Workshop on Third-Wave Data for Environmental Policy in China

Tuesday, March 1, 2016
9:00 am - 1:00 pm
Yale Beijing Center, 36th Floor, Tower B, IFC Building, 8 Jianguomenwai Avenue

Executive Summary

When it comes to our most pressing environmental challenges, much of the data needed to take effective action are missing. Water quality, recycling rates, toxic chemical exposures, land degradation – assessing these environmental issues is hampered by the lack of consistent, global information pathways. Where data exist, they are often incomplete, erratic or untrustworthy. Even if accurate, data may be based at a resolution or scale inapplicable to the policy question at hand. The flood of information poses new problems of its own, often creating more noise than clear signals.

In China, recent announcements by the government have signaled a new opportunity to scope the potential for novel data approaches, including big data (e.g., satellite data) analysis as well as bottom-up methods (e.g., citizen science) to help address these data challenges, utilizing a mix of inventive and cutting-edge techniques we dub "third-wave data." "Third-wave data" has the potential to engage new actors (e.g., citizens and technologists) and audiences in environmental data collection, improve linkages between policy-making and monitoring, and facilitate collaboration across private and public sectors in China.

On March 1, 2016, government officials, civil society organizations, and researchers gathered at the Yale Beijing Center to share their work assessing and responding to key data gaps in existing environmental information in China, and to scope potential opportunities for big data to measure, manage, and communicate environmental issues. This workshop aimed to build a baseline understanding of the status of data-driven initiatives in China, to identify opportunities for future collaboration and initiatives. Discussions were loosely organized around a gap analysis of data in China; the status of environmental monitoring in China; the current big data initiatives in China; and international cooperation in China's environmental monitoring.

A number of key themes and questions emerged from these presentations and conversations:

1. **The "central-local divide,"** or the difficulty of coordinating information between central, provincial and local governments. However, there are indications that more harmonization and data centralization is happening. The Information Center at Ministry of Environmental Protection is bringing disparate sources of data together in a single platform, which clarifies the existing data landscape, and helps inform efforts to create more legible and accessible monitoring results.

2. **Distinguishing data from usable data.** Difficulties in accessing, synthesizing, verifying, and interpreting data limit its overall utility to regulators, researchers, policy-makers, and the general public. Often where data may exist, political sensitivities render data opaque and unusable to the public. In other cases, data come in the form of highly normalized or aggregated statistics without transparency for raw underlying data.
3. **The need to determine the types of policy mechanisms and frameworks** that could help support efforts to address the central-local divide, and create more accessible and usable data. The creation of a 'big data lake" could bridge some of these gaps by allowing for multiple users from different sectors to contribute data.
4. **The need to harness citizen engagement and to help governments respond to these new sources of environmental information.** Determining how to incentivize citizens to collect environmental data, and how to evaluate and contextualize this information, will be crucial to efforts to generate third-wave data. Questions about data accuracy and reliability will be crucial to determining how governments can utilize and respond to citizen-generated data.

Presentations from the event are also available through a password-protected page on the Yale Data-Driven website.

Workshop Summary

Background and Project Overview

To begin the workshop, Dr. Angel Hsu, Director and Principal Investigator at the Yale Data-Driven Environmental Solutions Group and Assistant Professor of Environmental Studies at Yale-NUS College and Yale School of Forestry and Environmental Studies, traced the global evolution of environmental management. Increasingly, environmental management has shifted away from a command-and-control model, first to activities emphasizing economic and market-based measures, to now to the use of information disclosure to help protect the environment. Initiatives such as the United States Environmental Protection Agency's Toxic Release Inventory demonstrate the growing role information and disclosure play in driving environmental decision-making and incentivizing compliance with environmental laws. As the international community begins its work on meeting the recently agreed-upon targets outlined in the Sustainable Development Goals (SDGs) and the Paris Climate Agreement, finding data to measure and monitor progress will become even more crucial. Unconventional data, such as "big" data (i.e., satellite data) and citizen-generated data (i.e., birdwatcher's collection wildlife abundance and distributions data) can help complement conventional sources of environmental monitoring and fill data gaps. This type of "third wave data" offers new opportunities for driving environmental monitoring, management, and communication forward.

