Science & Technology Innovation Program

Authors
Anne Bowser
Alex Long
Metis Meloche
Elizabeth Newbury
Meg King

Filling Data Gaps:

A Citizen Science Solution

April 2020

Policymakers need data to make informed decisions. Local governments need data to justify policies like bans on single-use plastics. Federal agencies need information to set the conservation guidelines that protect endangered species. Data are also required to report on progress towards international policy targets, like the UN Sustainable Development Goals (SDGs).

But worldwide, we don't have enough data to understand the current state of our environment, or effectively evaluate the impact of interventions. In 2018, Washington, DC banned plastic drinking straws while citing evidence that 3,500 straws were picked up during Potomac Watershed and Earth Day cleanup campaigns. But raw data are not openly available to evaluate the effectiveness of this ban or understand the value of banning straws over other single-use plastics. A federal review across air, water, land, built, and sociodemographic environments found both general data gaps and spatial and temporal bias across all domains. Internationally, we lack enough information to track global progress against 68% of the 93 environmental Sustainable Development Goals.

Some additional data exist but are not yet open and Findable, Accessible, Interoperable, and Reusable (FAIR). To fill data gaps, we can open and remediate existing data, while using standards to collect new data through innovative methodologies that augment traditional reporting. Citizen science is one promising approach where members of the public voluntarily contribute to scientific research. By collecting information on scales and resolutions not achievable through professional activities alone, citizen science can help fill data gaps while engaging and educating public volunteers. A wealth of citizen science data already exists, though not all of it is open or FAIR. In addition, innovations in technology- like mobile applications (apps) and low-cost sensors for data collection, or data integration, visualization, and analysis tools- can support new data collection activities.

Building Blocks: Open and FAIR

Whether referring to government or data, "open" is an important ideal. In general, open strategies are associated with concepts like transparency, accountability, and participation. Within the US, open government, particularly open data, is an established policy priority with bipartisan support.

As early as 2009, the Obama Administration's Open Government Directive required that "*within 45 days, each agency shall identify and publish online in an open format at least three high-value data sets...and register those data sets via Data.gov.*" The 2018 OPEN Government Data Act (Kilmer, D-WA) took a step further. Signed into law by President Trump as the Foundations for Evidence-Based Policymaking Act of 2018, the Act establishes that the government "has the responsibility to be transparent and accountable to its citizens." In service of these ideals, all non-sensitive data will be open and machine-readable by default, and both sensitive and non-sensitive data will be documented in a federal data catalogue. The Act also requires each agency to develop an open data plan, allows agencies to collaborate with researchers, businesses, and private citizens on open data use, and designates a point of contact to assist the public and respond to complaints.

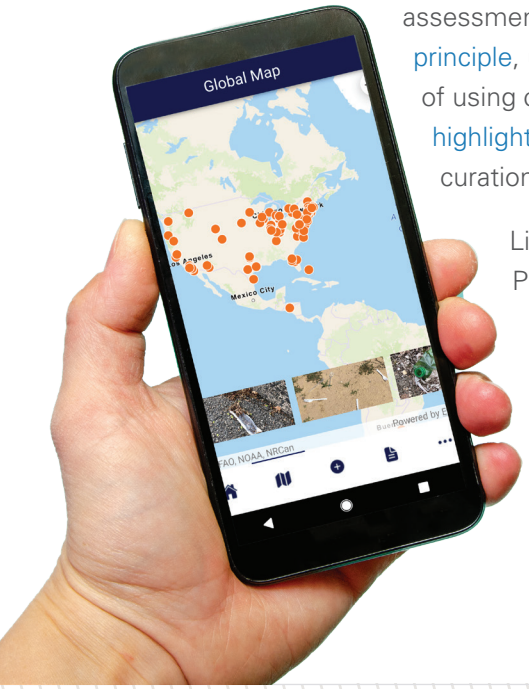
The Act contains some guidelines for implementation. For example, it defines open data as "*part of the worldwide public domain or, if necessary, published with an open license.*" This definition is important because it defines openness on a spectrum: the phrase "public domain" designates a legal waiver of copyright, while licenses advanced by organizations like the Creative Commons (CC) or Open Data Commons (ODbL) can make data open while preserving intellectual property rights. The Act also references many components of FAIR, a framework for describing how data can be made findable, accessible, interoperable, and reusable. As a complement to open, FAIR offers practical guidelines such as using metadata to document a data set, cataloging information in an open repository, leveraging open communication protocols like Application Programming Interfaces (APIs), and using disciplinary standards. While the Foundations of Evidence-Based Policymaking Act does not explicitly identify FAIR as a priority, all of the above provisions are addressed.

