



USAID
FROM THE AMERICAN PEOPLE

Crowdsourcing to Geocode Development Credit Authority Data: A Case Study



AUTHOR CONTACT INFORMATION

Shadrock Roberts (PPL/ST/GeoCenter) -shroberts@usaid.gov

Stephanie Grosser (E3/Development Credit Authority) - sgrosser@usaid.gov

D. Ben Swartley (PPL/ST/GeoCenter) -dswartley@usaid.gov

U.S. Agency for International Development
1300 Pennsylvania Ave. NW
Washington, DC - 20523
U.S.A – Planet Earth

NOTES

The subsection, “Phase 3: Accuracy Assessment” was written by –and used by permission from– the GISCorps.

ACKNOWLEDGMENTS

A large number of individuals volunteered their time for the realization of this project. In no particular order the authors would like to especially thank: Estella Reed, Patrick Meier, Kirk Morris, Bharathi Ram, Leesa Astredo, Melissa Elliott, Jeannine Lemaire, Shoreh Elhami, David Litke, Kate Gage, Chris Metcalf, Katie Baucom, Tyler Garner, Salim Sawaya, The Original Master Cluster, John Crowley, Andrew Turner, Lea Shanley, Radiohead’s “Optimistic”, all the volunteers who showed up for the live event and those who participated online, our bosses Alex Dehgan and Ben Hubbard for believing in this, and caffeine for giving us the strength to carry on.

DISCLAIMER

The authors’ views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

June 2012. Produced by USAID’s GeoCenter and USAID’s Development Credit Authority

Cover photo: Volunteers at the live crowdsourcing event in Washington, D.C on June 1st, 2012.

Photo Credit: Shadrock Roberts.

Table of Contents

EXECUTIVE SUMMARY	1
THE CONCEPT	3
WHAT IS CROWDSOURCING?	3
WHO WAS THE “USAID CROWD?”	3
THE DATA AND THE GOAL	4
THE DEVELOPMENT CREDIT AUTHORITY	4
THE GOAL: POTENTIAL IMPACTS FOR OPENING THE DATA	4
FINDING A SOLUTION TO MAPPING	5
THE INITIAL PROBLEM TO SOLVE: NON-STANDARD LOCATION INFORMATION	6
DRAWING ON WHOLE-OF-GOVERNMENT RESOURCES TO IDENTIFY PLACE NAMES	6
BRINGING IN THE CROWD	6
THE PLATFORM	7
POLICY ISSUES AND NECESSARY CLEARANCES	7
USING CROWDSOURCING IN THE GOVERNMENT	7
FREE LABOR	8
NON-DISCLOSURE ACT COMPLIANCE	8
RELEASING PUBLICLY IDENTIFIABLE INFORMATION	9
INFORMATION QUALITY ACT COMPLIANCE	9
WORKFLOW	10
THE CROWD’S TASK	10
ASSEMBLING THE CROWD	11
REACHING OUT TO VOLUNTEER TECHNICAL COMMUNITIES (VTCs)	11
DRAFTING THE SCOPE OF WORK TO DEPLOY VTCS	12
MARKETING THE CROWDSOURCING EVENT TO POTENTIAL VOLUNTEERS	12
IMPLEMENTING THE CROWDSOURCING EVENT	13
PRE-EVENT	13
PHASE 1: CROWDSOURCING	14
PHASE 2: DATA PROCESSING AND MAPPING	15
PHASE 3: ACCURACY ASSESSMENT	17
PUBLISHED MAPS AND DATA	19
DETERMINING WHAT TO MAP	19
ADOPTING IATI AND OPEN DATA STANDARDS	19
GEOGRAPHIC AND LICENSING ISSUES OF USING ADMIN1	19
SUMMARY AND LESSONS LEARNED	20
POLICY ISSUES	20
REACH-BACK TO CROWD	21
OPERATIONALIZING A CROWDSOURCING EVENT	21
PUBLISHING DATA AND MAPS	22
CONCLUSION	23
WORKS CITED	23

Executive Summary

The United States Agency for International Development (USAID) launched its first crowdsourcing¹ event to clean and map development data on June 1, 2012. At that time, no one predicted that all records would be completed in just 16 hours – a full 44 hours earlier than expected, which is precisely what happened. By leveraging partnerships, volunteers, other federal agencies, and the private sector, the entire project was completed at no cost. Our hope is that the case study will provide others in government with information and guidance to move forward with their own crowdsourcing projects. Whether the intent is opening data, increased engagement, or improved services, agencies must embrace new technologies that can bring citizens closer to their government.

USAID's GeoCenter, working in cooperation with the Agency's Development Credit Authority (DCA), identified a global USAID dataset of approximately 117,000 records that could be mapped and made open to the public. Significant data cleanup, however, was necessary before this was possible. USAID utilized a crowdsourcing solution for the data cleanup that had three primary advantages for the Agency:

- **Substantive Impacts:** The data describe the locations of loans made by private banks in developing countries through a USAID risk-sharing guarantee program. Making the data publicly available can lead to a variety of important analyses.
- **Transparency Impacts:** USAID is working to make more of its data publicly available. By doing so, the public can make significant and creative contributions to how USAID does business.
- **Establishing cutting-edge methods for data processing:** This is the first time that USAID has used crowdsourcing for help processing its data. This project serves as an example for future public engagement.

Prior to this event, the DCA database could only be mapped at the national level despite the existence of a very large amount of additional geographic data that has been collected since the inception of the program. At the national level, the entire data set can be mapped with an accuracy of 100 percent. The goal of this project was to add value to the data set by allowing users to map or query data at a finer level of granularity.

USAID partnered with federal colleagues in the Department of Defense (DoD) and General Services Administration (GSA), Socrata and Esri in the private sector, and volunteer technical communities (VTCs) Standby Task Force and GISCorps. In the end, these partnerships allowed USAID to automate geocoding processes that refined 66,917 records at 64 percent accuracy while a crowdsourcing process refined an additional 7,085 records at 85 percent accuracy. Our results confirm that crowdsourcing and using volunteered data can, indeed, be more accurate than other processes and establishes a promising precedent for future projects.

The reliability of crowdsourced and volunteered geographic information has been a persistent focus of research on the topic (Elwood, 2008; Haklay, 2010; Goodchild and Li, 2012). As this research notes, there is no reason to, *a priori*, suspect that these data are any less reliable than so called "authoritative data."

¹ Crowdsourcing is a distributed problem-solving process whereby tasks are outsourced to a network of people known as "the crowd."

As is true with any innovation, this project was a learning experience. Listed below are improvements and recommendations for any public sector, development, or humanitarian agency that would like to pursue a crowdsourcing path to data processing and public engagement.

- Agencies should involve their General Counsel from the outset to ensure that the project does not raise any legal issues and/or violate any policies/regulations. Every attempt should be made to disclose the nature of the data that volunteers are working on and ensure that they understand the purpose of the project. If certain information cannot be disclosed, these parameters need to be defined at the beginning of the project. When possible, a forum should be provided for questions to be answered to more completely engage volunteers in the goal of the project.
- Crowdsourcing a task should be understood as a project – like any other – that requires both management and a considerable amount of communication among partners to ensure a mutually beneficial experience and positive outcomes. Any organization that is planning to engage with crowdsourcing or VTCs regularly should build this management capacity into its organization.
- Agencies organizing crowdsourcing events should work closely with volunteer coordinators to provide the most appropriate guidance, for example by using several types of media (documents, videos, online chatting) to maximize volunteers' time.
- It is essential to have consistent and dedicated support for all technological aspects of such a project. All applications should be sufficiently tested to ensure that they can support more volunteers than expected.
- Development and humanitarian mapping projects would benefit from greater investment in existing initiatives to create and maintain updated, open, global boundary sets such as the United Nation's Second Administrative Level Boundaries or the United Nation's Food and Agriculture Organization's Global Administrative Unit Layers.
- Likewise, development and humanitarian mapping projects would benefit from greater investment in the National Geospatial Intelligence Agencies GEOnet Names Server (GNS) database in terms of content and usability.

