From: Pimenta and Canabarro, 'Digital Governance', 2014 http://www.ufrgs.br/cegov/files/livros/gtdigital.pdf

LESS IS MORE:
THE ROLE OF SMALL DATA FOR GOVERNANCE IN THE 21ST CENTURY

CATHERINE D'IGNAZIO

Artist, software developer and educator, she is the Director of the Institute for Infinitely Small Things, an interventionist performance troupe, and former Director of the Experimental Geography Research Cluster at RISD's Digital+Media MFA program. Catherine has a BA in International Relations from Tufts University (Summa Cum Laude, Phi Beta Kappa) and an MFA in Studio Art from Maine College of Art Catherine taught in the Comparative Media Studies program at MIT in 2009 and 2012.

JEFFREY WARREN

Fellow at MIT's Center for Civic Media, on the board of the Open Source Hardware Association, on the advisory board of Personal Democracy Media's WeGov. Jeff holds an MS from MIT and a BA in Architecture from Yale University. Creator of GrassrootsMapping.org and co-founder and Research Director for the Public Laboratory for Open Technology and Science.

DON BLAIR

PhD candidate at the University of Massachusetts at Amherst. His research involves problems in soft matter, condensed matter, and biophysics, as well as computational complexity theory and information theory. He has demonstrated a commitment to open science by organizing and running a successful workshop on open science hardware, and shepherding a group of technologists interested in solving scientific and environmental problems using open source tools.

INTRODUCTION

Transparency, open data, and data-driven approaches to governance have become popular in part due to the promise of closer engagement between government and the public. This trend has emerged in parallel to the use of Big Data in government – the aggregation and analysis of vast amounts of data *about* the public, in the hope that it may yield key insights about our society and provide the basis for better decision-making. Many of these concepts grew out of the desire for a more discursive mode of democracy, where information, shared openly, helps to bolster decisionmaking processes while also promoting accountability.

Unfortunately, the adoption of ideas of transparency and openness have been decidedly asymmetric, especially in environmental science – with a purely inward flow of data towards a central authority (experts or scientists) in whom we must trust to make decisions on our behalf. Likewise, the open government movement has been sidetracked by logistical questions of standard data formats and a focus on *visualization* rather than *participation*. Without a more participatory model, where members of the public may participate in collecting, analyzing, and interpreting data about issues important to them, we are left with a system which gives "data shepherds" – scientists, technologists, or analysts – sole authority in reading these "data tea leaves". Open data, provided on a voluntary basis by government or corporations, is typically self-reported (examples in WV chemical spill, others), making it a poor mechanism for accountability.

We propose that a *bottom-up, participatory, grassroots* approach to environmental investigation and data collection addresses the key issues of *inclusion*, *accountability*, and *credibility*, by building public participation into the data lifecycle. We envisage forms of participation in which members of the public take part in creating, analyzing, and understanding datasets, and using them to advocate for change. In the following FAQ essay, we refer to this as a "Small Data" approach, and examine the implications of this approach in a series of questions and proposed answers.

WHAT IS SMALL DATA?

When you hear "we can use Big Data to help us understand X," it is the definition of "us" in such statements that distinguishes Big Data from the Small Data approach. Both Big Data and Small Data use a data-driven approach to create understanding; and both may involve the aggregation of large data sets, contributed from a variety of sources. But Small Data is a practice owned and directed by those who are contributing the data – for example, a rural community collecting data on air pollution from frac sand mining, or a group of concerned residents investigating a nearby chemical spill. The essence of Small Data is that such communities may not just participate in, but can actually initiate and drive such data investigations towards the better understanding of an important local issue.

ISN'T IT ENOUGH THAT MY ORGANIZATION ALREADY MAKES DATA PUBLICLY AVAILABLE ONLINE?

There are many agencies and organizations that have begun, in the name of transparency, to embrace an 'open data' *ethos*. In many cases, this means that some of the organization's data, deemed to be of potential relevance relevant to the public, can be accessed online, in a web browser. But for such 'open data' to empower the public to make informed decisions, vote wisely, or wisely engage in collective action, more must be achieved: the data must be rendered legible, and meaningful, for the various public audiences.

