Guidance for Managing Crowd Sourced Information

Henry Reges,
CoCoRaHS.org
Colorado State University,
Fort Collins, Colorado, USA

Julian Turner,
CoCoRaHS.org
Colorado State University,
Fort Collins, Colorado, USA

Crowdsourced information continues to play an ever-increasing part of data available to NMHS office around the globe. Crowdsourced data is now new, it has been around for centuries (1714 – The Longitude Prize; 1848 – Matthew Fontaine Muary's *Wind Charts and Currents* in return for a standardized log of a sailor's voyage). With the power of the internet it has taken on an increasing role to provide data from a greater number of locations than ever before.

This presentation will briefly skim the surface, but will look at examples of best practices for managing crowd sourced data. We will specifically look at a citizen science network's practices as an example, that being the Community Collaborative Rain, Hail and Snow Network (CoCoRaHS).

Why the practices?
A set of practices potentially helps NHMS's to better incorporate crowdsourced data to their advantage. With similar practices across the board, NHMS's will have some key ingredients to leverage the data that is "out there" for their goals and customers.

The benefit of the practices
Having a set of practices to manage crowdsourced data will help NHMS's better set up their offices/websites to incorporate data that may ultimately benefit them. They will know in advance what may be required of them to be successful. With a set of practices (what is needed to manage this data) they will have a recipe to go by and not have to start from scratch.

What are some best practices
There are several best practices for managing the vast amount of crowdsourced data that may be available to a NMHS. We will focus on five ideas:

1. Decide what data you would like to capture.
2. Capturing and Displaying the data.
3. Quality Controlling the data.
4. Ingesting and Storing the data.
5. Disseminating/Sharing the data.

BEST PRACTICES

Before we go into detail, it is worth to note that Probably the best and most important piece in the management of crowdsourced data is having a "good" IT support, whether that be a single person (risky) or a team of developers. Some organizations have none and need to rely on software as a service.

Best Practices for Managing Crowd Sourced Information: Examples from the CoCoRaHS Volunteer Network
Our network, "CoCoRaHS", has a set of practices to provide the best possible outcomes for managing our crowdsourced data. This network is made up of over approximately 20,000 volunteer observers measuring precipitation daily in their backyards. It has become the largest source of daily precipitation measurements in the United States. The network is active throughout the United States, Canada, Puerto Rico, the U.S. Virgin Islands and the Bahamas (figure one)

1) WHAT DATA DO WE NEED, WHAT TO CAPTURE?
The first step for a NMHC is to ask what question they would like answered . . . do they wish to obtain precipitation information, data on temperature, wind information, or a

combination of these and other meteorological elements?

In the CoCoRaHS case we have chosen to only focus on precipitation as we follow the water cycle with manual observations.

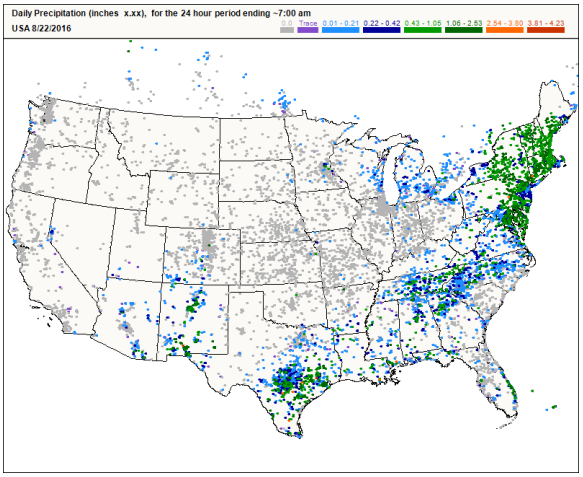

Figure 1: CoCoRaHS Daily Precipitation Map (www.cocorahs.org).

2) CAPTURING/DISPLAYING THE DATA

One of the keys is having an easy to use interface for receiving the data out there. An example of this would be to have a web page (figure two) or mobile app set up where an observer can enter their data.