This case study is meant to help individuals inside government looking to engage the public in new ways, and to individuals outside government hoping to understand some of the challenges and limitations the government faces in opening data. Ultimately taking risks with events such as this one is key to helping all parties achieve more in a smarter, more collaborative way.

The Concept

Crowdsourcing is a relatively new phenomenon that has evolved significantly due to the emergence of Web 2.0 technologies that facilitate assimilating several small contributions into a larger effort. In the humanitarian and development context crowdsourcing and associated themes rose to the forefront during the 2010 earthquake in Haiti. This was perhaps most visible in the “Ushahidi Haiti Project” through which the local population used text messaging to send requests for help.

Since then, the information landscape has continued to evolve. The humanitarian and development sector has identified innovative ways to incorporate new data and methods into well-established work flows, and leaders within “the crowd” have begun to formalize relationships and methodologies. While still nascent, increased public participation using new technology presents a shift in how the U.S. Government engages with its citizens and how citizens can participate in and direct their government.

The use of crowdsourcing for humanitarian or development interventions has spurred a lively debate about the attendant advantages and disadvantages of this approach including – justifiably – many questions surrounding data quality, security, and usability. Our experience will show that these questions were confidently addressed through careful planning and extensive dialogue with our partners. In addition to the substantive impact of having a clean dataset and map to release publicly, USAID was eager to explore this new way to engage with interested individuals anticipating that we would identify further applications of this methodology to further our work.

What is Crowdsourcing?

The neologism “Crowdsourcing” first appeared in 2006 to describe the phenomena whereby tasks are outsourced to a distributed group of people or “crowd” who is generally considered to be made up of non-experts and is further differentiated from formal, organized groups such as paid employees by their distributed nature (Howe, 2006). However, no set definition yet exists since it can be used to describe a wide group of activities that take on different forms. Reviewing the definitions currently in use, Estellés and González (2012) propose the following:

"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."

Who was the “USAID Crowd?”

A common question regarding crowdsourcing is who, exactly, makes up “the Crowd?” Put most simply, the crowd will be composed of individuals who are interested in the task at hand. Because most crowdsourcing involves access to a computer, internet, and mobile devices, certain characteristics can potentially be inferred about members of the crowd (e.g. those with access and capacity to use these tools).

Because USAID's project demanded the ability to quickly and thoroughly investigate partial or problematic locational data, USAID chose to partner with online volunteer communities – known more specifically as volunteer technical communities (VTCs) – to form the nucleus of the crowd while also soliciting general public engagement through various social media platforms such as Facebook and Twitter, and raising the awareness of this groundbreaking initiative. This had the benefit of ensuring that a minimum level of capacity for the task would exist in the Crowd while, at the same time, providing any interested individual with an opportunity to get involved. The two VTCs that partnered with USAID on this project were the Standby Task Force (SBTF) and GISCorps.

The Data and the Goal

The Development Credit Authority

All of the data in this project represent individual, private loans made possible by USAID's Development Credit Authority (DCA)². Through DCA, USAID issues partial credit guarantees to encourage lending to underserved sectors and entrepreneurs in developing countries. USAID typically shares fifty percent of any defaults as a result of the targeted lending with the financial institution.

Since DCA was established in 1999, more than 300 guarantees have been established with private financial institutions in developing countries. Over the years, up to \$2.3 billion in local capital has been made available for 117,000 entrepreneurs in all sectors. The default rate is just 1.64 percent across the portfolio, proving the profitability and creditworthiness of these new sectors and borrowers. USAID has only paid out \$8.6 million in claims, while collecting \$10.6 million in fees, for billions of private capital mobilized.

The Goal: Potential impacts for opening the data

Better Serving Entrepreneurs

By creating a map specifically listing available financing, USAID is making it easier for entrepreneurs to see where they could qualify for local financing. In addition, organizations working to help certain groups of entrepreneurs around the world access financing can take advantage of the USAID guarantee map to connect their networks with available financing. While the map does not list bank names or contact information, it provides a contact e-mail address (DevelopmentCredit@usaid.gov) so individuals can connect with local banks via USAID staff.

Targeted Lending

Visualizing loan data on a map can change the way USAID's in-country Missions plan for future guarantees. Guarantees are often targeted outside of capital cities or in certain regions of a developing country. By seeing where the loans are concentrated, USAID Missions can better analyze if the guarantees are fully reaching the targeted regions. In addition, the maps allow USAID to overlay additional open data sets on the USAID map. By adding a layer of open World Bank data on financial inclusion USAID can quickly see where needs and intervention align.

Analyzing Transnational Impact

² Visit DCA's web site at: www.usaid.gov/what-we-do/economic-growth-and-trade/development-credit-authority-putting-local-wealth-work [last accessed June 22, 2012].

For the first time, USAID loans can be easily analyzed across country borders. For example, if the map shows that in one country a region has all of its loans going toward agriculture but a bordering region in another country has all of its loans going toward infrastructure, it may suggest the need for future collaboration between USAID Missions. Without this type of analysis, USAID Missions in one country wouldn't have time to analyze the location of all loans for guarantees in surrounding countries.

Improved Partnerships

While USAID and other donors often try to collaborate to maximize impact; there is no overall database of active guarantees offered by all development agencies. By making accessible the map service layers, other donors can compare or even overlay their guarantee data to identify opportunities to increase collaboration.

Previous Steps in Releasing DCA Guarantee Data

Initial Public Data Release for DCA Guarantees

In December 2011, DCA released data on its 300 active and expired guarantees. The released dataset showed all partial credit guarantees that USAID has issued since DCA was founded in 1999. The spreadsheet detailed the full facility size of each guarantee, how much was lent under each guarantee, the status of the guarantee (i.e., active or expired), how much in claims the bank submitted due to losses it incurred for loans placed under the guarantee, which sectors each guarantee covered, and how many loans were placed under coverage of the guarantee. Since releasing that data USAID received and complied with requests from partners and the public asking for the Agency to release the percentage guaranteed for each record.

Releasing Additional Loan Data

In 2012, DCA decided to map its reach and impact, and release that information to the public, to improve programming and analysis of its work. To map activities more precisely than the country level, USAID needed to release information related to each individual loan for all active and expired guarantees. While the first dataset contained 314 guarantee records, the second data set contained 117,000 loan records. Previously, loan records were primarily used by USAID to ensure that banks were making loans to the correctly targeted sectors as per the guarantee agreement. By performing in-person audits of the transaction records, USAID staff was able to confirm financial institutions were inputting accurate data into the Credit Management System³, and could therefore pay claims related to those loans. USAID loan data has never been analyzed outside of the Agency and its independent evaluators.

Finding a Solution to Mapping

USAID performed an initial analysis to look for patterns that would inhibit or allow a crowdsourced approach for geocoding DCA data and conducted basic tests involving the number of records that an individual could process during a crowdsourcing event. It quickly became evident that manually processing 117,000 records would be a task that would require a minimum of several hundred volunteers working for months due to the amount of time it would take a volunteer to process each record. A majority of the records, however, contained information regarding the first administrative unit of that country (or "Admin1" in geographical terms), which in the United States is the state level. Indeed, this was the only information given for records

³ The Credit Management System is an online database where USAID's financial partners input data regarding the loans they make against USAID guarantee agreements.

in certain countries. Based on this feature of the data, Admin1 became the minimum mapping unit of the processed dataset, with finer scale resolution (place name) included as an ancillary entry where possible.

Although the idea of crowdsourcing the geo-tagging of the DCA database was present from early on, USAID also considered traditional approaches such as using existing labor or contractors. Each approach was evaluated on the basis of practicality, reliability, and cost. In the end, our approach was a hybrid method that involved contributions from other federal partners, private industry, and both the interested public and volunteer online communities that made up the Crowd.