China's Environmental Data: A Gap Analysis

A previous assessment of China's environmental data provides a baseline for understanding the country's progress as well as its persistent data gaps and challenges. In 2012, the Chinese Academy for Environmental Planning (an initiative within China's Ministry of Environmental Protection) contacted the

Environmental Performance Index (EPI), to explore the possibility of adapting this global index -- which ranks countries on their environmental performance on indicators of ecosystem protection and environmental health -- to the provincial level. From 2008 to 2010, Dr. Hsu worked with the Chinese Academy for Environmental Planning (CAEP), along with the Center for International Earth Science Information Network at Columbia University and City University of Hong Kong to develop a composite index and indicator framework to better evaluate environmental performance and policy implementation at the provincial levels in China. Dr. Hsu, along with members of the Yale Data-Driven team, outlined the findings of this report, along with the changes in China's environmental data and data availability since the completion of this work.

The study scoping the possibility of a China Environmental Performance Index (EPI) found the greatest amount of information around measures of economic sustainability, such as economic energy efficiency, and efficient use of water in agriculture and industry. Information availability for measures of environmental health, such as access to tap water, and levels of pollutants like nitrogen dioxide and sulfur dioxide per person, followed close behind. Information tracking ecosystem vitality finished last, suggesting this may be a low priority for environmental management. In addition to data gaps, discrepancies between and within data sources further fragmented the information landscape. As a result, a China EPI was never implemented, due to the difficulties of assembling comparable data and the lack of consistent management targets.

An updated survey of data currently available reveals that some progress has been made, particularly in tracking air pollution, but gaps persist, especially around water, climate change and biodiversity. The workshop aims to more closely map data challenges, and opportunities to leverage third wave data to help fill them.

Introduction to Environmental Monitoring in China

A number of speakers honed in on the state of existing monitoring strategies in China, to help build an understanding of the available data and some of the key considerations facing efforts to augment it. Dr. Peiyuan Guo, from SynTao; Dr. Yaling Lu, from the Key Laboratory for Environmental Planning and Policy Simulation, Chinese Academy for Environmental Planning; Dr. Bo Zhang, from the Information Center of China's Ministry of Environmental Protection; and Chris James, from the Regulatory Assistance Project all helped describe the gains and challenges around monitoring efforts across China.

Understanding the Rise and Value of Crowd-Sourced Data

Chris James framed the session in terms of three key questions around crowd-sourced data, asking:

1. What is driving the issue of crowd-sourced monitoring?

   The trend towards crowd-sourced monitoring reflects a desire to augment and verify official sources of data. Some distrust over official sources of data and discrepancies between different sources of data may also have helped to spur this movement. Crowd-sourced data also makes it possible to hone in on pollutants that can vary widely at a local level; the ability to create more

tailored and highly specific models of pollutant exposure also helps explain the rise in interest in crowd-sourced monitoring.

2. What are the advantages of crowd-sourced monitoring?

   Possible advantages of citizen science include its ability to validate or complement government data, by, for instance, measuring pollution hotspots around transportation or industrial facilities in more detail. Crowd-sourced data can also help assess temporary, emergency, or episodic conditions, providing a greater level of data depth around especially crucial or significant events or challenges.

3. What are the policy and technical considerations tied to crowd-sourced monitoring?

   A complex set of policy and technical considerations confront the trend towards crowd-sourced data. Realizing the full potential of third wave data will require: aligning data recorded from different sources (i.e., creating consistent protocols with a basic verification procedure); using transparent data collection methods; and evaluating the impact of responses to citizen-generated data. These actions, in turn, will rest on coordination between citizens, non-governmental organizations and businesses, and local and national governments. With large datasets, for instance, outlier points may require diagnostic procedures; determining who will carry these out, and how, will shape the reliability of the data. It will also be important to consider personal exposure levels and risks and the content and delivery of any necessary training.