CAN THE PUBLIC COMPREHEND SCIENTIFIC RESULTS WITHOUT SPECIAL TRAINING?

The usual approach taken by organizations and governments in attempting to render the data they have collected meaningful and relevant to the public is to provide summary digests of data for public in terms of simplified visualizations and infographics that depict overall trends and summary conclusions. While this practice has led to important insights into data, and has increased legibility, rendering data truly legible often requires dialogue with communities about what questions and modes of communication are most meaningful to them. Further, questions and communication styles will change over time; this calls for an ongoing, rich dialogue between organizations and the public.

However, we believe that there exists a more direct, and more effective path towards achieving the conveyance of truly meaningful and relevant information to the public: the facilitation of true, grassroots public *participation* in the entire data lifecycle. One of the best ways to ensure legibility and relevance, we suggest, is for communities to pose, frame, and find ways of generating answers to the questions themselves.

Consider a recent crowdsourced water quality monitoring initiative in China:

> This January, a few hundred employees of Alibaba, the massive online retailer and digital payments company, participated in an interesting experiment. Like many Chinese, they traveled home to celebrate the Lunar New Year. While at home, they used inexpensive water testing kits to sample water in their villages and uploaded their findings via smartphone to an environmental mapping website, Danger Maps. Employees measured water quality in 420 locations across 28 provinces, testing open bodies of water as well as sources of drinking water (ZUCKERMAN, 2013)[1].

Reflecting on this initiative, as well as the work of groups like Public Lab[2] and Safecast[3] – organizations which have employed similar crowdsourcing and citizen science initiatives – Zuckerman writes:

> Their work raises questions of whether we want citizens to be cooperative sensors, or citizen "scientists". The latter is a high bar to cross – we need citizens not only to collect data but to formulate and test hypotheses. What we gain in exposing participants to the scientific process, we may lose in terms of data quality and believability .... There's a balance between accessible sensors, high-quality data and the ability for users to formulate and test hypotheses that crowdsensing projects need to wrestle with going forward (ZUCKERMAN, 2013).

Zuckerman nicely highlights the most important questions raised by a 'grassroots' approach to scientific investigation, and the apparent trade offs between full citizen participation and accessibility, on the one hand, and scientific credibility or 'believability', on the other. This leads us to tease out and attempt to address these important questions about the relationship between 'citizen science' and 'science'.

IS A BOTTOM-UP, GRASSROOTS SMALL DATA APPROACH COMPATIBLE WITH 'REAL SCIENCE'?

We believe that a truer, expansive notion of 'science' is one based not merely to professional credentials, academic institutions, governmental accreditation, or cultural prestige, but rather on the judicious application of the scientific method and scientific reasoning – by individuals and communities – in an attempt to

(1) <http://www.ethanzuckerman.com/blog/2014/04/20/water-monitoring-in-china-and--the-changing-role-of-citizenship/>. Last access: Oct. 19, 2014.

(2) <http://publiclab.org/>.

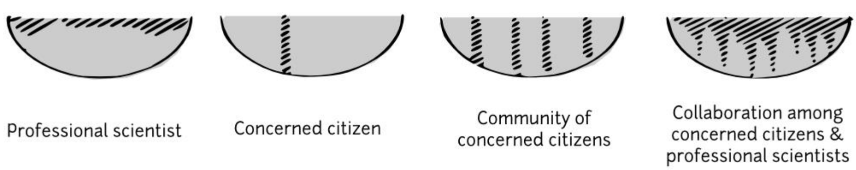
(3) <http://blog.safecast.org/>.

revise and improve our understanding of the world. This 'our' points to the fact that the scientific process is, inherently, one of dialogue, a process that consists in an investigator (professional researcher, or citizen scientist) attempting to convince others (professional colleagues; other citizens in a community; governmental agencies; posterity) that it is proper to revise their beliefs in such a way, by making reference to mutually agreed-upon standards of evidence. It is this more general understanding of 'science' that informs our answer to the above question – an emphatic *yes!* – and leads us to assert:

Small Data complements Big Data and professionalized science. The Small Data approach to data collection and scientific investigation does not dismiss traditional approaches to environmental data collection. In many cases, the expense and difficulty of more precise, lab-based testing can be augmented by a broader, more participatory approach. For example, air quality monitoring by hundreds or thousands of low-precision, but affordable, home monitoring devices can extend the reach of the few-and-far-between tests performed with more precise instruments. Furthermore, we propose that investigations that leverage local knowledge through close collaborations between the public and professional environmental scientists throughout the data lifecycle can result in better science and better outcomes for everyone.

A 'Do It Together' approach. Most scientific problems, especially those involving the environment, are best addressed through an approach that combines broad (and shallow) syntheses of experience ('energy is conserved in a closed system; plants require nutrients to survive') with deep, specific, local knowledge ('my drinking water has started to smell like licorice, recently'; 'I'm coughing a lot recently, and am getting migraines whenever I use the shower')[4].

Fig. 1 - Research scientists, concerned citizens, and grassroots community organizations provide complementary forms of knowledge when addressing scientific problems together[5].



(4) See Osnos (2014).

(5) Adapted from Vandermeer (2014).

Following John Vandermeer's view exposed in a 2014 blog post , we might represent a given environmental concern – suspected contaminants in water sourced from a well located, say, in rural Peru – metaphorically, as in Figure 1, above: all relevant knowledge required to address the problem might be considered as a 'lake'. A typical research scientist (in this case, a hydrologist) might possess many broadly-applicable principles (connecting insights on the 'surface' of the 'lake of knowledge') which are vital to solving the problem at hand. But even after a lifetime of accruing knowledge through academic training and professional practice, such researcher will likely be unable to match the deep, local knowledge possessed by a homeowner with a lifetime of direct experience of the taste, smell, and color of their own drinking water (represented by a deep, narrow band of knowledge in the 'lake'). A grassroots community of such homeowners can then be considered to have, among them, an important and unique collection of such 'deep, direct' experiences. In some cases, a community is able to use such collective, local knowledge to address the problem without any additional expertise ('we should stop drinking water from the well near the farm, because we all agree that it has recently been tasting very bad'); and for cases that benefit from a broader perspective, adding in the broader (but necessarily more shallow, without local, lived experience) knowledge and techniques of professional scientists to this local, deep knowledge makes for a powerful, collaborative approach to problem solving.

BUT WHAT ABOUT DATA VALIDITY?

Addressing important scientific questions requires acquiring good data. Data quality is, arguably, a nebulous concept, and whether data is acceptable as evidence in support of a hypothesis depends quite obviously on the explicit or implicit standards of evidence employed by the intended audience for a given investigation. Some aspects of data quality include: whether the source of the data is deemed trustworthy (data provenance); whether the data acquisition methods were performed in accordance with established procedure; instrument validation; and reproducibility. Various institutions and agencies have evolved an array of mechanisms for evaluating the quality of data, including training, certification, and peer review. Is it possible for such institutions to admit data submitted by non-experts, without special training, as evidence on par with their own?

Chains of trustworthiness. While the diverse array of backgrounds, equipment, and methods that might be employed by non-experts in Citizen Science/Small Data investigations certainly presents a challenge, techniques exist for 'bridging' data sets from such disparate sources. As a simple example: 'low-veracity' scientific instruments created or improvised by citizen scientists might initially be 'validated' against 'professional, trusted' instrumentation, before deployment; further periodic checks using this method can ensure that calibration is maintained subsequently.

Small Data as a 'first pass', 'early warning system', and 'conversation starter'. Another approach is to rely on inexpensive, 'low veracity' instrumentation to provide an initial, first-pass assessment of the places wherein more expensive, 'high veracity' instrumentation, ought to be deployed. This 'prosthetic' approach has enormous potential for extending the reach and capacity of monitoring agencies (see, for example, the Alibaba water quality monitoring project, referenced above). Here, we should also underline the important social function involved in organizing a community to collect and analyze a data set. While the data collected itself may not hold up to agency standards, the process of participatory collection serves a discursive function to inject these topics into public space and public action, engage new participants and spur more investment into interrogating the questions at hand.