The Initial Problem to Solve: Non-Standard Location Information

The DCA database was originally structured to capture information regarding the amount, sector, and purpose of each loan as per the guarantee agreement and paid less attention to the geographic specificity of each loan. Users who entered data were given a single field marked “City/Region” and all geographic information was stored as free-form text in a single column in the database.

Typically databases have detailed geographic information collected in separate fields that are machine readable. The DCA database, on the other hand, did not originally envision a demand for mapping its data and did not separate these fields. Moreover there was no standardization given for how to enter various pieces of information (e.g., spelling of place names, abbreviations to use, separation of discreet pieces of information by commas). This unstructured, non-standard input translated into a column of information containing only partial geographic information that could not be automated for mapping. USAID’s first task was to rectify this.

The DCA database has now been updated to have partners input location information into three separate fields: national; Admin1; and a populated place, which is generally the name of a village, town, or city. In order to keep the database standardized, its fields are now linked to an external gazetteer, or geographical directory (GNS).

Drawing on Whole-of-Government Resources to Identify Place Names

Working with U.S. Government partners in the Department of Defense (DoD), USAID developed a basic automated process that standardized and searched each record for any identifying features in the National Geospatial Intelligence Agency’s (NGA) online gazetteer “GNS Names Server.” Because the national scale information was correct, the automated process searched only for matches within the specified country. In cases where a place name was found, this was added to the record and used to generate both the Admin1 information and populate latitude and longitude based on the geographic center of that place (or centroid).

There were 66,917 records that could be automated to derive the needed Admin1 information and an additional 40,475 records contained no sub-national geographic information or could not be mapped at the Admin1 level. The remaining 9,607 records, which contained the most problematic and partial geographic data, required human processing.

Bringing in the Crowd

After thoroughly and carefully conceptualizing the remaining problem, it was determined that crowdsourcing would be the best approach to move forward. USAID would present the Crowd with a clearly defined, bounded task that could be completed in approximately 15 minutes per record. As the project methodology developed, USAID conducted two practice runs with small groups of volunteers. One early test showed that volunteers grew frustrated by processing multiple entries with duplicate geographic

information. Based on this, and with the help of DoD, USAID developed a method of pre-processing whereby duplicate entries were collapsed into a single, “parent” entry to be given to the crowd. The parent entry then would be used to populate its associated duplicate records.

In sum, the final hybrid approach was a mixture of automated processes and crowdsourcing. Pre-processing involved stripping potentially sensitive data from the entire dataset, using the automated process to generate Admin1 and place name information where possible, and grouping multiple entries of duplicate information into a single record.

The Platform

Once USAID decided to move forward with a crowdsourcing solution, an appropriate platform was needed to enable the Crowd to view and edit the data. Internal market research turned up the following options:

- 1) Building a crowdsourcing platform for USAID to host similar projects in the future. This way the Agency would be able to build and cultivate an interested community within an engagement platform.
- 2) Using an existing tested platform on the market, for example, Amazon’s Mechanical Turk.
- 3) Utilizing a pre-existing government option. USAID discovered that data.gov has the potential to be a platform for crowdsourcing. Data.gov currently hosts data in order to increase transparency with the public. This platform is already built and paid for by the General Services Administration (GSA) and is available for all U.S. Government (USG) agencies to use. By uploading the dataset as “private”, then inviting the crowd to access it, the platform could be used at no cost.

Besides cost and utilizing pre-existing platforms, USAID also had to decide to either give volunteers a form where they would only see one record at a time, or give volunteers access to a spreadsheet to view multiple records at a time and have access to all records they previously geocoded. Ultimately a spreadsheet format, available through data.gov, made more sense so people could reference records they had already completed or make corrections to past records if necessary.

For other government agencies interested in emulating this process, it should be noted that the setup for the USAID crowdsourcing application which connected to the data.gov site was a one-off proof of concept and not a permanent part of the data.gov contract.

Policy Issues and Necessary Clearances

When thinking through using crowdsourcing to clean previously non-public government information some initial flags were raised:

- Whether the government may use crowdsourcing;
- Which steps the government must follow to use volunteers;
- What non-public information the government is able to release; and
- How to ensure data cleaned by external volunteers met the Information Quality Act Compliance.

Using Crowdsourcing in the Government

The White House Office of Management and Budget (OMB) published a Technology Neutrality memo in January 2011 stating that, "...agencies should analyze alternatives that include proprietary, open source, and

mixed source technologies. This allows the Government to pursue the best strategy to meet its particular needs."

Even before the OMB memo was published, other Agencies were utilizing crowdsourcing. For example, since 1999, the U.S. Geological Survey (USGS) Earthquake Hazards Program has used the "Did You Feel It?" internet questionnaire to collect information about shaking intensity and damage from the Crowd. This qualitative crisis information from the public turned into quantitative metrics that fed into the other USGS earthquake products for emergency response purposes (Wald et al. 2011).⁴

Similarly, in 2010, the Federal Communications Commission (FCC) used crowdsourcing to help populate the National Broadband Map⁵. The FCC provided the public with a mobile application to test and report their speeds, which were then used to populate the broadband map.

Finally, at the same time that USAID launched this project, the U.S. Department of State's Humanitarian Information Unit launched an experiment to map roads and footpaths in 10 refugee camps that contain a population of over 600,000 people to better support humanitarian response and logistics. As with the USAID effort, they partnered with a well-known VTC – the Humanitarian OpenStreetmap Team – and the general public who spent 48 hours tracing satellite imagery to generate the maps. This short list is by no means exhaustive but illustrates the point that these new methods have already made an important contribution to the U.S. Government.

Free labor

It is within USAID's purview to accept services without offering compensation if they are other than those performed by a U.S. Government employee as part of his or her scope of work. Assuming that is the case, the Agency could accept gratuitous labor after receipt of a written affirmation from volunteers (prior to their performing the service) that:

- They understand they have no employment relationship with USAID or USG;
- They understand and affirm that they will receive no compensation; and
- They waive any and all claims against the USG with respect to the services being provided.

Because the project was taken on by the USAID team in addition to their regular duties, USAID's Development Credit Authority did not have the time or resources to go through 100,000 records for the purpose of geocoding the data. In order to use volunteer labor, USAID included the language above in the crowdsourcing application that every volunteer checked off prior to seeing any data.

Non-Disclosure Act Compliance

USAID's Development Credit Authority has partnerships with private financial institutions in developing countries. Due to the Non-Disclosure Act, the U.S. Government is not legally allowed to release private financial data of these partners. Therefore USAID deleted all private and strategic information prior to releasing the data. More specifically, USAID deleted columns including bank names, borrower names, borrower business names, borrower business asset size, interest rates charged to the borrowers, purpose of the loan, fees charged to the banks, and whether or not each individual borrower defaulted on his/her loan.

⁴ More information can be found at: <http://earthquake.usgs.gov/research/dyfi> [last accessed June 26, 2012].

⁵ More information can be found at: <http://www.broadbandmap.gov/> [last accessed June 26, 2012].

Items remaining in the dataset included the location of each transaction at the state level, and where possible at the city level; the sector of each loan; the amount of each loan in U.S. dollars; the gender of the borrower; whether the loan went to a first-time borrower; the currency of the loan since USAID guarantees both local currency and U.S. dollars; and which records were geo-tagged by the crowd.

Releasing Publicly Identifiable Information

For privacy reasons, USAID wanted to ensure that a business supported by a DCA guarantee could not be identified based on the data USAID released. Therefore prior to the crowdsourcing event, USAID partnered with the DoD to remove all exact addresses from the records. This was achieved by replacing all numeric data with a pound symbol (“#”) throughout the database. Concern remained, however, that in some rural areas of certain countries even a single street name could be used to identify the one business on that street. Therefore USAID decided to take additional precautions.

