Crowdsourcing methodology: establishing the Cervid Disease Network and the North American Mosquito Project

Lee W. Cohnstaedt[1*], Darren Snyder[2], Elin Maki[1] and Shawn Schafer[3]

[1] *Arthropod-Borne Animal Diseases Research Unit, Agricultural Research Service, United States Department of Agriculture, 1515 College Avenue, Manhattan, Kansas 66502, USA.*
[2] *Department of Entomology, Kansas State University, 123 W. Waters Hall, Manhattan, Kansas 66506, USA.*
[3] *North American Deer Farmer's Association, USA.*

[*] *Corresponding author at: Arthropod-Borne Animal Diseases Research Unit, Agricultural Research Service, United States Department of Agriculture, 1515 College Avenue, Manhattan, Kansas 66502, USA. Tel.: +1 785. 537.5592, e-mail: Lee.Cohnstaedt@ars.usda.gov.*

Veterinaria Italiana 2016, **52** (3-4), 195-200. doi: 10.12834/VetIt.577.2765.2
Accepted: 01.07.2016 | Available on line: 30.09.2016

IV International Conference on *Bluetongue and Related Orbiviruses*. November 5-7, 2014 - Rome, Italy - Selected papers

Keywords
Citizen science,
Continental scale,
Disease detection and spread,
Geographic distribution,
Mosquito and biting midge collection,
National surveillance,
Sentinel farms.

Summary
Crowdsourcing is obtaining needed services, ideas, or content by soliciting contributions from a large group of people. This new method of acquiring data works well for single reports, but fails when long-term data collection is needed, mainly due to reporting fatigue or failure of repeated sampling by individuals. To establish a crowdsourced collections network researchers must recruit, reward, and retain contributors to the project. These 3 components of crowdsourcing are discussed using the United States Department of Agriculture social networks, the Cervid Disease Network, and the North American Mosquito Project. The North American Mosquito Project is a large network of professional mosquito control districts and public health agencies, which collects mosquito specimens for genetic studies. The Cervid Disease Network is a crowd-sourced disease monitoring system, which uses voluntary sentinel farms or wildlife programs throughout the United States of America to report the onset and severity of diseases in local areas for pathogen surveillance studies.

Fondazione del Cervid Disease Network e del North American Mosquito Project: due esempi sull'utilità del metodo crowdsourcing

Parole chiave
Aziende sentinella,
Citizen science,
Distribuzione geografica,
Raccolta di zanzare e *Culicoides*,
Rilevamento e diffusione delle malattie,
Scala continentale,
Sorveglianza nazionale.

Riassunto
Il *crowdsourcing* è un metodo per ottenere servizi utili, idee o contenuti sollecitando il contributo di un largo numero di persone. Questo è un metodo innovativo per l'acquisizione dei dati, che funziona molto bene per singoli report. Il *crowdsourcing* si rivela poco efficace per le raccolte di dati nell'ambito di progetti a medio e lungo termine, a causa dell'elevato impegno necessario per la comunicazione dei dati. Per stabilire una rete di scambio *crowdsourced*, i ricercatori devono reclutare, ricompensare e mantenere i contatti con i collaboratori del progetto. Il presente studio analizza questi tre aspetti del *crowdsourcing* in relazione a 2 progetti sviluppati utilizzando il social network del Dipartimento dell'Agricoltura degli Stati Uniti: un progetto sulle patologie dei cervidi (Cervid Disease Network) e un progetto sulla presenza di zanzare nell'America del Nord (North American Mosquito Project). Quest'ultimo progetto comprende una larga rete di distretti specialistici per il controllo delle zanzare e di aziende sanitarie che raccolgono campioni di zanzare per studi genetici. La rete Cervid Disease è, invece, un sistema di monitoraggio *crowdsourced* per le malattie dei ruminanti selvatici che si avvale di aziende sentinella o di dati derivanti da progetti sulla fauna selvatica realizzati in tutto il territorio degli Stati Uniti per divulgare dati sull'insorgenza e la gravità delle malattie che si verificano localmente utili per programmare piani di sorveglianza nei confronti dei vari patogeni responsabili.