Small Data is often sufficient to answer fully the question at hand. In many cases, meaningful answers to important questions do not require highly-accurate techniques or sensors. While a question like 'What is the precise level of contaminant X in my water?' might require expensive, high-veracity instrumentation and expertise to answer, a question like: 'Does there seem to be an unusually high, worrisome level of X in my water, so that I might reconsider drinking water from that source?' might require only a very simple, binary sensor reading, for which purpose a low-cost, low-veracity instrument is quite sufficient.

Indeed, sometimes even what might be considered 'expert' questions, posed by members of the scientific establishment, can be fully and directly addressed with 'low veracity' techniques. For example, hydrology researchers interested in the impact of urban road salt practices on the nearby river ecosystem require only an assessment of the extent to which an observed pulse of rainwater through the river network after a storm is correlated with an associated pulse in relative salinity; both the river depth measurement and the conductivity measurement, so long as they are consistent, can be relative, rather than absolutely accurate. They need not be particularly precise so long as the measurements are sufficient to distinguish relative increases and decreases in conductivity and depth over time.

For example, the CATTFish Project[6] at Carnegie Mellon University, in which conductivity probes were deployed in toilet tanks, concluded that relative spikes in conductivity above a pre-measured baseline was likely an indication that the house water was being contaminated by nearby fracking operations, so that

(6) <http://www.cmucreatelab.org/projects/Water_Quality_Monitoring/pages/CATTFish>.

residents ought not to use the water for bathing and drinking when conductivity readings were high. The demonstrated improvements in health outcomes as a result of this intervention were achieved using low-cost, accessible equipment; further, no governmental agency or institution was directly involved.

The professionalization of the practice of 'Science', and high school educational practices focused on training students to follow procedures, reproduce results, and score well on standardized exams has led lead most non-experts to feel that they are not 'allowed' to conduct investigations on their own. We believe that by facilitating grassroots participation in the entire scientific process, from hypothesis generation to data analysis, important new societal capacities for answering difficult, systemic problems will emerge, allowing novel types of scientific questions to be asked and addressed.

WHAT DOES SMALL DATA LOOK LIKE IN PRACTICE?