The Act is a major step forward. But while it will help increase the availability and documentation of existing data, it does not directly address the collection of the new data that will be required to fill environmental data gaps.

In general, open strategies are associated with concepts like transparency, accountability, and participation.

Citizen Science and Complementary Approaches

Citizen science projects have been supported by government agencies, educational institutions like museums and schools, community non-profits, and other non-government organizations (NGOs) for hundreds of years. Some citizen science projects, like community-led low cost air quality sensor assessments in North Carolina, are designed to collect data and drive change in local communities. Others, like the U.S. National Phenology Network's Nature's Notebook, contribute to national and even global research and assessment. Citizen science researchers and practitioners recognize data sharing as a core principle, unless there are significant security or privacy concerns. However, the process of using crowdsourcing to facilitate data collection at greater scales and resolutions also highlights the need to understand and support good practices around data validation, curation, and management.



Like open data, citizen science is supported by high-level public policy. In 2015, President Obama issued a Memorandum on Addressing Social and Scientific Challenges through Citizen Science and Crowdsourcing. More recently, the Crowdsourcing and Citizen Science Act, included as part of the 2017 America COMPETES reauthorization, offers federal agencies explicit authority to use these approaches to advance their missions. The Crowdsourcing and Citizen Science Act also states that data should be open and machine-readable by default. This, together with provisions in the Foundations for Evidence-Based

Policymaking Act that highlight the value of citizen engagement, suggests that open data and citizen science are complementary approaches.

Some practical resources to support citizen science exist. The website CitizenScience.gov includes a toolkit, a catalogue of federally-supported projects, and information on a community of practice and discussion list. The toolkit also contains information on data management that reminds readers to adhere to federal open data policies, find an authorized repository for long-term storage and catalogue data in "directories of data of similar types" as well as "the appropriate federal and agency open data catalogues." This information provides general, high-level guidance to a primarily federal audience, and is supported by case studies of citizen science to demonstrate practical, on-the-ground gains. But, policies and practical guidance through CitizenScience.gov stop short at pointing to, or providing, the necessary technological infrastructure to make citizen science data open and FAIR.

Citizen Science in Practice: Earth Challenge 2020

April 22nd, 2020 marked the 50th anniversary of Earth day. In recognition of this milestone the Wilson Center, U.S. Department of State, Earth Day Network, and many other partners launched Earth Challenge 2020 ("Earth Challenge") as the world's largest coordinated citizen science campaign to date. The emphasis on coordination means that Earth Challenge prioritizes working with existing citizen science projects to help make their data more open and FAIR. In addition, a new mobile application offers a concrete opportunity for public participation while providing an additional source of information.

2018 Citizen Science Camp, June 11-14, 2018
Summer nature camp for rising third through fifth graders at the Environmental Institute of Houston.
www.eih.uhcl.edu

Analyzing this initiative can demonstrate the value of citizen science, and reveal opportunities for acting on recent policy directives to help fill environmental data gaps. The initiative addresses six research areas: air quality, water quality, insect populations, plastic pollution, food security, and climate change. This case study will explore how citizen science can help understand the extent of plastic pollution measured at local, national, and global scales.

Working with existing data. Citizen scientists contribute to monitoring plastic by taking pictures of litter, collecting water samples, and reporting on pollution found during scuba dives. In addition, three citizen science projects-- NOAA's Marine Debris Monitoring and Assessment Project (MDMAP); the European Environmental Agency (EEA)'s Marine Litter Watch (MLW); and, Ocean Conservancy's Trash Information and Data for Education and Solutions (TIDES)-- ask volunteers to report on pollution after beach cleanup events. These projects work with individuals and communities to collect data and improve the health of local beaches. Many target their efforts within a national boundary. But, stakeholders including the UN Environment Program recognize the opportunity for this data to also be re-used in global assessments, like progress against SDG 14.1.1, which assesses the health of our oceans.

