

Introduction to Citizen Science & Scientific Crowdsourcing

Citizen science data management issues
8th February 2018

Luis FP Velasquez – Earthwatch Institute
lfvelasquez@earthwatch.org.uk



Creating Knowledge. Inspiring Action.



Hi I'm Luis Velasquez. I'm the GIS specialist at Earthwatch Institute and I have been working with citizen science data for over 5 years. My interest in data management lies in trying to understand the different processes that enable the relationship between professional scientist and citizen scientist to go beyond data collection, I have an especial interest in the areas of analysis and visualization. I believe that we all as part of the citizen science movement have the duty of striving towards delivering the best outputs possible as a results of the efforts made by volunteers, and that is why understanding how to make the best use of the data through the implementation of data management plans is a great step towards this essential aim.

Lecture Outline

What is data management?
Data management - life cycle



Creating Knowledge. Inspiring Action.



We will start by looking at the concept of Data Management and how this is related to citizen science. Subsequently we will go through the different stages of the data management life cycle and what type of considerations are needed when creating or reviewing the project's data plan.

Data Management

The array of data management techniques used in traditional research can be easily extended to citizen science projects.



Source: <http://www.digitalistmag.com/technologies/analytics/2014/11/07/do-you-have-enough-information-to-make-a-sound-decision-01719714>

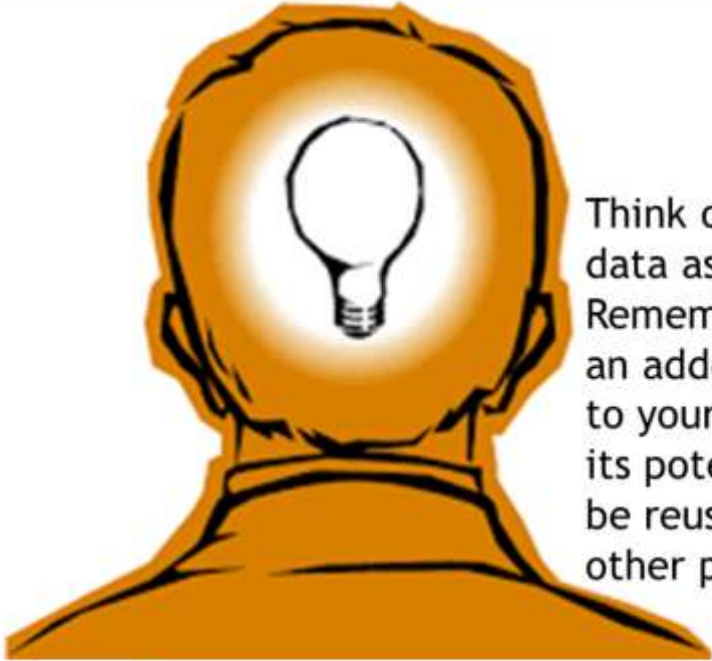


Creating Knowledge. Inspiring Action.




2


The array of data management techniques used in traditional research can be easily extended to citizen science projects however, due to the nature of the project – volunteer involvement – there are certain aspects that need further attention. The majority of people involved in in citizen science projects do not have formal training in data management hence the importance of understanding the considerations and implications of working with citizens to generate scientific outputs. It is understood that by implementing data management practices the outcome of citizen science can transcend from the creation of new scientific knowledge to support policy and decision-making process.



Think of your data as an asset. Remember that an added value to your data is its potential to be reused for other purposes

 EARTHWATCH INSTITUTE

Creating Knowledge. Inspiring Action.

 3

Thinking of your data as an asset ensures its effectiveness in the short, medium and long term, this line of thought will benefit if you understand the data management life cycle, which will be exploring next. It is important to always keep in mind that an added value to your data beyond its intended use is its potential to be reused for other purposes.

Data Management - Life Cycle



Creating Knowledge. Inspiring Action.





As previously mentioned we will be describing the different stages of the data management life cycle and we will be using the Data Management Guide for Public Participation in Scientific Research developed by DataONE (the guide has been added as part of the reading list). I tend to find this guide particularly useful as it goes beyond project goals and sees data as one of the vehicles through which a citizen science project can generate a greater impact.



At this stage you should will be thinking of the data life cycle as a whole entity, during the planning stage you need to review and make decision regarding the project's data. The type of questions you will be asking at this stage include:

What data are you collecting?

Does the data already exist?

Who is responsible for creating the metadata? A simple way of understanding metadata is to see it as information about the data i.e who is collecting it? how have the data been collected - methodology? how often is the data collected?

Another very important question at this stage is: are there any specific requirements for sharing or storing data? The best way to understand this questions is by imagining the data collected by citizen scientist is part of a big research project, this could mean that you won't be able to share the data until the outputs of the research have been made public, with regards to storing the data, you will be considering aspects such as the creation of a database - we will be exploring this concept later in the presentation – or perhaps the possibility of using existing offerings such as citsci.org

Either way it is important to answer all these questions considering the project aims as you want your data to answer the questions posed during the design of the project, remember to think how much you want your project to transcend?



Most people working or wanting to work with citizen science are familiar with this stage of the data management life cycle. In here it is common practice to talk about a **data model**, what this means is the description of a process that identifies what you will be collecting, the data format, and how this will be organized and processed. When going through this stage you will be thinking:

How will data be collected?

How will you be storing the data?

If you are obtaining data from an outside source, the question you need to ask yourself is: will I be able to store a local copy of the data?