First, all non-location columns were deleted from the Crowd’s dataset so there would not be access to any additional information about each client. Then USAID took the additional precaution of not disclosing what the data represented to the crowd. Instead of telling volunteers the data represented loans to businesses, they were told that they were geocoding “certain USAID economic growth data.” That way even if a business was identified due to a street that only had one business in a rural area, volunteers would not know anything about the USAID project in which the business was involved. After the crowdsourcing event, the non-location columns were merged back into the dataset.

Next, in the final dataset released to the public, all specific addresses were removed, such as street names and street numbers, and USAID only released place names such as towns or villages where possible associated with each record. This is so the public would not be able to identify a specific business that benefited from a guarantee without the borrower’s and bank’s permission.

Finally, USAID was initially planning on releasing the original location field without numbers to the public in case anyone wanted to perform his or her own quality control/quality assurance on the records the crowd completed. To fully protect the clients, however, USAID ultimately deleted the original location field from the released dataset and instead released only the parsed out place name (i.e. nearest town, village, city, administrative unit, etc. to the address of the loan) and Admin1 name (i.e. state).

Information Quality Act Compliance

Federal agencies are required to comply with the Information Quality Act (Section 515 of P.L. 106-554) by maximizing the quality, objectivity, utility, and integrity of the information they disseminate. USAID ensured that the plan to use volunteers to improve the data using the data.gov platform would comply with the Information Quality Act. During the crowdsourcing project, the data being worked on was visible only to the volunteers and was not publicly disseminated. USAID worked within the confines of the Information Quality Act with the data.gov Program Management Office (PMO) in the U.S. General Services Administration (GSA). In the end, it was determined that there was not an Information Quality Act prohibition on using volunteers to reformat the data used in the project as long as USAID was able to certify that the resulting information complies with the Information Quality Act and internal USAID procedures for assuring the quality of information disseminated to the public.

Workflow

The Crowd's Task

As mentioned previously, the DCA database is structured in such a way that all geographic information is stored in a single field (labeled “City/Region” in the original database and later changed to “Original Location” for processing) and not standardized across all records (see example below). Sometimes the city is given, sometimes a street address, and sometimes only the first administrative (or “state”) level is provided. To resolve this problem that had arisen from manual data entry, the information had to be broken out into different fields to be mapped at the lowest level of granularity across all records. Once parsed out, the place name would be used to capture the first administrative level unit by using a gazetteer such as GNS.

Because a DCA record often contains an Admin1 name and a city/town name within the Original Location field, it was deemed feasible to develop an automated process that used a computer script to parse out the Admin1 name and/or place names and validate them against an authoritative database. The script first looked for matches for place names and Admin1 names against the NGA database. If no match was found, the text of Original Location was entered into the Google Geocoding API to see if it would return an Admin1 name that was valid in the GNS database. The roughly 10,000 remaining records – representing the most complex and/or partial data would require human processing by way of the Crowd.

Fig. 1 The “original location” column would be used to fill in the preceding columns and the status updated accordingly:

Status	Country	Original Location	Admin 1	Admin 1 Code	place name
Assigned	Vietnam	Mac Thi Buoï Ward, Wi Kan Gam Dist Ha Noi Viet Nam			
Assigned	Haiti	Port au prince			
Assigned	Haiti	Sud			
Assigned	Paraguay	### calle, campo ####			

Would then become...

Status	Country	Original Location	Admin 1	Admin 1 Code	place name
Completed	Vietnam	Mac Thi Buoï Ward, Wi Kan Gam Dist Ha Noi Viet Nam	Ha Noi	VM44	Ha Noi
Completed	Haiti	Port au prince	Ouest	HA11	Port au prince
Completed	Haiti	Sud	Sud	HA12	
Bad Data	Paraguay	### calle, campo ####			

The Crowd's task was to mine the data for clues to the appropriate Admin1 and "place name" or name of a populated area or feature that would allow a volunteer to determine the Admin1. Volunteers, therefore, would be given the country name and the "Original Location" (known in the original DCA database as "City/Region") with the task of deriving the Admin1 name, the Admin1 Code (based on international standards to eliminate problems of transliteration between disparate language), and place name if possible. Because of incomplete data, not all records could be processed and it was important to allow volunteers to flag such records as "bad data." This process can be seen in figure 1.

Assembling the Crowd

Reaching out to Volunteer Technical Communities (VTCs)

Because the primary task of the project was to mine geographic information and prepare the data to be mapped, USAID partnered with VTCs known for their capacity in this domain. They included:

The Standby Task Force (SBTF): <http://blog.standbytaskforce.com>

Launched in 2010, the SBTF has roots in the ad-hoc groups of tech-savvy volunteers who had begun to engage the humanitarian sector around mapping, information management, and other technical challenges. The goal of the SBTF is to harness the power of ad-hoc volunteers "into a flexible, trained and prepared network ready to deploy." The main objective of SBTF, and its 855 members, is to assist "affected communities through co-operation with local and international responders." To this end, capacity building for SBTF volunteers is paramount and supported by dialogue and coordination with other tech and crisis mapping volunteer efforts. SBTF members sign a code of conduct based on best practices in the field, including the Code of Conduct of the International Red Cross and Red Crescent Movement and NGOs in Disaster Relief and the United Nation's Office for the Coordination of Humanitarian Affairs (OCHA) Principles of Humanitarian Information Management and Exchange.

One of the USAID team leads on this project, Shadrock Roberts, is an experienced member of SBTF who has participated in previous deployments. Shadrock therefore had built trust with several key points of contact and understood both the culture of the organization and its methods. Questions of motivation and trust can figure prominently in discussions between large public sector, humanitarian, or development agencies and VTCs especially when processing or collecting sensitive data. In this instance a common trust had already been well established. Using this prior contact as a point of departure allowed both groups to focus all of their energy on achieving the best results possible.

GISCorps: <http://giscorps.org>

Founded in 2003, GISCorps grew out of The Urban and Regional Information Systems Association (URISA: a nonprofit association of professionals using Geographic Information Systems or "GIS") and other information technologies to solve challenges in government agencies and departments. GISCorps coordinates short-term, volunteer based GIS services to underprivileged communities by deploying any number of its 2,672 members. GISCorps specifically aims to help improve the quality of life by engaging in projects that support humanitarian relief and encourage/foster economic development. GISCorps members sign a code of conduct that fully incorporates URISA's GIS Code of Ethics adding specific obligations for Volunteers, Project Sponsors, Donors, and GISCorps Administrators.

Drafting the Scope of Work to Deploy VTCs

Both partner VTCs have gone to great lengths to streamline requests for their services in the most professional manner possible. An online “request for activation” form can be found on both organizations’ web sites. Requests are generally reviewed and responses generated within 24 hours. For this project the response from both organizations came within minutes. The request process is the important first step in defining the scope of work (SoW) for the project. The SoW is important for three primary reasons:

- Volunteer coordinators need to understand, precisely, the demands of the project to allocate appropriate resources and budget their management time.
- Well-defined tasks are more achievable than vague, partial notions: both SBTF and GISCorps place an emphasis on after-action review to learn from the project and to better prepare for the next.
- The above points mitigate volunteer frustration and provide a more rewarding experience.

For this project, USAID drafted a two-page, online document that explained, as precisely as possible, the task at hand. This document, along with the request for activation, became the starting point for a series of e-mails and calls between VTCs and USAID to further refine the SoW and collaborate on how best to engage the public.

Marketing the Crowdsourcing Event to Potential Volunteers

USAID framed the message around this being a first-of-its-kind opportunity to engage with the Agency on its pilot crowdsourcing event to geo-tag and release economic growth data. While USAID utilized listservs and social media to publicize the event, half of volunteers came from established volunteer communities interested in geographic information and data. By partnering with the Standby Task Force and GIS Corps, the Agency had an automatic pool of thousands of internationally-based and interested volunteers eager to work on international development data with the government.