Earth Challenge 2020 began by working with these three citizen science projects to make their data interoperable (Box 1). Once this was achieved, additional steps were taken to provide access to well-documented information. A data integration and processing platform published the citizen science data through two APIs: One, exposing the data in its initial structure; and, one exposing data in an interoperable, Open Geospatial Consortium (OGC) compliant format. Publishing two APIs allows external users to understand and make tradeoffs between working with the initial, more granular data, and working with interoperable data.

In practice, interoperability is complex and requires resources to do well; the example of integrating plastics data from three projects reporting data on beach cleanup events demonstrates this. Because MDMAP, MLW, and TIDES all use different terms to describe or classify different types of plastics pollution, one initial challenge was cross-walking these terms to create a common vocabulary or schema. Consulting with experts in the field helped identify a schema for categorizing plastic debris published by Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO) as the most authoritative.

Through an iterative process, researchers mapped each existing data set to CSIRO's schema, adjusting the schema as needed. This process often required taking granular data and making it more generic. For example, MDMAP asks volunteers to share information about "cups" and "silverware"; MLW, asks about "cups and cup lids" separately from "cutlery and trays"; and, TIDES asks volunteers for information on "cups, plates (plastic)" and "cups, plates (foam)." Cross-walking these schemes required

aggregating all of these categories into "plates, bowls, cups, or silverware" to allow for comparison across the three data sets.

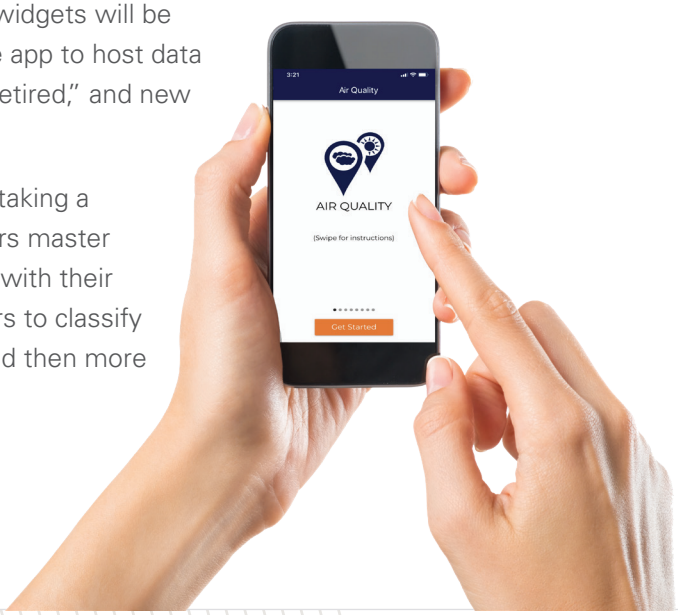
After developing an interoperable model for classifying debris, the Open Geospatial Consortium (OGC)'s SensorThingsAPI standard was consulted to help make other information included in these data sets interoperable. SensorThingsAPI was initially developed for an Internet of Things (IoT) use case. Citizen science scenarios often rely on data collected through hardware sensors and low-cost mobile phones, so SensorThingsAPI is a natural fit. In addition, earlier efforts using SensorThingsAPI for citizen science, and a community of practice investing in the standard helped establish its value for interoperability.

After cross-walking the information in each data set with SensorThingsAPI, researchers referenced the data standard to find out what additional information was required to reach "compliance." In all cases, collecting supplemental information was needed to help document information on things like data quality and designate intellectual property rights through a standardized license.

Information on the three data sets is documented in an emerging Citizen Science Cloud. This Cloud offers a catalogue function, where people seeking to use citizen science data can find, understand, and access high-value information. Ultimately, to maximize value, the Citizen Science Cloud will share metadata with other catalogues and repositories, as discussed later.