Introduction

Crowdsourcing is a cost-effective tool to gather data, services, ideas, or content by soliciting contributions from a large group of people and especially from the online community. This may include professionals already conducting the work or citizen scientists (untrained volunteers). Crowdsourcing works because contributors, individuals providing samples or data, are willing to use their time to answer larger questions than individuals can address. Successful crowdsourcing projects must recruit contributors and then carefully maintain their connection during the project lifetime. More difficult tasks required by the project protocol leads to fewer contributors despite initial generosity. Therefore after recruitment, the contributors must be retained through rewards and continued connection to the project via recognition, education, and reciprocal information exchange. For long-term crowd-sourced data acquisition the 3 most important components (recruitment, retention, and reward) must be carefully planned. Networks do not self-assemble and will quickly fall apart if not actively maintained. Furthermore, contributors must never be placed in harm's way or asked to do dangerous tasks, so the crowdsourcing plan must also have explicit safety features to protect contributors. In this article, crowdsourcing methodology will be explained using 2 examples of crowdsourcing networks for long-term data gathering, the North American Mosquito Project (NAMP) and the Cervid Disease Network (CDN). Both networks were designed by the United States Department of Agriculture and focused on studying pathogen vector insects (mosquitoes and biting midges) of medical or veterinary importance.

The North American Mosquito Project was designed to collect mosquito specimens throughout single species' known distribution range in a single mosquito season. This required hundreds of simultaneous collections throughout the continental United States and in other countries. The network used mosquito abatement and control district professionals, public health agencies, and citizen scientists. Later the project evolved into the Invasive Mosquito Project (Thackrah *et al.* 2016). The NAMP project focused on pathogen vector insects, whereas the CDN was created to aggregate various sources of captive cervid data from private farmers and wild cervid disease data from government agencies. The CDN focused on the greatest threats to cervid health: biting midge transmitted Epizootic haemorrhagic disease virus (EHDV), and Bluetongue virus (BTV). The network collected disease onset and severity data of outbreaks on cervid farms and in national and state forests. Both networks carefully recruited contributors based on the requirements of insect monitoring or disease surveillance plans (Cohnstaedt *et al.* 2012). The methods are discussed to recruit, retain, and reward contributors to maintain long-term participation in crowd-sourced data collection network and to insure contributor safety.

Methods

The recruitment plans for the NAMP and CDN networks used clear and simple research questions to communicate the study purpose and clarify to potential contributors why they should donate time and resources to help the projects. After developing a clear concise message, contributors were identified from personal connections, at professional meetings, and Internet searches. Contributors were then contacted using face-to-face meetings where possible, phone calls, and emails. When recruiting contributors, the project purpose and how each contributor might benefit had to be communicated first before discussing the contributor's inputs. After understanding the project contributors often offered to contribute to the project and were willing to offer almost unlimited help. The initial CDN contributors were recruited at the North American Deer Farmer's Association's annual meeting. Later, to get a more even distribution, targeted recruiting via word-of-mouth using selected connections (known figures in the industry) were used to build upon the core group of contributors. Targeted recruiting via email and telephone was used to cover the missing areas. Contributing farmers contacted each other and increased the network size through natural growth.

The retention plan for each network was as important as the recruitment phase because only a finite number of skilled individuals or farms were available in an area to be recruited. Furthermore, losing a contributor was permanent. Therefore the retention plan was carefully calculated to minimize contributor effort and maximize contributor reward. The key aspects of the retention plans were clearly defining the project scope and limiting the expectations of the contributors. Most contributors were happy to be part of a larger process if only a single collection was requested. However, continued data collection required continued correspondence and return of information in a tit-for-tat approach. In addition to returning information, contributors were given yearly updates via email to keep them involved. A yearly presentation at an annual meeting of the associations (the American Mosquito Control Association or the North American Deer Farmer's Association) was also an opportunity to acknowledge the contributors and return information. The presentations helped recruit additional individuals to the networks.

The reward plan is a part of the retention plan but should receive special consideration because hundreds of contributors cannot be contacted or rewarded equally. In general, the most effective reward was recognition for the help. Published papers and presentations were used to acknowledge the contributors' efforts. Additionally, when research papers were not produced, the contributors were emailed updates or progress reports. The contributors benefited from the information that was distributed. For example, the mosquito control associations gained knowledge of their mosquito populations and the deer farmers were informed of the past year's disease outbreaks. In some cases, letters of support and acknowledgement were written to the municipal funding sources for a few public agencies. The letters thanked the agencies for their efforts and participation in national disease surveillance and mosquito monitoring studies.

The safety plan was important to insure contributors were never placed in danger or at risk of harm. In the case of the mosquito collections, protocols insured all mosquitoes were dead and traps were placed during the least likely time of mosquito exposure. This was unnecessary in most cases because the mosquito control professionals were already aware of the hazards, but for the citizen scientists the safety plan was crucial. With the animal diseases for the CDN, the safety plan was needed to safely preserve the biological samples, which degraded quickly if not properly handled. This also insured high quality samples from collection to transport. Furthermore, anonymity of the contributing farms was important to the farmers and this was obtained by masking their locations.