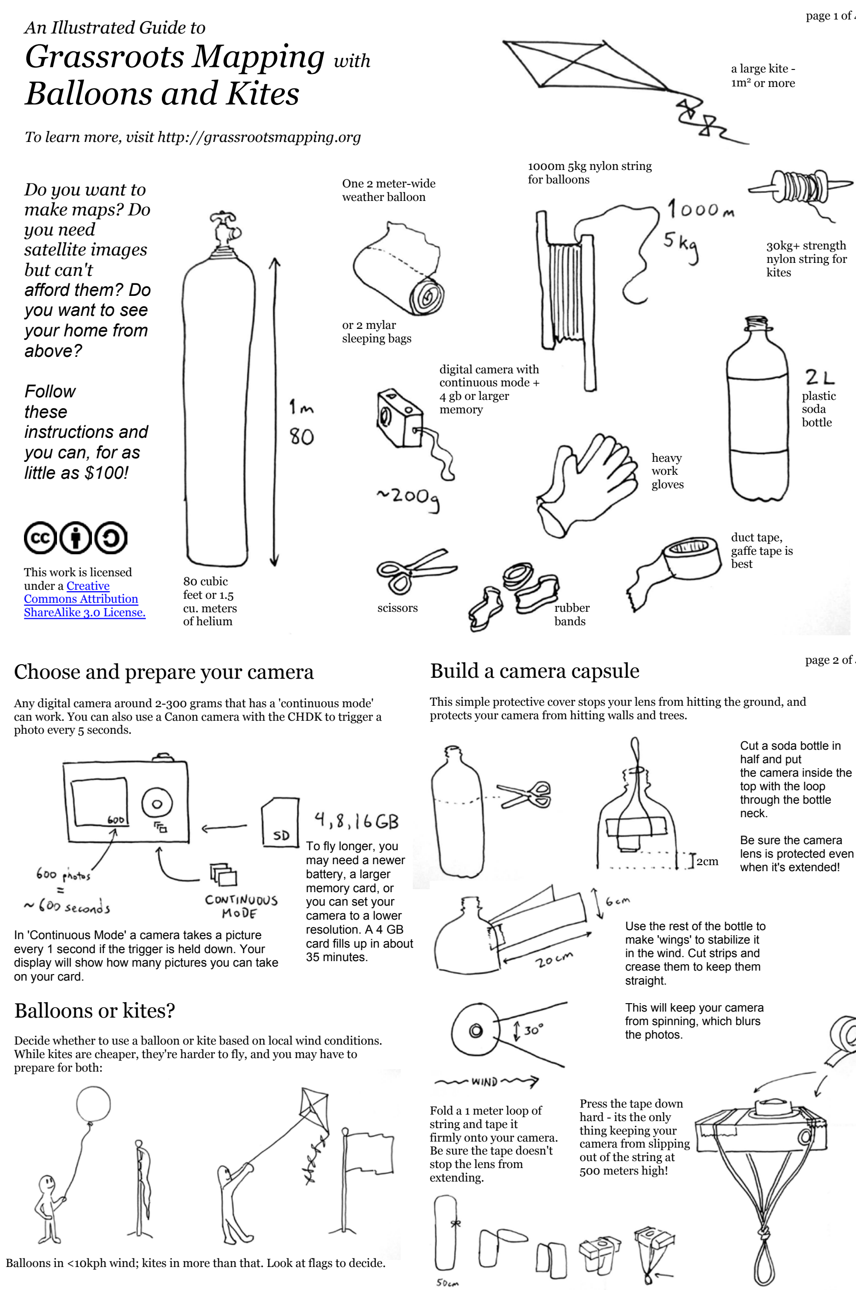
One early Public Lab project was a grassroots project to document the effects of the BP oil spill at a time when journalists were being kept away from spill areas, and a no-fly-zone prohibited close aerial photography of the spill-affected Barataria Bay. Local "civic scientists" used balloons and kites to lift cameras up over a thousand feet in the air, documenting some of the worst-hit wetlands and public beaches along the Gulf Coast. The photographs and maps were republished widely in the press, and maps made before, during, and after the spill helped both wetlands researchers and the public to better understand the scope and severity of the disaster. Later, *Google* chose to publish the locally produced oil spill maps on its *Google Maps* platform. With the high-resolution imagery they collected, hundreds of local residents were able to shape the public's understanding of the BP oil disaster through the use of affordable open source techniques.

Fig. 2A - A Public Lab map of Lake Borgne, Louisiana, made using open source photo stitching and rectification software MapKnitter.org and Public Lab's balloon mapping[7] techniques. Photos (green, colored strip) were taken via balloon by Erin Sharkey, and stitched together in MapKnitter (using underlying *Google Maps* satellite imagery as a base layer, for rectification) by Stewart Long.

Rapid progress. Because of their community-based, discursive nature, Public Labs' Small Data projects rely heavily on online forums, wikis, blogs, and other similar modes of communication that have proved useful in open source software communities. 'Peer review' is accomplished through comment threads on blog posts, through 'likes', and similar forms of recognition. But rather than focusing on a one-time assessment of the merits of an academic paper, the Small Data research cycle begins with ideas posted on mailing lists, proceeds to prototype designs described in blog posts, and easily leverages rapid-fire constructive feedback, non-competitive collaboration, and community support that are encouraged by an open source *ethos*. As in other realms of open source collaboration, innovation (and the correction of mistakes) proceeds at a rapid pace in such communities (SCHWEIK; ENGLISH, 2012).