   In addition to these logistical and design questions, determining how to identify, engage and reach out to citizens, and to get governments to embrace and see the advantages in their involvement, will shape the future of crowd-sourced data.

Companies' Environmental Performance

Dr. Yaling Lu, of the The Key Laboratory for Environmental Planning and Policy Simulation, Chinese Academy for Environmental Planning, and Dr. Peiyuan Guo, representing SynTao, focused their remarks on efforts to track and measure environmental performance of companies operating in China.

Dr. Lu focused on the Key Laboratory for Environmental Planning and Policy Simulation's efforts to track the environmental performance of listed companies. In addition to conducting general data, the Laboratory publishes an assessment report(Titled:上市公司环境绩效评估研究报告)in 2015, focusing on key industries within Shanghai's stock exchange market. Since its start in 2009, with 156 companies, the yearbook has grown to encompass nearly 270 companies across seven industries. The method for ranking these companies is also improving over time, through adjustments to indicators, sub-indicators, and their respective weightings in the overall score. The Laboratory's experience creating this report has revealed data gaps and shortcomings. Some datasets include discrepancies and inconsistencies between different data sources (such as the gaps between company-reported data and government-reported data). While open data sources, such as the information collected by the Institute of Public and Environmental Affairs, has helped construct the yearbook, there is a need for additional research.

Dr. Guo highlighted similar challenges in his description of SynTao's work on MQI (Material and Quantitative Indicators Database), which collects nearly all of the corporate sustainability reports (CSRs) from companies operating across China. In 2007, SynTao gathered less than 100 CSRs; as more and more companies started to report their environmental performance, SynTao developed the MQI platform based on the original Sustainability Reports Center. SynTao expects the total number of reports collected for 2015 to exceed 2,500. The MQI database tracks disclosure and transparency as well as companies' actual environmental performance, as measured by 30 material indicators across 10 sectors, and all the data is searchable by year, region, indicator, and sector, among other features. SynTao's goals for strengthening the platform include making the platform easier to use; comparing companies in the same sector; and assessing progress at the sector level. Dr. Gao noted that while investors may not be worried about their absolute environmental performance, they are often motivated by their relative rank with their competitors.

Big Data in China's Ministry of Environmental Protection

Dr. Bo Zhang provided an overview of the Ministry of Environmental Protection's use of big data solutions, highlighting three key objectives guiding the Ministry's considerations of big data initiatives: (1) scientific decision making; (2) accurate supervision; and (3) public service. To achieve these objectives, the government is working to: strengthen sharing of data resources and their integration into scientific decision-making; innovate new methods of data collection; improve public services for the environment; and develop a big data platform. Initiatives to analyze and predict particulate matter (PM) pollution or to predict air quality using satellite data, demonstrate some of the potential use of big data with the Ministry of Environmental Protection.

Current Big Data Initiatives in China

A number of big data initiatives are working to address the gaps in the current landscape of environmental monitoring in China. Clement Dai, from IBM China; Fangyi Yang, from the Alibaba Foundation; Kate Logan, from the Institute of Public and Environmental Affairs (IPE); Jianyu Zhang, from Environmental Defense Fund (EDF) China; Ying Long, from Beijing City Lab and Tsinghua University; and Xiangying Irene Shi, from the Shan Shui Conservation Center provided updates on their work gathering third-wave data. Their presentations, and the conversation that followed, identified common obstacles, challenges, and areas of opportunity in efforts to access, synthesize, and apply data, as well as to understand the motivations driving these efforts forward.

Overview of Big Data Initiatives

The presenters outlined the initiatives their organizations are leading, also often citing other examples, case studies, or complementary efforts occurring across China. These initiatives span different sectors, from biodiversity and air and water pollution to corporate compliance, but reflect a common recognition of the potential of third wave data to help understand and bolster the effectiveness of current environmental management efforts.