The primary webpage USAID used to talk about the event was a Facebook event page. The page can still be accessed at www.facebook.com/events/395012370542862/. One hundred ninety-one individuals signed up for the event on Facebook. This forum provided a platform to send volunteers quick informal updates about the project. During the week of the event, the USAID DCA Facebook page reached 4,200 people. The page had a 15 percent increase in “likes” in the two months preceding the event, increasing from 522 to 599. USAID also established an open data listserv so people could sign up to receive updates about the event.

To engage more people, USAID sent Twitter updates about the event using the hashtag #USAIDCrowd. The hashtag was widely used by the volunteers, interested observers, and other U.S. Government agencies and officials. Two months before the event @USAID_Credit had 830 followers and by June 1 (the starting point for the event) it had surpassed 1000 for the first time, a 20 percent increase.

In order to inform people beyond social media, USAID sent out a press release about the event, which can be found in the appendix of this report. USAID also presented and disseminated information about the event through USAID’s University Engagement group. USAID invited other government agencies to participate by presenting the project at a White House Open Data working group meeting. Finally, a crowdsourcing blog post was published on USAID’s Impact Blog to call attention to the event.

USAID’s partners were instrumental in getting the word out by blogging and tweeting to their followers, putting out their own press releases, and mobilizing their volunteers through their own listservs.

Implementing the Crowdsourcing Event

The event was organized in four stages, each having its own unique characteristics in terms of the partners involved, data quality assurance and quality control (QA/QC) methods, technical challenges, and outputs. The overall workflow was designed to capitalize on each partners' strengths to achieve the best possible outcome. The stages, described in detail below, were pre-event, Phase 1 (data mining and cleaning using the Crowd), Phase 2 (data cleanup and mapping), and Phase 3 (independent accuracy assessment of Phase 1).

Pre-Event

Partners involved: DoD, Standby Task Force, GISCorps, Data.gov, Socrata, Esri.

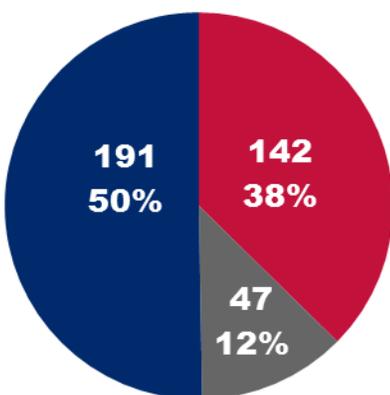
This stage was entirely focused on planning. For the VTCs this meant refining the SoW, preparing resources, and test-runs of the workflow (discussed in Phase 1). Both VTCs spent considerable time communicating the established tasks to volunteers. This included creating detailed instructions for volunteers, establishing means of communication, scheduling, etc. For SBTF this included ensuring continuous, around-the-clock management of a globally distributed volunteer network. USAID was closely involved in this effort, continually refining technical documents for volunteer use, coordinating marketing strategies, and replicating the use of SBTF communications strategies, such as using google documents and skype chat channels, for the general public.

One important aspect of the pre-event planning was asking volunteers to indicate their participation. Volunteer coordinators used this information to allocate appropriate management resources while USAID used the figures to gauge the likelihood of completion. Both SBTF and GISCorps kept their own internal "sign-up" lists for this purpose. USAID used DCA's Facebook web page.

Fig. 2: Anticipated Volunteer Participation by Affiliation

"Public" data comes from the DCA Facebook page, which likely included volunteers from each of the other groups.

■ Standby Task Force ■ GISCorps ■ Public



It was during this time that volunteer management decisions were made regarding Phase 1. Because maintaining crowd motivation and input was critical for the envisioned three-day period, there was considerable discussion about how best to handle a "crowd" of two VTCs and a then unknown number of volunteers from the general public. Phases 2 and 3 were less of a concern since they consisted of a small, self-directed team. It was initially decided that SBTF volunteer coordinators would focus primarily on managing SBTF volunteers while USAID would manage GIS Corps staff and non-affiliated members from the general public. However, SBTF volunteer coordinators eventually went on to assist in the management of all volunteers in Phase 1.

During this time USAID also worked closely with data.gov, Socrata, and Esri regarding the web applications that were developed for the project. Socrata, the contractor for data.gov, undertook the design of a custom application that allowed for the use of data.gov as a platform for tabular data editing and generation. This considerably extended the capabilities of data.gov which previously solely acted as a platform for data *viewing*. The Socrata

application allowed users to check out up to ten individual records from the database at a time for processing. Using data.gov's spreadsheet, the application captured the volunteers' e-mail and time of check-out, and presented the user with the necessary fields for filling in Admin1 names, codes, and place names. The application further allowed users to flag "bad data": meaning that the geographic information provided was simply not good enough to permit proper geocoding of the record. USAID worked closely with the DoD regarding technical issues such as the necessary elimination of sensitive data discussed previously. The DoD also performed data cleanup that was necessary to ensure consistency within the dataset and all instances of certain text occurrences (e.g. "P-A-P," "Port au Prince," "Port-Au-Prince," etc.) were standardized.

USAID initially wanted volunteers to use GNS as the primary tool for searching text within the original geographic information and establishing a first administrative unit match. However, initial user testing found that the user interface for this database was problematic because it did not return the Admin1 code alongside the Admin1 name in search results. In general, volunteers did not find it to be as user friendly or extensive as other online tools. With this in mind, USAID also partnered with Esri, a mapping software company, to develop a custom web map application on Esri's ArcGIS Online platform that allowed users to easily and quickly find administrative names and codes with the click of a mouse. The properties and capability of this geocoding tool can be found at: www.arcgis.com/home/item.html?id=991a730ac41248428b48584ccf77b583.

Phase 1: Crowdsourcing

Partners involved: SBTF, GISCorps, Socrata, Esri.

Phase 1 was the most visible stage of the event. To coordinate and manage volunteers, USAID adopted the SBTF model including:

- A publicly available, online, Google document that detailed instructions and included screenshots of the applications and a log for frequently asked questions.
- A dedicated chat room using the freely available commercial software Skype. The chat room acted as a "call center" where volunteers could receive real-time instructions, advice, and ask questions. This is a highly social environment and a great number of volunteers used it: SBTF reports that 85 percent of their volunteers actively used the Skype chat room. The chat room becomes a space for sharing information - especially when certain volunteers have regional expertise - and relieving tension by interacting with other volunteers.

Moreover, USAID and the VTCs actively promoted the event and kept volunteers motivated with updates via the use of social networking tools (e.g. Twitter and Facebook) and regular e-mails. As a result of careful planning, Phase 1, which was scheduled to take place over a period of 60 hours (from noon on June 1 until midnight on June 3) was completed in roughly 16 hours with most records having been completed by 3 a.m. (Eastern Daylight Time) at which time the application crashed. When the application came back online, it took only another hour to complete all records.

In all, 145 volunteers took part in geocoding at least one record. While more had signed up to participate, because the event finished so early many volunteers never had the chance to clean records.

Phase: 1 Quality Assurance/Quality Control (QA/QC)

To participate, volunteers had to register an account on data.gov, which was then linked to each record they geocoded. By linking records to volunteers at the individual level, USAID staff members were able to perform “spot checks” during the crowdsourcing event to look for anomalies in how the Crowd was entering data. If it was determined that any individual volunteer was incorrectly - whether purposefully or not - entering bad data, that volunteer could be contacted directly or their records could be redacted from the final product. It should be noted that at no time did USAID staff detect any suspicious activity. There were some initial mistakes made by volunteers that were rectified by communicating, en masse, the problem via the volunteer coordinators.

Phase 2: Data Processing and Mapping

Partners involved: DoD and GISCorps

This phase was largely designed to adjust for any problems in Phase 1 and to begin mapping the data. A small number of records (69) remained “assigned” but had not been completed. This is likely due to a bug in the application or problems while the application crashed. USAID had initially worked with GISCorps to ensure that a small team of volunteers was available in the event that all records were not completed during the two and half day Phase 1. This team was, instead, activated to complete the 69 remaining records. Once those records were finished, they were delivered to DoD staff who played an important role by populating duplicate records based on the “parent” record that was given to the crowd. This essentially involved identifying multiple records with duplicate data information in the “original location” field. These records were then given a unique identifier and only one of them was given to the Crowd. The processed records would then be used to populate the necessary information for the duplicates.