Results

The NAMP network was used to collect preserved mosquito specimens in 2011 and 2012 (Maki and Cohnstaedt 2015) but was shut down in 2013 to reduce burnout and because sufficient mosquito specimens were received in prior years. In 2014 and 2015, selected nodes of the network were activated and asked to collect live mosquito specimens for the next set of NAMP studies. Details of the NAMP contributors, the recruitment plan, the snowball effect during recruitment (i.e. when contributors, not contacted by the network organizer, recruit additional contributors it is called the snowball effect), and the specimens are reported by Maki and Cohnstaedt (Maki and Cohnstaedt 2015). A yearly update is presented at the American Mosquito Control Association meeting and the contributors were emailed the first published paper in 2014. Despite the retention plan, contributor burnout was still a problem, however, single sample geographic contributor distribution was comprehensive throughout the continental United States (Figure 1).

The CDN was established in 2014 and consists of 47 sentinel captive cervid farms in 16 states (Figure 2). Additionally, wildlife reports from state and federal agencies will be added to the database

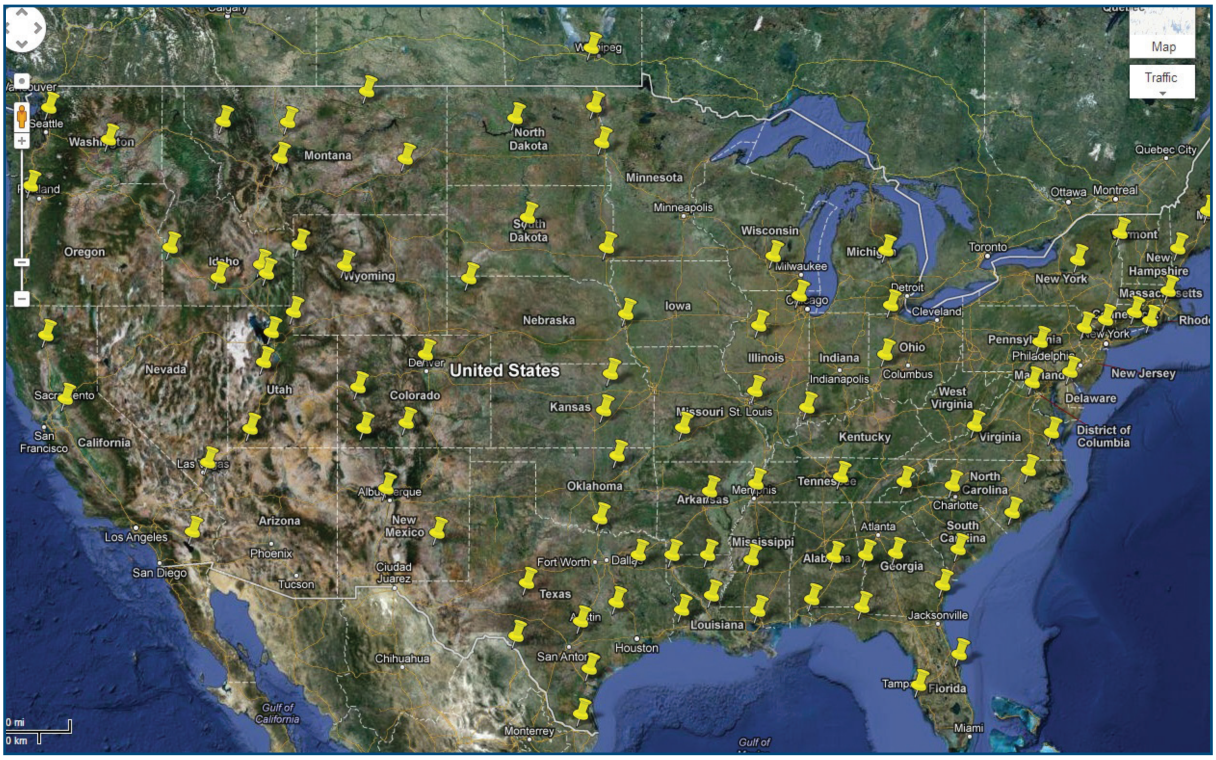

Figure 1. *Crowdsourcing contributor' collection sites for the North American Mosquito Project in 2011 and 2012.*

Figure 2. *Map of current sentinel farms participating in the Cervid Disease Network.* Deer icons denote approximate location of participating members to protect their anonymity, which consists of 47 sentinel farms in 16 states.

to incorporate wild deer disease cases. Members of the cervid community, professionals, industry, and natural resource managers were targeted for recruitment. Contrary to the NAMP network, anonymity was a vital contingency for recruitment of private farms. Most individuals were recruited by word-of-mouth at the annual meeting and during state workshops and association meetings. An update is planned for the next North American Deer Farmer's Association annual meeting for retention and recruitment purposes.

