1.5 million tax-exempt organizations in the United States with $1.51 trillion in revenues,[10] comprising about 9.2 percent of US wages and salaries. The IRS gathers information from these organizations on Form 990, which includes financial information on boards, investments, and other factors, depending on type of the form submitted. Several institutions, such as GuideStar and the National Center for Charitable Statistics (NCCS) at the Urban Institute historically acquired this data for a fee (PDFs on DVDs) and converted it into machine readable formats, which they in turn sold to libraries and other users via a subscription database. But for years this was certainly not open data: at best users could freely download PDFs a few at a time.

After many years of activism and a lawsuit filed by Carl Malamud, in June 2016 the IRS released this data in ASCII, JSON, and XML formats onto Amazon cloud servers.[11] Since then things have been happening fast. A Github repository, "Open Data for Non Profit Research," was released by researchers at Syracuse University,[12] and the files uploaded into a *Nonprofit Initiative for Open Data* dataverse; Charity Navigator created a toolkit that allows users to clone the IRS dataset as a relational database; and the National Center for Charitable Statistics at the Urban Institute opened their historic IRS data files onto a *National Center For Charitable Statistics Data Archive.*[13] The files are in CSV format and accompanied with data dictionaries and a helpful user guide. While great kudos is due to data activists like Karl Malmud, the Aspen Institute, and Jesse Levy and Nathan Grasse at Syracuse University, these formats present challenges to novice users. Recognizing that preparing and maintaining open data in an easy-to-use formats for citizens has costs, hopefully our best minds will continue to find ways to make civil society data even easier to use and access in the future.

Jim Church (jchurch@library.berekeley.edu) is the International Documents Librarian at the University of California Berkeley.

References

1. "International Statistics: Helping Library Users Understand the Global Community" (preconference, ALA Annual Conference, Chicago, 2013), http://connect.ala.org/node/204097.
2. Leon Gordenker and Thomas Weiss, "Pluralizing Global Governance: Analytical Approaches and Dimensions" in *NGOs, the UN and Global Governance*, edited by Thomas G. Weiss and Leon Gordenker, 17–47 (Boulder, CO: Lynne Rinner 1996), 17–47.
3. Igbal Safarov, Albert Meijer, Stephan Grimmelikhuijsen, "Utilization of Open Government Data: A Systematic Literature Review of Types, Conditions, Effects and Users," *Information Polity: The International Journal of Government & Democracy in the Information Age* 22, no. 1 (2017): 1–24, https://doi.org/10.3233/IP-160012.
4. Ibid., figure 4.
5. Ibid., figure 6, p. 7–8.
6. See sources at "Sources and Methods," Stockholm International Peace Research Institute, accessed October 11, 2017, https://www.sipri.org/databases/armstransfers/sources-and-methods.
7. Zara Rahman, "The Road to Hell is Paved with Brightly Coloured Bubble Maps," UNCHR Innovation Service, June 10, 2015, http://www.unhcr.org/innovation/the-road-to-hell-is-paved-with-brightly-coloured-bubble-maps.
8. D. Zeller et al., "Still Catching Attention: *Sea Around Us* Reconstructed Global Catch Data, Their Spatial Expression and Public Accessibility," *Marine Policy* 70 (August 2016): 14–162.
9. "Oceana Study Reveals Seafood Fraud Nationwide," Oceana Report, February 2013, http://oceana.org/reports/oceana-study-reveals-seafood-fraud-nationwide.

10. Beth Simone Noveck and Daniel L. Goroff, *Information for Impact: Liberating Nonprofit Sector Data* (Washington, DC: Aspen Institute, 2013), http://thegovlab.org/wp-content/uploads/2013/06/psi_Information-for-Impact.pdf.
11. "IRS 990 Filings on AWS," Amazon Web Services, accessed October 11, 2017, https://aws.amazon.com/public-datasets/irs-990/.
12. "Open-Data-for-Nonprofit-Research," GitHub repository, last updated September 5, 2017, https://github.com/lecy/Open-Data-for-Nonprofit-Research.
13. *National Center for Charitable Statistics Data Archive*, Urban Institute, http://nccs-data.urban.org/index.php.