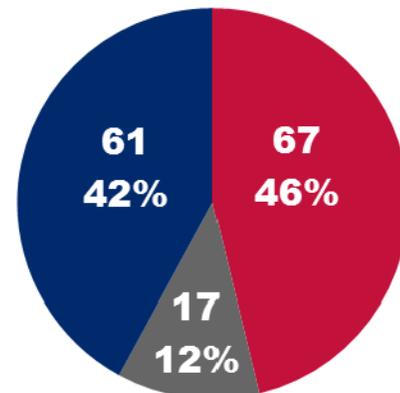
It was also during this stage that initial data was processed concerning crowd performance. Records were sorted by volunteer, and then volunteer affiliation, to compare the intensity with which each group performed in Phase 1.

The volunteer data confirm an oft-cited axiom for crowdsourcing that a small portion of the crowd is generally responsible for a disproportionate share of the work. In figure 4 we see that this is confirmed across all groups. However, it is interesting to note some differences between them. While the general public had the volunteer with the single largest number of records processed (89) there is a steep decline and a sharp curve in the distribution meaning that there is greater variability between high-producing and lower-producing volunteers. The GISCorps, on the other hand, had many fewer volunteers (participating in Phase 1), yet there is a more gradual slope in their output versus a steep curve indicating less extreme

Figure 3: Phase 1 Volunteer Participation by Affiliation

Data are for Phase 1 only and do not reflect GISCorps volunteers for Phases 2 and 3, nor volunteers who had signed up to work but could not due to early completion.

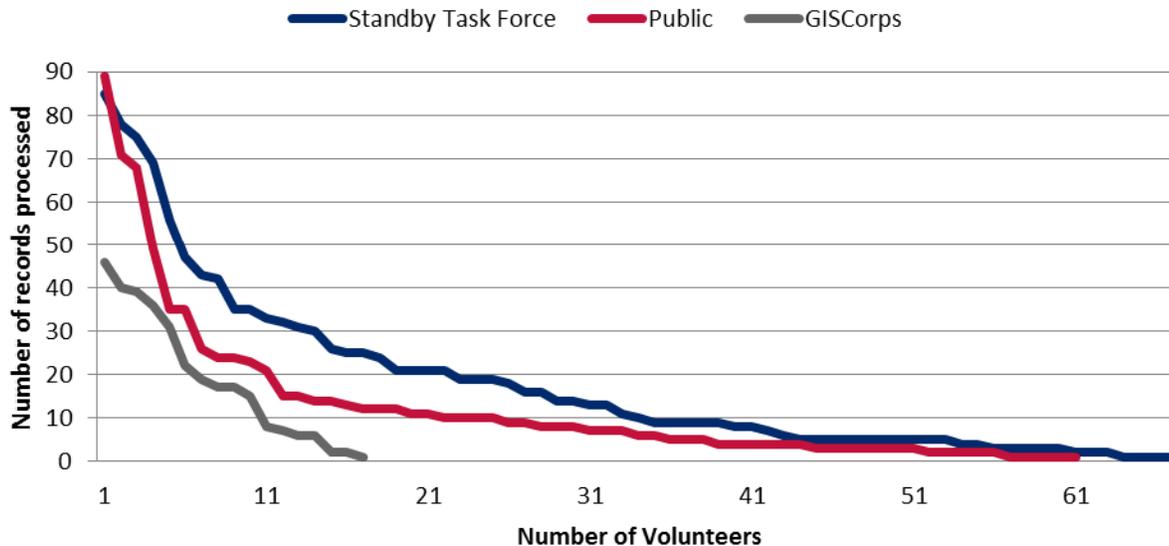
■ Standby Task Force ■ GISCorps ■ Public



variability between volunteers. Finally, SBTF had the largest overall number of volunteers (67) and – while there is variability between high-producing and lower-producing volunteers – shows the consistently highest output. While every single volunteer made an important contribution to the greater sum of the project, these data suggest the importance of incorporating VTCs with the appropriate expertise that translates into, generally, more consistent output due to a greater understanding of both the tools and the task at hand.

Figure 4: Number of Records Processed by Volunteer and Affiliation

It is possible that some GISCorps or SBTF volunteers were counted as “public” volunteers since these metrics were taken from VTC volunteer e-mail lists and volunteers may have used alternative e-mail addresses.



Phase 2: QA/QC

Both USAID and the DoD reviewed the data to look for any anomalies or patterns that might indicate systematic error. This included an automated process that checked Admin1 codes in each record against the country designation in the record to ensure that all reported administrative units were indeed located in that country. USAID staff found 66 records that had not been completed correctly but the error was largely due to slight deviations in transcriptions when Admin1 codes were entered. These records were easily corrected. In all, 2,312 records were processed by the crowd, of which 480 (20 percent) were labeled as “bad data” and could not be mapped below the national level.

During phase 2 preliminary QA/QC was performed by comparing a set of 400 records that were processed using both crowd and automated processing. Of the 400 records used, the crowd labeled 12 of them as “bad data.” When comparing the remaining 388 there was agreement in 61 percent of the records for the administrative code (237 agreed, 151 disagreed) and 49 percent for the name of an administrative unit. The difference in agreement between Admin1 codes and names is likely due to small differences in input, including diacritical marks for pronunciation. It was for this reason that Admin1 codes were used as the basis for mapping. At first this finding was confusing because it did not communicate as much as a more polarized finding (e.g. only 10 percent agreement or 90 percent agreement) might have, however, it would later confirm a greater than expected discrepancy between the accuracy of the automated and crowdsourced data. It also highlights the very subjective nature of the process and underscores the need for multiple methods for assessing the quality of the data.

Phase 3: Accuracy Assessment

Partners involved: GISCorps

To better understand the limitations of the data provided at the sub-national level, USAID asked the GISCorps to perform an independent accuracy assessment of the data. As geographic data is increasingly produced by, collected from, or processed by ‘non-experts’ the question of assessing the accuracy of these data has become a focus in scientific literature (Goodchild and Li, 2012; Haklay, 2010; , Elwood, 2008). Following Congalton (2004), accuracy assessment is important for the following reasons:

- It provides a method of self-evaluation to learn from mistakes;
- It allows for comparison of the two geocoding methods quantitatively; and
- It better positions the final products to be used in decision-making processes where understanding the accuracy of the data is critical.

It is important to judge the relative accuracy of each data set independent of the other because volunteers had records containing much less, or much more difficult, geographic information than was available for the automated process.

Phase 3 Design

Phase 3 volunteers were tasked with creating a Quality Control (QC) dataset of high-quality geolocated records with which to do an accuracy assessment of the automated and crowdsourcing methods of geolocation. A random sample of records was drawn from both datasets; 382 records were drawn from the automated database, and 322 records were drawn from the Crowdsourcing database. These sample sizes were chosen to ensure that sample estimates would correctly represent population metrics.

The 17 phase 3 participants were selected from among highly-experienced GIS professionals in GISCorps; participants had an average of eight years of GIS experience. In addition to professional experience, participants were chosen who had experience in this specific geolocating process. Of these participants, 13 had taken part in previous phases. In addition, participants were preferentially assigned records for countries in which they had personal experience, or spoke the language of the country. Participants were instructed to geolocate records with the greatest possible care, since their results were to be considered true and accurate. Phase 3 participants used the same geolocating resources as were used for Phases 1 and 2. Participants were not exposed to the earlier automated or crowdsourced results for geolocated records, so as to not bias their determinations. Participants were asked to quantify the difficulty and certainty of their determinations based on a 1 to 5 point scale. For example, a difficulty rank of 1 indicated that correctly spelled city/town name and Admin1 name were present in “Original Location” data, while a difficulty rank of 5 indicated that neither city/town name or Admin1 name were present and had to be inferred. A certainty rank of 5 indicated that the volunteer was completely sure of the Admin1 assignment, while a certainty rank of 1 indicated that the assigned Admin1 name was a best guess.