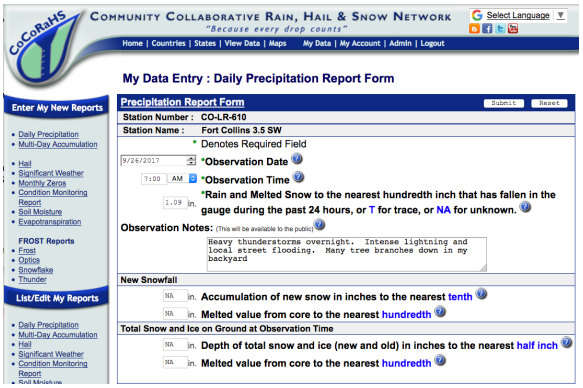

Figure 2: CoCoRaHS Web interface for entering data.

A place to display the crowdsourced data as it comes in is very helpful (figure three). Instantaneous public viewing of the data helps motivate participation of observers and allows for data comparison over a specific area. It also makes for a good QC tool.

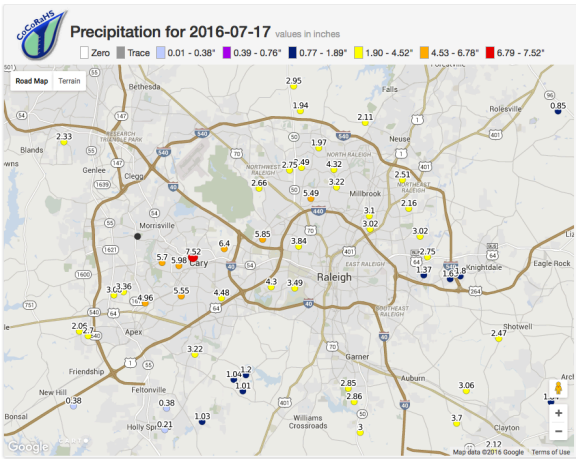

Figure 3: Web display of the volunteer's data.

For a volunteer network, observer training on how to 1) take measurements and 2) use the data interface can help facilitate the receiving of data and provide for more accurate measurements. Observers know what they are measuring and why.

3) QUALITY CONTROLLING THE DATA

It is important that the data is quality controlled for accuracy once it comes in. The better the data, the more reliable it will be to use in making important decisions. High quality data is important to the end users.

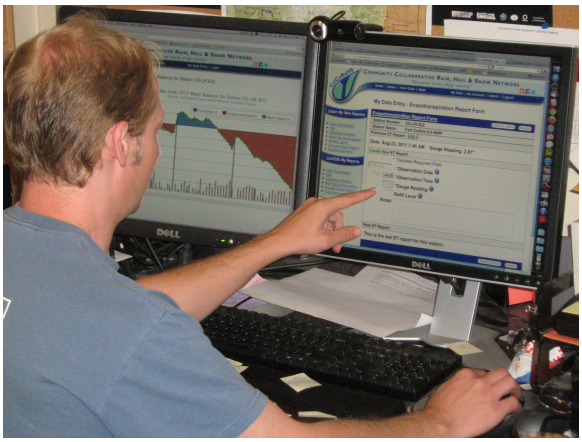

Figure 4: Quality Control Team at CoCoRaHS Headquarters.

Quality control can be done effectively in several ways:

a) As previously mentioned, training can really help with getting the data entered accurately.

b) Another is to set up your system so that it will disallow certain parameters when data is entered. An example would be to prohibit 1,000mm of rain, or display a question after a suspect entry: *"Are you sure this is the correct amount?"*

c) Lastly, it is good to is to have a team (figure four). to quality control the observations once they are received. If the observation is questionable, contact the observer. Often after several contacts, observer's data seems to improve over time.

4) INGESTING/STORAGE OF THE DATA

There are several recommendations regarding the ingesting and storage of your data. They include:

a) Invest in reliable IT services whether on-site or off-site (hosting companies, the cloud, etc.). If you do host your system on site, be sure to have reliable dependent services like power and network connectivity.
b) Make sure the data are securely stored, easier said than done.
c) Find a simple way to ingest the vast amounts of data you might encounter. Ingest from other supplementary networks.
d) Be aware that IT hosting and IT personnel might represent a significant part of your NMHS budget.

- Data Flows

Successful networks will have a lot of data flows (figure five) with other organizations to manage. While those arrows do represent actual digital connections and data transfers, it also represents relationships that have developed with data users and providers. They also represent a connection between your IT staff and outside IT staff because even with a flexible and modular system, those data feeds need to be set up and maintained.