- **IMB China** has worked with the government to develop the Green Horizons Initiative, which merges weather data with chemical modeling to predict air pollution. Recently, the initiative has also begun to focus on the relationship between energy and the environment, exploring ways to investigate the value of environmental services; apply energy use forecasting data to support the management of renewable energy (specifically through a project based in the city of Zhangbei); and establishing baseline data in areas including energy efficiency and carbon monitoring.

- **The Alibaba Foundation** has begun exploring the possibility of creating an Environmental Cloud Map, a data platform that would deliver data reports and assess environmental impact. The Foundation may also host an event at the Shanghai Open Day Apps (SODA) event in Shanghai this May, focusing on environmental protection. If it occurs, the event might invite governments to provide public data, and offer NGOs, the research community, and other users the opportunity to use this data to create new kinds of tools, visualizations, and other kinds of results. The Foundation's presentation also noted the work that the World Resources Institute is conducting in Hangzhou, citing it as an example of how big data can address environmental questions (in this case, around transportation).

- **The Institute of Public and Environmental Affairs (IPE)** provided an overview of the development of the Blue Map App and China Pollution Map Database, and outlined the the goals of IPE's Pollution Information Transparency Index (PITI), Green Supply Chain, and Green Finance initiatives. The presentation also previewed the update of the Blue Map App (Blue Map 3.0). IPE's work with businesses have allowed them to trace the growth in corporate supervision records since IPE began collecting them in 2006, along with the growing number of investors utilizing their platforms to assess businesses environmental records. In addition to the rise of the private sector's engagement with environmental data, some local Environmental Protection Bureaus (EPBs) have also taken important steps forward, notably by establishing Weibo platforms that allow citizens to tag EPBs as they flag potential environmental issues or compliance gaps, bringing a new level of transparency to this process.

- **The Environmental Defense Fund (EDF) China's** presentation focused primarily on the framework for data collection, noting the need to design a mechanism that both enables the government to disclose data and creates an incentive for data improvement. The presentation also highlighted the growing importance that will be attached to emissions data (particularly in light of the government's plan to establish a scaled-up emissions trading scheme by 2017) and to green finance data.

- **The Beijing City Lab** uses big data and open data to observe urban space and understand urbanization in China. The Lab both uses and produces open data in sources such as platforms, working papers, research data, slides, city rankings, and visualizations. Specifically, recent projects have focused on ranking the road intersections of 10 Chinese cities; monitoring ground level PM 2.5 pollution; using Weibo to track the social impact of odor from landfills; and using online searches in Baidu as a proxy for tracking levels of water pollution in rivers.

- **The Shan Shui Conservation Center** runs the China Nature Watch Biodiversity Information Platform (hinature.cn), which supports data collection around wildlife conservation, particularly in the western provinces of China. While researchers and NGOs have accumulated a great deal of biodiversity data, much of it is not publically accessible, or is submitted to donors, rather than to public or big data repositories. The potential for mobilizing citizens to fill gaps in biodiversity data is significant. The China Birdwatcher Society, for instance, has at least 6,000 registered members, a level of mobilization that helps explain why the Aythya baeri (a duck found in Northern China), currently has more data entries from citizen scientists than from government sources.

Key Challenges, Questions and Next Steps Facing Big Data Initiatives

A number of common challenges and questions facing big- and citizen-generated data emerged from the presentations, and the discussion that followed them. These challenges face both individual big data initiatives and attempts to link or merge them to create a more integrated network of information.

- **Data access and quality** emerged as a consistent challenge across the presentations. The state of ecological data in China, for instance, is emblematic of the challenges facing many other areas of environmental management. Efforts to track, study, and protect biodiversity are hindered by the lack of systematic monitoring (especially of wildlife and wild plants) and the absence of an open nation-wide database on biodiversity. In this and in other sectors, the lack of universal requirements for data disclosure; uneven or poor data quality; and gaps in disclosure consistency frustrate efforts to reliably compare environmental performance over time.