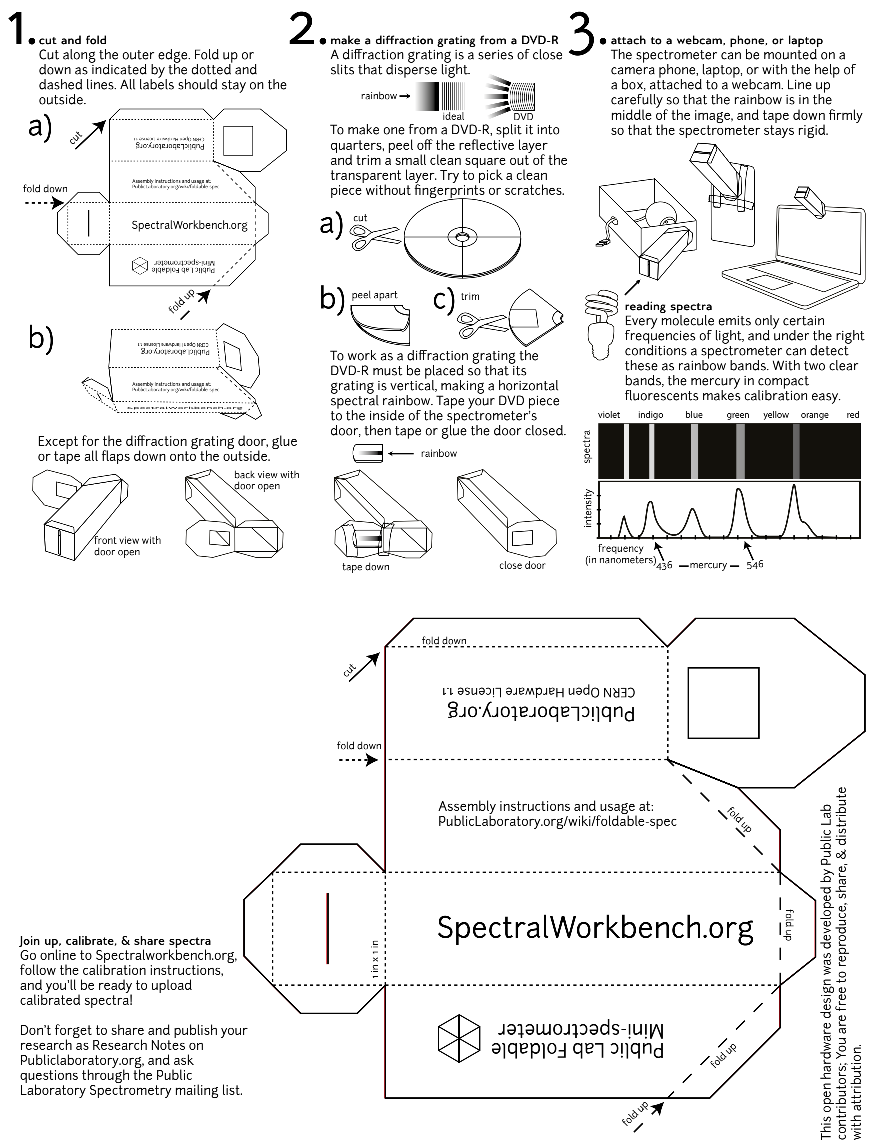
(7) <http://publiclab.org/wiki/balloon-mapping>.

Fig. 2B - A Public Lab Balloon and Kite Mapping guide[8], useful for enabling citizens to acquire their own aerial imagery, using inexpensive cameras, of the sort depicted in Fig 2A.



(8) <http://publiclab.org/wiki/revisions/balloon-mapping>.

Fig. 3 - Public Lab's "Foldable Spectrometer"[9] – developed through a community dialogue among citizens concerned by the Gulf Oil spill's effects on their beaches, research chemists, high school teachers, and grassroots activists – requiring only a cardboard cut-out, a discarded DVD, and a smartphone or other digital camera. The spectral imagery captured by this device, used conjunction with a $5 laser, can identify crude oil contamination in a sample.