Phase 3 Results

Accuracy of results was calculated by comparing the resulting Admin1 Codes with the previously determined Admin1 Codes. The Codes were used rather than the Admin1 Names, because there is some variation in the spelling of Admin1 Names among the three geolocation resources. The Automated method was found to be 64 percent accurate, while the Crowdsourcing method was found to be 85 percent accurate.

Automated Method Details

Of the 382 records in the QC dataset for the automated method, 136 were in disagreement with the automated method results. The median certainty rating of records in the QC database (the degree to which volunteers were sure of their assignments) was 5: the highest rating of certainty. It is therefore highly certain that the automated method results were inaccurate for these records. The median difficulty ranking of records in the QC dataset was 2, which indicates that the “Original Location” field contained a valid Admin1 name or City/Town name, but that these valid values may have been difficult to parse out from among a long string of data. There were two records where the automated method accomplished a geolocation, but our experts were not able to do so.

These results suggest that the automated method script might be re-evaluated and improved by examination of the 137 records where invalid assignments were made. Many of the invalid assignment records contained a complex series of words in the “Original Location” field and quite sophisticated logic might be needed to find the correct keywords for deciphering this location. In other cases the “Original Location” was not as complex, but the Automated method was too simplistic in its evaluation; for example for the “Original Location” of: “# DE JUNIO Y CALDERON ANTIGUA BAHIA”, the Automated method recognized the word “Bahia” as a valid Admin1 Name, while the expert discovered that Antigua Bahia is the name of a neighborhood in the city of Guayaquil in the Canton Guayaquil Admin1 unit.

Crowdsourcing Method Details

Of the 322 records in the QC dataset for the crowdsourcing method, 46 were in disagreement with the crowdsourcing results. The median certainty rating of records in the QC database (the degree to which volunteers were sure of their assignments) was 4 (the second-highest rating of certainty), so the experts were only slightly less certain of their designations than they were for the automated method dataset. This is to be expected, since these records were more complex to evaluate (as suggested by the fact that the automated method was unable to find matches). Surprisingly, however, the median difficulty ranking of records in the QC dataset was 2, the same as for the automated records, which indicates that the “Original Location” field contained a valid Admin1 name or City/Town name, but that it takes some degree of sophistication by the experts to find these correct key words.

There were 63 records in the crowdsourcing QC dataset which the experts were not able to geolocate. Of these, 27 were geolocated by the crowd, which suggests that the phase 1 participants, in their zeal for success or inexperience, might have produced a result where one was not warranted. This may indicate that for best results, expert volunteers are needed.

Of the 46 inaccurate records, 15 mismatches were due simply to transcription errors; for example, where an Admin1 Code of “11” was typed instead of the correct code of “TZ11” (the country code was omitted). These errors are quite easy to fix by visual inspection of the database. After correction of these errors, the accuracy rate of the crowdsourcing method improved from 85 percent to 90 percent.

Phase 3 Summary

The high accuracy rate for crowdsourcing method is a promising indicator of the quality of crowdsourced data, especially when experienced professional volunteers are recruited. The smaller accuracy rate for the automated method suggests that sophisticated algorithms need to be developed to impart a higher degree

of intelligence to the computer – one way to develop this machine intelligence is through a QC check such as that done here where mismatches can be examined to capture the human thought process.

Following spot-checks during Phase 1 and the completed accuracy assessment in Phase 3, it was determined that, overall, the crowd performed very well with a high degree of reliability and only a small number of records were corrected.

Published Maps and Data

Determining what to map

As noted earlier, the goal of the project was to achieve a greater resolution than the national scale. It was determined that Admin1 would be the minimum mapping unit but that, where possible, “place name” level data would be provided. This means that the dataset works at three geographic scales: national, first-administrative unit, and place name. The data are complete at the national level but become progressively less so at lower levels. After final processing, the first administrative unit was identified for 74,002 records, or 63 percent of the final dataset. Since some records are available at a finer or coarser geographic resolution, both “Place Name” and “Admin1 columns” are included in the released dataset.

Adopting IATI and Open Data Standards

The International Aid Transparency Initiative (IATI) aims to make information about foreign assistance spending easier to find, use and compare (IATI, 2012). To this end, there exists a set of geographic precision codes that can be used with point data⁶. These codes are a valuable international standard that allows data to be compared across entities such as the World Bank, whose “Mapping for Results” uses this standard (World Bank, 2012). Using a fixed point, however, rather than an administrative area polygon, presents certain geographic challenges such as not accurately capturing the geographic extent of an activity.

While the web-maps that USAID created for public viewing display all data within aggregated administrative polygons, we have included centroid point data for all records in the transactions data set available on data.gov and denoted each with the most appropriate IATI precision code.

Earlier this year, the White House published a Roadmap on Digital Government⁷ which emphasized that federal agencies are to “fundamentally shift how they think about digital information.” Rather than focus on final products, the U.S. Government should focus on providing data through a web Application Programming Interface (API)⁸ in order to “make data assets freely available for use within agencies, between agencies, in the private sector, or by citizens.” To comply with this information-centric approach, all data published through this project has been made available on data.gov in multiple formats and with enabled APIs.

Geographic and Licensing Issues of Using Admin1

The difficulty with maintaining a global database of internal boundaries is that a) these boundaries are the purview of individual countries and b) these boundaries can and do change. Boundaries or names used are

⁶ More information available at http://iatistandard.org/codelists/geographical_precision [last accessed June 26, 2012].

⁷ More information available at: <http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html> [last accessed June 18, 2012].

⁸ Web APIs are a system of machine-to-machine interaction over a network. Web APIs involve the transfer of data, but not a user interface.

not necessarily authoritative. While several open administrative boundary sets are available online, such as the United Nation’s Second Administrative Level Boundaries, or the United Nation’s Food and Agriculture Organization’s Global Administrative Unit Layers, they are often incomplete and not regularly updated.

The DCA map uses a U.S. Government created global data set of Admin1 units that contains both open and commercial purchased products that are protected by license. While the results derived from these boundaries are public, USAID cannot share the commercially protected shapefiles that contain each boundary’s geometry.

The combination of Admin1 codes and names in the open data set can be used with open or commercial boundary sets acquired by the user to create new maps. Additionally, by providing centroid locations for all records and adopting the IAIT precision codes, users can alternatively choose to view the data as point locations instead of administrative units.

Summary and Lessons Learned

Prior to this event, the DCA database could only be mapped at the national level despite the existence of a very large amount of additional geographic data. While the entire data set can still be mapped at the national level with an accuracy of 100 percent, value has been added to the data set by automated geocoding processes that refined 69,038 records at 64 percent accuracy while crowdsourcing processes refined an additional 9,616 records at 85 percent accuracy, detailed in the following table.

Processing Method	Records Processed	Records Mapped at Admin1	Accuracy at Admin1
Automated	107,392	66,917	64%
Crowdsourcing	9,607	7,085	85%
Total	116,999	74,002	

This provides the public with some options for using the data set at a finer geographic scale. Moreover, the process itself broke new ground by engaging the public, for the first time, in processing to map and open USAID data. The project attracted the attention of more than 300 volunteers worldwide, 145 of whom far exceeded all expectation by finishing their portion of the processing in roughly 16 hours: less than one-third of the time anticipated for this task. The additional 155 volunteers who had signed up to help on Saturday and Sunday logged in to find the project had completed before they were able to geocode any records.

DCA increased Twitter followers by 20 percent and Facebook friends by 15 percent during the preparation and launch of the crowdsourcing event. Moreover, the project created a strong relationship with two vibrant VTCs and was completed without any public expenditure. As is true with any innovation, this project was a learning experience. Listed below are improvements and recommendations for any government, development, or humanitarian agency that would like to pursue an exciting new path to data processing and public engagement.