Producing new information. The Earth Challenge 2020 mobile application is a generic mobile application (app) that hosts a series of data collection widgets. For Earth Day 2020, the app was launched with two widgets, Plastics and Air Quality. Over the course of 2020, four additional widgets will be added. After 2020, the design decision to create a generic mobile app to host data collection widgets could allow any of the first six widgets to be "retired," and new widgets created, on an as-needed basis.

The Plastics widget first invites volunteers to "Report Plastic," by taking a picture of any piece of plastic pollution, anywhere. After volunteers master this task they unlock a second opportunity– "Train AI." Beginning with their data, and progressing to other photos, Train AI task asks volunteers to classify pictures from Report Plastic first generally (e.g., "soft plastic"), and then more specifically (e.g., "cigarette butt").



Data from the Plastics widget will be used in two ways. Once labeled, information collected through "Report Plastic" will be combined with data from MDMAP, MLW, and TIDES to help create a top ten list of global plastic pollution broken down by country. Pictures labeled through Train AI will be used to teach a machine learning (ML) solution to identify different types of plastic found in the environment. Over time, a mature ML solution linked to the Plastics widget will give citizen science volunteers real-time information about what they are seeing. In line with broader ideals of openness, the ML solution will also be made publically available.

One impact of Earth Challenge 2020 will be elevating the value of citizen science in six research areas. The initiative should provide evidence for the value of working with existing citizen science projects while creating new tools for data collection, analysis, sharing, and documentation to bridge work on local, national, and global scales. A second impact will be the contribution of key technologies to facilitate this approach. In partnership with other OGC members, we plan to continue our work with the SensorThingsAPI standard. As consensus around the value of SensorThingsAPI continues to grow, so will the utility of this standard for organizing information within the citizen science community. The data integration platform and Citizen Science Cloud will help leverage this standard to expose citizen science data to broader research and policy communities working with a range of different information types.

Filling Data Gaps Through Citizen Science

MDMAP, EEA, and TIDES all work with individuals or community groups that are cleaning beaches and sharing data. All three data sets are open to varying degrees, though, at the time of writing, none have independently published an API. Earth Challenge 2020 offers new forms of access to their data, as well as better documentation, through the use of standards, APIs, and cataloguing. Similar approaches can, and should, be applied to make existing citizen science data more open and FAIR in a range of environmental research domains.

In addition to analyzing data on pollution, visualizations analyzing the location of citizen science activities are helpful for understanding current trends of participation and gaps. For example, across MDMAP, MLW, and TIDES, Australia has engaged 10,774 volunteers in 525 cleanup events from 2015-2018. In contrast, Vietnam has documented participation from 3,490 people and 49 cleanup events in the same time frame.

There are probably a number of reasons for this discrepancy. Regarding differences in the populations of the two countries, Vietnam has a larger and more dense population, but less coastline (this is also reflected in the proportionately higher number of volunteers at each Vietnamese cleanup event). But factors not related to population differences are also worth considering. Citizen science is relatively mature in Australia, with a Citizen Science Strategy recently advanced by the government of New South Wales. Australian researchers are also supported by a professional association to help support the citizen science community and foster knowledge exchange, but no similar entity exists in Vietnam.

There are a number of opportunities to invest in citizen science to help fill data gaps. Capacity building efforts could focus on launching citizen science projects in different communities. Research shows that citizen science

volunteers are often involved in the design of new data collection and analysis activities, using their local knowledge to shape what is studied, and how. An important consideration is therefore how to make space for local knowledge while also recognizing the value of data sharing and interoperability.

One opportunity is found in the use of different citizen science platforms. iNaturalist can be used to collect, curate, and share any type of biodiversity data. Zooniverse supports data analysis across research domains by asking volunteers to classify or annotate images. Each of these platforms can be incorporated into a local citizen science campaign, though the degree that different platforms allow for customization, or adhere to standards to promote interoperability, varies.

Another opportunity lies in sharing information about metadata in a standardized way. Even if data are not initially open or interoperable, cataloging information in a community (like the Citizen Science Cloud), a domain specific repository (like the Global Biodiversity Information Facility (GBIF)), or a government repository (like Data.gov), is an initial step towards responsible reuse. For US federal agencies, cataloging information is not only important but legally required through the Foundations for Evidence-Based Policymaking Act.