(9) <http://publiclab.org/wiki/foldable-spec>.

It was this sort of collaborative, online community – including collaborations among high school teachers, chemists, and grassroots activists – that collectively developed Public Lab's "foldable spectrometer" (Fig. 3) device, consisting of a piece of cardboard, a segment of a throw-away DVD, and a smartphone or webcam – a device designed to allow citizens to determine, in the wake of an oil spill, not only whether crude oil is present in a sample, but also potentially identify the unique spectral signature that ties that oil to the corporation responsible for producing it.

WHAT ABOUT PRIVACY AND DATA OWNERSHIP?

In the 'aggregate and analyze' Big Data approach, information flows from individual actors, who collect and provide (wittingly or not) data quanta, into large, centralized databases for subsequent processing. Typically, these aggregate datasets are not made available to the public, as their contents often play a key role in Big Data business models. Thus, the individual actors who provided the data in the first place are often not granted access to aggregate data sets; even their own data may be difficult or impossible to access (e.g. witness the complexity associated with ordering one's own 'credit report').

In a Small Data project, individuals and communities attempt to retain control over their own data. Individual contributors are reminded frequently (as on the Publiclab.org main site) of their 'licensing options' when contributing data to a central database, and efforts are made to ensure that contributors are recognized, and can easily manage the associated data.

In some cases, a balance must be struck between a community's interests ('Is there a systematic and ongoing contamination of the private well water in our region?') with an individual's interest ('If the high levels of arsenic in my private well were made public, my property values would plummet'). A Small Data *ethos* recognizes and respects such issues, and attempts to provide an array of potential solutions (anonymizing data contributions, or providing secure, accessible mechanisms for ensuring data privacy).

CAN MY ORGANIZATION INVEST IN BOTH SMALL DATA AND BIG DATA?

The Big Data approach, when employed wisely, represents an important new set of analytical tools that allows organizations and institutions to generate new insights into systemic problems. An organization wishing to use these tools will need to invest in the cultivation of sophisticated mathematical and statistical expertise, as well as in the acquisition of significant computational resources – i.e. hardware, software, and programming skills.

The Small Data approach, we believe, represents an equally important and promising approach to addressing systemic problems faced by organizations and society at large; however, the types of investments required to enable this grass-roots, participatory approach look fundamentally different from those of Big Data. Instead of a focus on hiring expertise and increasing computational capacity, a Small Data approach is supported through investments in the social structures that empower more of the public to usefully frame, pose, investigate, and discover explanations and solutions for the problems they face: direct investments in schools, community centers, and grassroots science education; and, indirectly, through support of any initiatives that are believed to support a thriving, democratic society.

WHAT PROBLEMS MIGHT WE USEFULLY ADDRESS WITH A SMALL DATA APPROACH?

A grassroots, community-led, Small Data approach can be used to address any scientific concern in which individuals or communities feel invested in the outcome of the investigation. This is often the case in the context of environmental issues – individuals and communities often have a clear stake in having clean air to breathe, clean water to drink, or food without chemicals and pesticides.

A Small Data approach can be self-contained, within a local community; it can bridge several communities; it can be deployed in critique of insufficient state efforts, or abuses of corporate power; or it can be used to complement the environmental, extending the spatial resolution of current monitoring efforts and engaging the public in a deeper understanding of the ecological systems upon which they depend.

Below, we briefly highlight several ongoing efforts that employ a Small Data approach in addressing important environmental problems.

Water quality. Public Lab has recently launched the Open Water Project[10], which aims to make water quality information accessible, easily sharable, and more directly meaningful to communities. Typical water monitoring efforts have relied on expensive, proprietary technologies, severely limiting the scope and accessibility of water quality data. Homeowners interested in testing their own well water, watershed managers concerned about fish migration and health, and

(10) <http://openwaterproject.io/>.

global communities facing toxic contaminants in their water supply might all benefit greatly from an open source, inexpensive, accessible set of technologies and methods in water quality monitoring. Public Lab is working with the larger hydrology and environmental science community in the development of open data standards and accessible educational materials, with the idea of making water quality data easier to collect, understand, and leverage in a dialogic process of civic engagement. Importantly, while the Public Lab community is actively engaged in the development of novel low-cost, open water quality monitoring hardware, it aims to build collaborative bridges with other open source communities (like the CREATE Lab at CMU) that are focused on similar efforts, aiming to help curate, support, and promote such efforts.