- The challenges of accessing and comparing data from the same source are often compounded by the challenges of **synthesizing data** from multiple sources. The difficulty of accessing and combining data across the institutional barriers separating different agencies, as well as those separating national, provincial, and city or local monitoring systems, and the need to eventually consolidate this data, emerged as a consistent theme. Combining information from different data providers -- NGOs, companies, and the scientific and academic community, in addition to the government -- adds another level of complexity to this challenge. Several presenters highlighted the need to evaluate and assess who holds different types of data, and to determine the willingness of different data owners to collaborate with others and to share data publicly.

- Understanding and harnessing different **motivations for data collection** also stood out as a key question. For instance, developing potential business models or funding sources for data collection may be especially crucial to more fully involving companies and the private sector. If the funding comes in the form of support from the National Development and Reform Commission (NDRC), what form will the resulting cooperation with the Ministry of Environmental Protection take? What industrial solutions can companies offer to promote the government's goal of sustainable development and Ecological Civilization Construction (生态文

明建设)?[1,2] The emerging carbon market offers interesting insights into the possible intersection of data collection and financial incentives: will the current annual reporting schedule be enough to support an emissions trading scheme, or will a more frequent reporting system be necessary? Developing a case study on the benefits of open data, one presenter suggested, could be an important tool for addressing some of these questions and building momentum for the movement.

Understanding the motivations of existing data providers -- such as their desire to protect the data they have collected -- will also be key to facilitating more integrated and effective data gathering and analysis. Determining ways of keeping citizens engaged and incentivized is necessary to maintaining and mobilizing widespread data collection.

- Beyond the challenges of data collection, questions about **data application** remain. If data exists, how can it be integrated into the government's decision-making process? Big data efforts will be successful when they not only encourage the disclosure of additional data, but also facilitate the continuous improvement of data collection and environmental management efforts. Monitoring the degree of greening that occurs as a result of additional data disclosure -- in addition to simply tracking environmental compliance with existing regulations -- will make it possible to realize the full potential of more robust information.

International Cooperation in China's Environmental Monitoring

The workshop closed with presentations that explored ideas and opportunities to be gleaned from ongoing efforts to build new kinds of data networks. Kong Chiu, Senior Program Manager of United States Environmental Protection Agency (EPA), provided an overview of the EPA's management of the United States' Greenhouse Gas Reporting System for large facilities, while Kevin Hart, from TZOA, and Yann Boa, from AirVisual, provided an overview of their companies' air quality sensors. All three presentations included discussions about ways to leverage these tools towards new ways of measuring and responding to air quality and the impacts of air pollution. In particular, a number of cross-cutting themes emerged from this portion of the workshop's discussion:

- The creation and use of a "data lake," a place for storing data in its native form, before processing it for further analysis, could help map and utilize existing environmental data across China. This approach could move information from silos into a common repository, facilitate the creation of more comprehensive analyses, and help translate data into more accessible forms of information.
- Efforts to measure air quality using portable, citizen-managed sensors will need to develop ways to identify the difference between indoor and outdoor environments, and to identify and determine the levels of exposure associated with different modes of transportation.
- Integrated platforms - that link sensors to mobile devices and larger databases -- are helping to build larger networks of citizen-generated information. Economies of scale could help amplify

[1] *Ecological Civilization.* (24 October 2007). *China Daily*. Retrieved from: http://www.chinadaily.com.cn/opinion/2007-10/24/content_6201964.htm.
[2] Opinions of the CPC Central Committee and the State Council on Further Promoting the Development of Ecological Civilization. (25 April 2015). Retrieved from: http://environmental-partnership.org/wp-content/uploads/download-folder/Eco-Guidelines_rev_Eng.pdf.

the usefulness of these platforms, validating data, identifying outliers, and creating detailed resolutions of air quality.

- Communicating the information generated by air quality sensors requires translating data into actionable information, and ensuring that data is contextualized to avoid unnecessary false alarms or panic.