Discussion

Where complex research questions require large data sets, which single individuals cannot obtain due to a lack of personnel or resources, crowdsourcing the data collection is an economical and efficient method of addressing the logistical challenges. Crowdsourcing using the contributions of professional or citizen scientist worked very well for both the North American Mosquito Project and Cervid Disease Network. Specific and detailed contributor recruitment, retention, and reward plans helped to reduce burnout and to maintain the networks for long-term data collections.

Recruitment plan

When recruiting contributors, first address the contributors needs before stating the project's goals. Throughout the conversation clearly identify how the project is mutually beneficial to the contributor and the project. Contact with an individual is not sufficient to insure conversion from contact to contributor. The NAMP network had only a 9% (110 of 960 contacts) conversion rate of contacts to contributors. The method of the contact had the most significant impact on conversion rate. The NAMP demonstrated the importance of personal contact (face-to-face); 29% of contributors were known contacts and personally recruited. Emailing without another form of follow up was nearly useless. Cold calling was not effective or efficient without a personal connection to the contact. If any personal connection can be established the conversion rate increased. For example, in the South-East United States, a single individual was known to all local areas and after the individual endorsed the project, all the others joined. In another case, 1 well-known individual made all the calls to cover another region. Knowing the right people with connections was vital for natural growth during recruitment of contributors. Natural growth of the network each year was needed to maintain the network size. Without a yearly organic growth, the network would have failed because of the constant flux of contributors. In 2016, the original network of 880 email address includes over 200 that do not function, emphasizing the importance of maintaining contacts as a quarter of the emails have gone dormant in only 4 years. However, with recruitment, such as the Invasive Mosquito Project (Thackrah *et al.* 2016), new addresses have been added to the list.

During NAMP recruitment, contributors recruited 26% (251 of 960) of the contacts in NAMP and these additional recruits contributed over 60% of received samples (Maki and Cohnstaedt 2015). Although nobody requested anonymity during NAMP, that was the single most important assurance when recruiting individuals for the CDN. Likely this has to do with the differences in types of data. Contributing mosquito specimens throughout their known distribution is not harmful to the NAMP contributors, but contributing disease data was perceived to be harmful to the cervid farmers and so anonymity was important for the CDN. One inducement to participate in the CDN arose from the federal program to provide indemnity to farmers that lost deer to bluetongue and/or epizootic haemorrhagic disease. The network offered a service to participating farmers to store their yearly disease data and create official records for such indemnity claims.

Retention and reward plans

After recruitment, contributors must be retained with rewards such as receiving data and acknowledgements in a tit-for-tat system. Therefore, in addition to a low barrier to participation, a good reward system and a well-thought contributor retention plan are necessary to encourage continued participation. Initially, the NAMP did not have a good retention program and of the 84 original contributors in 2011, only 21 contributed to a second request for samples in 2012. This datum shows collection fatigue is highly likely if repeated sampling is necessary and contributors are insufficiently rewarded and retained. Requesting information without returning information quickly disenchanted contributors. The principal way to retain contributors is to list them in publications and send them information (manuscripts) they helped generate. This worked successfully with the NAMP network, where contributors were listed in the appendix and received articles via email. Consequently when the second step of the NAMP was needed the contributors volunteered again. Two-way information exchange is very important for continued contributions.

In the CDN, data by year were stored at the United States Department of Agriculture. Having a central repository of data for all contributors benefits the project in 2 ways:

1. the contributors develop "inertia" and familiarity with the data entry; and
2. there is a single source of historical data by location and year.

Therefore, once contributors are accustomed to entering data, there is little incentive to change the method as long as they receive some benefit or positive feedback, such as will be reported each year at the national Association meeting and in other nationally relevant meetings. This inertia, the indemnity aid, and data tracking will aid in creating retention of involvement. The longer a contributor sends samples, the more information they have for future indemnity claims.