Air quality. Public Lab has also recently developed a focus on accessible approaches to monitoring local air quality. By sharing designs, particulate matter sensor hardware 'tear downs', programming techniques, and design ideas in the publiclab.org online wikis[11], research notes, and mailing lists – and by linking to similar resources elsewhere online – a collection of research scientists, journalists, and concerned citizens has made rapid progress towards determining the lowest-cost, accessible and effective technologies[12].

Agriculture and land use. FarmHack[13] is an "open source community for resilient agriculture". As an organization, it has developed an extensive network of farmers and engineers[14] interested in applying an 'open source *ethos*' to the development and sharing of low-cost, DIY farming technology. Many of the designs developed or employed within this extended community have been curated and shared via FarmHack's 'Tools Wiki'[15], with tools that range from bicycle-based root washing systems to biodiesel generators housed in old soda delivery trailers. FarmHack's recent focus has been on identifying points of leverage and mutual interest among climate scientists interested in collecting data about environmental trends and farmers interested in collecting data about their own crops – a set of interests that broadly intersects with water quality, soil health, and air quality. Members of FarmHack have collaborated extensively with Public Lab's community on developing Infragram[16], an open source approach to capturing DIY plant health imagery using inexpensive filters and commercially-available cameras; further collaboration on water quality issues common in an agricultural setting are on the immediate horizon.

(11) <http://www.publiclab.org/wiki/air-quality>.

(12) <http://www.publiclab.org/search/air-quality>.

(13) <http://farmhack.net/home/>.

(14) <http://farmhack.net/shops>.

(15) <http://farmhack.net/tools>.

(16) <http://infragram.org/>.

WHAT ARE THE CHALLENGES FOR THE SMALL DATA APPROACH, GOING FORWARD?

We believe that the application of an open source, grassroots, participatory, bottom-up, and fully collaborative *ethos* – this "Small Data" approach – in the realm of environmental science has enormous potential to both reinvigorate society's notion of how 'science' is done and, at the same time, address some of the most pressing issues we face today.

We do not feel that it is as yet possible to definitely answer the questions posed in this essay. Finding ways of validating data across communities with disparate standards of evidence; achieving productive dialogues between citizen scientists and professional scientists; finding effective ways of negotiating hierarchies in expertise and background within grassroots scientific communities; engaging and educating a diverse public in ways that enable these sorts of grassroots scientific investigations; and establishing useful dialogues among activists, educators, professional scientists, policymakers, and the public; all of these achievements will require new methods, new forms of collaboration, and an effective 'community organizing' effort across the various backgrounds that must be represented and work together. We are, however, deeply optimistic and excited by the progress we have witnessed already in this 'Small Data' realm, and about the significant advancements we anticipate for the future.

REFERENCES

OSNOS, Evan. Chemical valley: the coal industries, the politicians, and the big spill. (The) **New Yorker Magazine**, Apr. 7, 2014. Available at: <http://www.newyorker.com/magazine/2014/04/07/chemical-valley?currentPage=all>. Last access: 20 Oct. 2014.

SCHEICK, Charles M.; ENGLISH, Robert C. **Internet Success**: A Study of Open Source Commons, 2012. Available at: <http://mitpress.mit.edu/books/internet-success>. Las accesss: 20 Oct. 2014.

VANDERMEER, John. The Sixth Developmentalist Myth of Agriculture (From Lewontin and Levins, 'Biology Under the Influence') and Comments on its Resolution. **John Vandermeer's Blog**, March 31, 2014. Available at: <http://johnvandermeer.blogspot.com.br/2014/03/the-sixth-developmentalist-myth-of.html>. Last access: 20 Oct. 2014.