In addition to filling data gaps directly, research conducted through citizen science can help advance new methodologies.

In addition to filling data gaps directly, research conducted through citizen science can help advance new methodologies. In the case of plastic pollution for SDG monitoring, a second emerging methodology involves the use of drones to survey areas more quickly and comprehensively than on-the-ground monitoring allows. The opportunity here is also the challenge, as vast quantities of drone data are difficult to analyze. Through approaches like the Train AI module of the Earth Challenge 2020 plastic widget, citizen science can not only collect but also label the images needed to train machine learning approaches. These approaches can then be applied to analyze other types of big environmental data. Beyond drone data, satellite images collected by NASA and NOAA in the US, and foreign collaborators like EEA, are another underutilized source of information that citizen science and ML can help unlock.

Conclusions and Next Steps

For citizen science. The Earth Challenge 2020 case study demonstrates the value of existing citizen science data and the opportunity afforded by new technologies and methodologies to help fill environmental data gaps. In the US, we have a strong start for citizen science for open data through policy support. The agency open data plans required by the Foundations for Evidence-Based Policymaking Act present an additional opportunity to include guidance on citizen science approaches. Agency Open Data Coordinators could work with their counterparts in the Crowdsourcing and Citizen Science Coordinator role to align the two approaches within a specific agency.

Looking across agencies, more concrete guidance is needed on technological resources such as platforms to support citizen science data collection, analysis, management, and documentation. For a light touch approach, the CitizenScience.gov resources could be expanded to include links to such tools. In the future, the FedRAMP processes could help ensure that all relevant privacy and security concerns are met.

Given the federal investments allocated by granting agencies to universities, museums, and schools, we also need to support citizen science outside of the beltway. Procuring or developing technologies, and making these broadly available, could be one contribution. Equally important is finding mechanisms to offer researchers and practitioners community support (e.g., through the Citizen Science Association, a professional organization connecting practitioners, researchers, and others that design, lead, manage, and study citizen science).

For FAIR and open data. To elevate the value of existing data, it may make sense to begin with FAIR principles that documentation and discovery through, for example, metadata and cataloguing. Documenting existing information assets forces data stewards to think critically about data quality, privacy, and security concerns, including how suspect data are flagged for review, what constitutes personally identifiable information (PII), and what data license could best enable ethical reuse. Focusing on FAIR before open also gives data stewards time to assess interoperability goals.

Within and beyond citizen science, there is often a tension between meeting specific, local data collection needs and making information and interoperable. Earth Challenge 2020 shows one path forward: beginning with one

Mount Rainier National Park, Cascade Butterfly Project
https://creativecommons.org/licenses/by/2.0/

or more standards and a relatively small number of local data sets it is possible, though labor intensive, to create interoperable information. In the future, data providers like citizen science projects could work together to build consortia dedicated to achieving interoperability in a particular area of environmental research, leveraging cross-cutting standards like OGC's geospatial data standards in the process.

For coordinated environmental monitoring. Different local communities, whether individual data collectors or communities of practice, have unique needs. If the goal is efficiency and reuse, technologies designed to support citizen science should be as flexible as possible. The Earth Challenge app, iNaturalist, and Zooniverseplatforms all take modular approaches by providing a generic "framework" that "widgets," "projects," or "campaigns" can build on. Modular structures encourage innovation and enable future developments around a diversity of environmental concerns more efficiently than single use technologies. Reusable infrastructures also provide opportunities for communities to become accustomed to a certain tool or platform.

As a community of practice, researchers who use citizen science share norms around openness and attribution, as well as the challenge of defending data quality. These norms and challenges inform how data are collected, analyzed, and most documented to maximize value for reuse. Technologies developed specifically for and with the environmental citizen science community will often show diminishing value when used out of context.

Zooming out, a modular information architecture would recognize the unique value of citizen science technologies and platforms while linking these to other systems. Ultimately, a range of systems should be brought together under one or more federations. The Global Earth Observation System of Systems (GEOSS) is one federation created to support the broad community of Earth observation data producers and users. A UN-backed initiative, the Digital Ecosystem for the Environment, outlines a second. A federated system-of-systems approach for environmental data collection, management, and exchange will enable local (regional, disciplinary, or sector-specific) platforms to emerge as centers of excellence, while enabling the combination of different types of information for a range of uses.