Policy Issues

Initially USAID was going to message the event around the impact of the data. However, in order to protect the borrowers’ personal information, USAID delayed disclosing details concerning the exact nature of the data to the crowd until after the data-processing was closed. There was concern that this would have a negative impact on the amount of volunteers and, indeed, some volunteers would have preferred to more

clearly understand both the nature of the data and the final intent. While USAID was ultimately able to garner sufficient interest and participation for the event even with the more generic messaging, it is preferable that there be full public disclosure about the data prior to any crowdsourcing event. Ultimately in this case, the opportunity for the public to engage with their government in a new way and make a difference in international development mattered more to people than the exact nature of the data.

Recommendations:

- Every attempt should be made to disclose the nature of the data that volunteers are working on and ensure that they understand the purpose of the project. If certain information cannot be disclosed, define these parameters at the beginning of the project. When possible a forum for volunteer questions to be answered should be provided to engage them in the project.

Reach-back to Crowd

Crowdsourcing often requires performance by a large group of individuals who are working remotely without any direct contact with the project convener. VTCs should be viewed no differently than any business partner: all parties are working toward a shared goal with limited resources. In both cases communication is paramount. Interacting with volunteer coordinators preceding and during the crowdsourcing event required a significant amount of time. This was time well spent as it involved continually refining workflow and communications and preparing trial runs of the workflow and applications. Any crowdsourcing project should include adequate time for this critical interaction. VTC coordinators must fully understand the task, workflow, and potential pitfalls that volunteers can encounter to best assist them during the project. Greater time spent preparing will directly maximize the efficiency of the volunteer's time.

Recommendations:

- The Crowd is a resource and crowdsourcing should be understood as a project – like any other – that requires both adequate time dedicated to management and a considerable amount of communication between partners to ensure a mutually beneficial experience and positive outcomes. Any organization planning to engage with crowdsourcing or VTCs regularly should build this management capacity into their organization. While the entire event took place at no additional cost to USAID, the Agency did “spend” the time that three of its employees dedicated to the project.
- People volunteer to make a difference but they also volunteer to connect with other people. The social elements of the event- both the chance to crowdsource live from USAID and joining a chat room to converse with others- fulfilled this important role. Many people who joined the crowd at USAID said they were volunteering to meet other like-minded people. The social element of a crowdsourcing event creates community that lasts beyond the immediate event.

Operationalizing a Crowdsourcing Event

A simplified set of instructions would have enhanced the crowdsourcing event, and increased the likelihood that volunteers read the instructions fully. Additionally, a short online video showing volunteers the workflow and helping them understand what they were supposed to do would have been helpful. Nevertheless, having run through two trial runs and gaining volunteer leads to help run the Skype channels

proved invaluable. Allowing volunteers to be able to choose which country they wanted to work on would enable individuals with a familiarity with a certain country to better complete those tasks.

The event would have also benefited from having volunteers use one gazetteer rather than searching among a wide range of online gazetteers. This would have helped to standardize the updated database and released data. Taking this a step further, ideally USAID, Socrata, and Esri would have linked the data.gov dataset to the Esri USAID crowdsourcing application so when a volunteer identified a location, with one click the Admin1 name and code could be filled in within the dataset.

Finally, the Socrata application became overwhelmed with traffic at several points, causing some volunteer frustration. Although many volunteers are used to working in technically challenging conditions, every effort should be made to mitigate technical problems. In this case, Socrata's volunteer support was excellent, but was operating on a time-zone different from some volunteers and could not provide around-the-clock coverage. Aside from load-related issues, the application also had bugs – most notably for the registration process – that had to be addressed on more than one occasion.

Recommendations:

- Agencies undertaking a crowdsourcing event should work closely with volunteer coordinators to provide the most appropriate guidance to volunteers. Several types of media (documents, videos, etc.) should be used to maximize volunteers' time.
- It is essential to have consistent and dedicated support for all technological aspects of such a project. Sufficiently test all applications to ensure that they can support more volunteers than anticipated.
- Follow the crawl, walk, run approach to crowdsourcing. By starting slow and running incrementally larger tests, organizers are able to hone the workflow. In both test runs, USAID was able to refine the workflow, instructions, and applications.

Publishing Data and Maps

This project was necessary because USAID had a dataset with non-standardized location information. The database where the data is stored has been updated to geocoding standards. The location fields are parsed out and the database is linked to an external gazetteer that ensures the standardization of all place names and Admin1 names. Now that the database has been corrected, USAID's DCA will not need to manually geocode any of its records in the future.

This project did not have access to a global administrative data set that is regularly maintained and that could be distributed publicly. Thus, USAID was unable to make the admin1 polygon shapefiles publically available as a service. Another challenge was that the GNS database did not contain information for some countries. This information had to be created individually and documented.

Recommendations:

- All offices should review location data in databases and confirm that they are up to geocoding standards. At the very least, location data should be separated into different fields for different geographic areas (i.e. city versus state).
- More Agencies should take advantage of the section of the data.gov platform that allows for a dataset to be uploaded to an interactive platform where it is enabled with an API.

- Development and humanitarian mapping projects would benefit from greater investment in existing initiatives to create and maintain updated, open, global boundary sets such as the United Nation’s Second Administrative Level Boundaries or the United Nations Food and Agriculture Organization’s Global Administrative Unit Layers⁹.
- Likewise, development and humanitarian mapping projects would benefit from greater investment in the GNS database in terms of content and usability.

Conclusion

Throughout the Obama Administration there has been a commitment to make government more transparent. This pilot USAID project sought to find the most efficient way to make a dataset available and more useable for the public by utilizing existing platforms, new and existing partnerships, and online volunteer communities. Though used by other U.S. Government agencies, this was the first time USAID deployed crowdsourcing to process Agency data and the first time data.gov was used as a crowdsourcing platform anywhere in our government. It is USAID’s hope that our experience has helped blaze a trail to make crowdsourcing a more accessible approach for others. The project has the potential to encourage more agencies to publish more data in a cost-free manner and engage an interested and experienced public directly in U.S. Government work. This “data-as-dialogue” has transformative power not only for data processing, but also building greater awareness of USAID’s mission, goals, and work.

⁹ Web links for these boundary data are listed on page 24.

Works Cited

Congalton, Russell. 2004. Putting the map back in map accuracy assessment. In *Remote Sensing and GIS Accuracy Assessment*, pgs.1-13, eds. Lunetta, R.S., and J.G. Lyon, CRC Press, Boca Raton: FL.

Elwood, Sarah. 2008. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72 (3): 173-183.

Estellés-Arolas, E. & G. Ladrón-de-Guevara, F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science* (forthcoming).

Goodchild, M. and L. Li. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*. 1:110–120.

Howe, J. The Rise of Crowdsourcing. *Wired*, June, 2006. Available online at: <http://www.wired.com/wired/archive/14.06/crowds.html> (last accessed 10 June 2012)

Haklay, Muki. 2010. How Good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37 (4): 682–703.

Wald, D.J, V.Quitoriano, B. Worden, M. Hopper, and J.W. Dewey. 2011. USGS “Did You Feel It?” Internet-based macroseismic intensity maps. *Annals of Geophysics*, 54 (6): 688–707.

Web sites

International Aid Transparency Initiative (IATI), 2012: www.aidtransparency.net/ (last accessed 14 June 2012).

World Bank, Mapping for Results, 2012: maps.worldbank.org/ (last accessed 16 June 2012).

Publicly Available Global Administrative Boundary Data

- United Nation’s Second Administrative Level Boundaries: www.unsalb.org (last accessed 16 June 2012)
- United Nation’s Food and Agriculture Organization’s Global Administrative Unit Layers: www.fao.org/geonetwork/srv/en/metadata.show?id=12691 (last accessed 16 June 2012)
- Global Administrative Areas: www.gadm.org/ (last accessed 16 June 2012)