Safety plan

A very important aspect of the whole process is educating contributors about diseases and their insect vectors. This is particularly important for citizen scientists or non-professionals, for the contributors and the community. Therefore, insect-borne pathogens and risk factors must be clearly communicated and repeatedly emphasized throughout collection instructions and materials returned to contributors. Community and contributor education are 1 of the principle strategies to recruit individuals. Consider, for example, the Invasive Mosquito Project (IMP)[1]. This is a subsequent citizen scientist network of the North American Mosquito Project. It will use high school students to collect mosquito eggs for a mosquito species distribution study. Although the primary goal of the IMP is to map the geographic area of mosquito species, the secondary goal is to educate individuals and improve their public health awareness. To this end, safety is a huge emphasis in the project (Cohnstaedt *et al.* 2016)

Data problems

Data fidelity will always be a problem because individuals may misreport information accidently or intentionally. Therefore, network creators must curate both the content and the contributors depending on the problems the network addresses. In the case of the North American Mosquito Project, the network was designed originally to gather genetic material for a population genetic study. The contributors provided preserved specimens, which were often identified to species. When checking these specimen identifications, several of the mosquito specimens were misidentified and excluded from the analysis. Data fidelity was difficult to ensure within the CDN.

In this case, to confirm animal disease diagnosis, tissue samples were shipped to diagnostic laboratories. Therefore, the network could use the onset data from the self-reporting farmers but the diagnosis was confirmed from the laboratory samples. If case reports did not have associated laboratory samples, they were discarded. Going directly to the laboratories for the data, without the permission of the submitters, is prohibited. In both cases, we had a high level of confidence in the data because of the questions asked or the confirmation by an accredited laboratory.

Prior to establishing a network, developers must develop a robust questionnaire and approach that does not require subjective data entry by contributors. If this is the case, the contributors must be limited to experts. Despite limiting contributors to experts, stringent data cleaning and confirmation will be needed to have a high confidence in the data. More complicated statistical analysis can be used in some cases, but decisions based on the data should always be cautious and confirmed by other sources than contributors. Voucher specimens are another form of validation which is used by the IMP and NAMP networks.

Although the network data may not be reliable for a study, this does not make it invaluable to the contributors because the network is a two-way process. Data, regardless of their scientific usefulness, are exchanged between contributors and network creators. If the contributors are happy receiving data, even with the errors and warnings, than they may continue to send data. This was the case with the CDN. Individuals wanted case data from other farms

[1] www.citizenscience.us.

in the state, regardless of the accuracy of that data. Therefore, researchers need to balance the efforts in time, money, and resources of the contributors with the usefulness of their requests. If data are gathered but not for a continuous question, the network will collapse as it was the case with the CDN when EHD and BT were rare epizootics for 2 consecutive years. Without the need for large indemnity claims, the CDN collapsed and no longer functioned in 2016. On the contrary, the NAMP network is active and growing as the Invasive Mosquito Project. These points are subjective and should be discussed with the contributors prior the establishment of the network and contribution of resources or data.

Carefully planned long-term crowd-sourced data collection is possible if recruitment, retention, and reward plans are carefully constructed to avoid contributor burnout and to return data in a timely and useful manner. Always keep in mind that the contributors have no obligation to help and therefore every effort must be made to accommodate their needs and to answer their questions. When done properly, crowd-sourced data can be a very powerful tool for researchers, especially as quite often these data cannot be obtained without the help of the collaborators.

Acknowledgements

The USDA is an equal opportunity employer.

Grant support

Funded by the USDA CRIS project 5430-32000-007-00D and the North American Deer Farmer's Association Distinguished Scholars Program.

References

Cohnstaedt L.W., Rochon K., Duehl A.J., Anderson J.F., Barrera R., Su N-Y., Gerry A.C., Obenauer P.J., Campbell J.F., Lysyk T.J. & Allan S.A. 2012. Arthropod surveillance programs: basic components, strategies, and analysis. *Ann Entomol Soc Am*, **105**, 135-149.

Cohnstaedt L.W., Ladner J., Campbell L.R., Busch N. & Barrera R. 2016. Collecting mosquito distributions data as a citizen science research project for high school biology students. *The American Biology Teacher*, **78**, 317-322.

Maki E.C. & Cohnstaedt L.W. 2015. Crowdsourcing for large-scale mosquito (Diptera: Culicidae) sampling. *Canadian Entomologists*, **147**, 118-123.

Thackrah A., Cernicchiaro N. & Cohnstaedt L.W. 2016. The Invasive Mosquito Project: a public education tool. *Wing beats*, **27**, 23-24. http://www.citizenscience.us/imp/docs/Thackrah%20et%20al%202016%20IMP%20wingbeats%20article.pdf.