The environment transcends national boundaries, and public policy helps determine how data can be shared across borders.

Broader policy implications. Data and public policies will shape how this unfolds in practice. Recent open data policies primarily target the federal government, and citizen science engages the general public, but more support is needed. On the data policy side, increasing the amount of open data will require revisiting traditional, one-size-fits-all approaches to intellectual property, in order to effectively engage the private sector. As one example of a more flexible approach, the satellite data provider Planet.com often makes proprietary data available for specific use cases, such as disaster response. Developing open templates for data sharing under such cases, with legally enforceable provisions, could help encourage good behavior across private sector stakeholders that are not naturally incentivized to make their information open for reuse.

The environment transcends national boundaries, and public policy helps determine how data can be shared across borders. Europe's General Data Protection Regulation (GDPR) is becoming the de facto standard for preserving personal privacy. Citizen science and other methods of data collection should have the ability to meet GDPR compliance if their goal is to facilitate aggregation and reuse in international assessments like the SDGs. While GDPR may limit data sharing in the short term, once data providers adjust, benefits may include increased citizen trust-- and therefore, in the case of citizen science, more voluntary data contributions.

The recently ratified US-Mexico-Canada trade agreement (USMCA) for the first time in history enables data sharing through data transfers across borders. Working out the details of implementation between North American governments could pave the way for other multilateral agreements, and potentially help serve as a model for future public-private data sharing mechanisms. Citizen science projects collecting data across North America could offer added value through science diplomacy, connecting people from diverse communities with different cultural backgrounds.

Ultimately, the environment is a shared asset. Understanding and preserving it requires data. Governments and citizens should be enabled to work together to secure the information we need.

Thanks to Alison Parker and Erin Rohn for assistance and helpful advice.

SCIENCE AND TECHNOLOGY INNOVATION PROGRAM

The Wilson Center's Science and Technology Innovation Program (STIP) brings foresight to the frontier. The Program examines emerging technologies through vital conversations, making science policy accessible to everyone -- not just scientists. The modern era is defined by exponential leaps in scientific understanding and technological breakthroughs. But the potential impacts of these breakthroughs – in research, in policy, in society – are not always clear. We analyze and translate how emerging technologies will impact international relations, from cybersecurity to artificial intelligence to Big Data. We equip decision-makers and the public with the tools to help understand advancements in science and technology. From hack-a-thons to seminar series for Congressional staff, our goal is to help communicate complex policy topics in an engaging, accessible way. Science policy is not just for scientists. STIP engages with audiences from middle school classrooms to the Hill to help educate and inform the American people about leading issues in public policy. We value science that is experiential and participatory, whether captured through citizen science or serious games.

WOODROW WILSON INTERNATIONAL CENTER FOR SCHOLARS

The Woodrow Wilson International Center for Scholars, established by Congress in 1968 and headquartered in Washington, D.C., is a living national memorial to President Wilson. The Center's mission is to commemorate the ideals and concerns of Woodrow Wilson by providing a link between the worlds of ideas and policy, while fostering research, study, discussion, and collaboration among a broad spectrum of individuals concerned with policy and scholarship in national and international affairs. Supported by public and private funds, the Center is a nonpartisan institution engaged in the study of national and world affairs. It establishes and maintains a neutral forum for free, open, and informed dialogue. Conclusions or opinions expressed in Center publications and programs are those of the authors and speakers and do not necessarily reflect the views of the Center staff, fellows, trustees, advisory groups, or any individuals or organizations that provide financial support to the Center.

Woodrow Wilson International Center for Scholars
One Woodrow Wilson Plaza
1300 Pennsylvania Avenue NW
Washington, DC 20004-3027

The Wilson Center

www.wilsoncenter.org
wwics@wilsoncenter.org
facebook.com/woodrowwilsoncenter
@thewilsoncenter
202.691.4000

STIP

www.wilsoncenter.org/program/science-and-technology-innovation-program
stip@wilsoncenter.org
@WilsonSTIP
202.691.4321

Wilson Center

Science and Technology Innovation Program