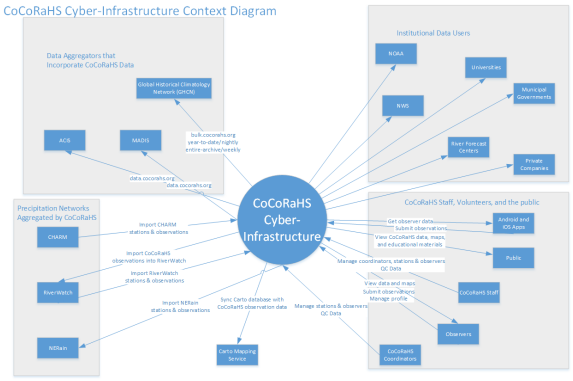

Figure 5: CoCoRaHS Cyber-Infrastructure Context Diagram.

Any modern system will most likely be a combination of custom applications and third party services (figure six). It could be possible to manage a very complex system with nothing but software as a service offerings. However, that will not eliminate the complexity of the system, it will only shift it.

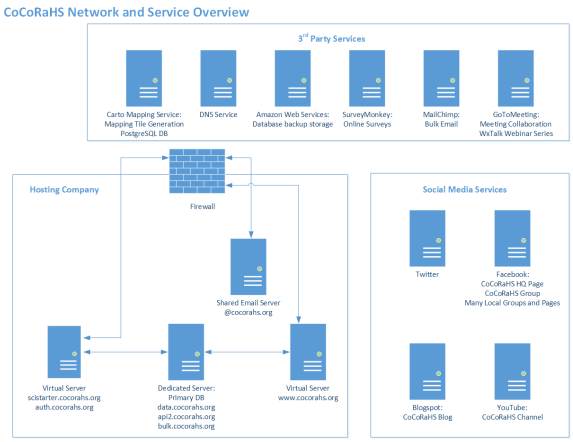

Figure 6: CoCoRaHS Network and Service Overview.

There will be many ways to interact with your different types of users and partners (figure seven).

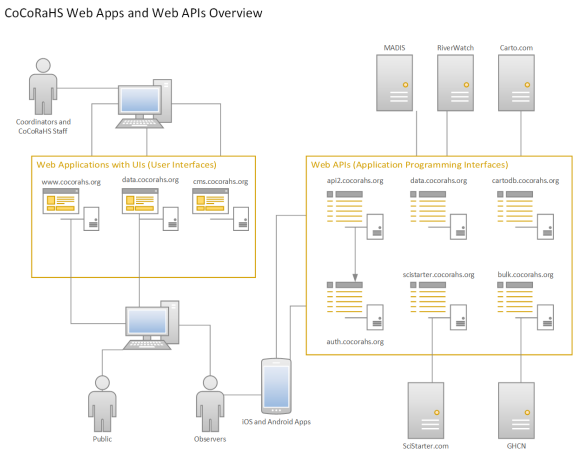

Figure 7: CoCoRaHS Web Apps and Web APIs Overview.

5) DISSEMINATING/SHARING THE DATA

Dissemination and sharing crowdsourced data is the final step in the list of best practices of data management. We suggest the following:

- Develop multiple ways for data users to access your data based on their needs. You want to make it as easy as possible for others to access and use your data.
- Be technology agnostic. Don't require users to support a specific technology. Use standard HTTP and HTTPS endpoints that can be called from any programming language and web browser.
- Support whatever formats are popular (for your users). JSON, XML, CSV are common examples. Some CoCoRaHS data products support SHEF, an older text based format for sharing climate data. By doing so, CoCoRaHS data could more easily be integrated into legacy systems.
- Support different data formats at the field level if possible. For instance, the CoCoRaHS data export system allows data users to specify how they want dates formatted, if they want times in local or UTC, and if they want values in metric or English units. For some data users, having to do those conversions themselves would be a prohibitive barrier to using the data.

CONCLUSION

Managing crowdsourced data does not have to be a daunting task. With the right guidance and following best practices developed by others, you can succeed in harvesting the data available, manage it, and finally shape it into products that your stakeholders/customers can use to meet their needs.

REFERENCES

Reges, H., N. Doesken, J. Turner, N. Newman, A. Bergantino, and Z. Schwalbe, 2016: COCORAHS: The evolution and accomplishments of a volunteer rain gauge network. Bull. Amer. Meteor. Soc. doi:10.1175/BAMS-D-14-00213.1, in press.

IMAGES

Figures 17: Image/Photo credits - CoCoRaHS