A good example of this will be the use of data offered by the Environment Agency in the UK or NASA, these two agencies offered the data to the public free of charge and allow any use, in other words you will be able to store it locally – The best way to understand this process is by familiarizing yourself with the concept of copyrights licensing especially the Creative Common license – check the link in the slide.

You need to see this stage as a very important part of the data management cycle, in here you will be aiming to find the answer to your initial hypothesis through the data collected by citizen scientist. It is also important to be thinking beyond the scope of the project; for example what other data can be indirectly collected that can aid the impact generated by the project, as well helping other lines of research such as social and computer sciences.

Great care must be taken when designing this stage within a data management plan some of the most significant issues in citizen science projects origin at this stage e.g. data storage, use of technology and project outreach amongst others. A clear understanding of the mechanisms are involved in this stage could help to avoid future

headaches when delivering a cit-sci project.

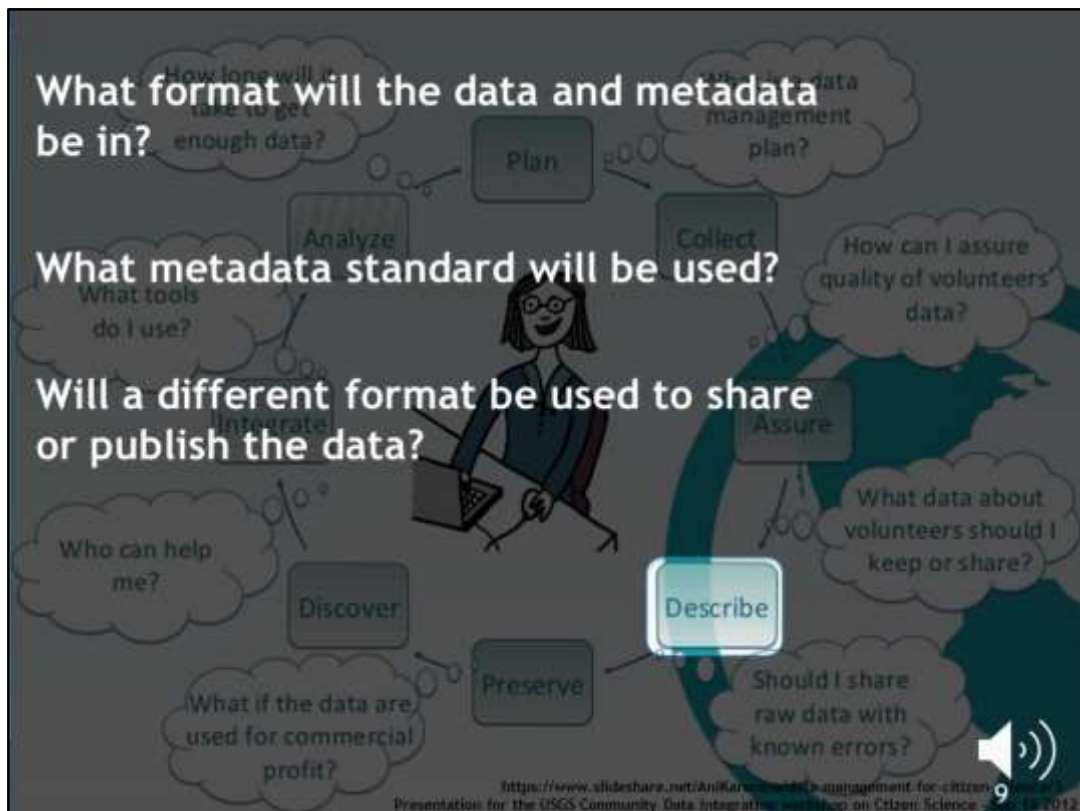


At this stage you are thinking of what are the project goals as well as what type of scientific standards you would like the project to follow, you are basically defining the quality of the data you will be expecting to collect. It is common practice for citizen science projects to define the methodology use before and during data collection (Quality Assurance – QA) as well as what process the data should go through after its collection (Quality Control – QC). Typical questions at this stage are:

Are the QA/QC steps being documented?

Is QA/QC occurring throughout the data life cycle?

Is data that is transcribed or copied checked for errors against the original data?



In recent years this stage has been gaining more and more importance in citizen science projects especially as the cit-sci community is aiming to use this type of initiatives to have a higher level of influence in policy and decision making activities. Having say that, this is also one of the main issues in data management particularly with the boom of citizen science initiatives across the globe. It is important to highlight that it is through this stage that you will enable the data collected as part of your project to go beyond its initial purpose. Here we are writing how the project will create metadata, which as we previously mentioned is just information about the data i.e. who collected it? How was data quality ensured?

When writing this section, the sort of questions you will be asking are:

What format will the data and metadata be in?

What metadata standard will be used?

Will a different format be used to share or publish the data?

This is a very complex topic and there are several groups working in identifying metadata standards for citizen science projects, such as the European Citizen Science Association (ECSA) Data, Tools and Technology working group and the U.S Citizen Science Association (CSA) Data and Metadata Working group.



This requires thinking in the medium and long term life cycle of the data. Normally, projects only think in the short-term preservation of the data, and even though this is especially important, establishing a long-term archive or data repository for the data collected during the duration of the project should not be underestimated. Any decision made regarding the preservation of data in the short term should also considered the long-term implications. The type of question you will be asking include:

How is the data being stored and backed up?

Is there someone checking to ensure that backups are being done properly?

How long will backups be kept?

With new regulation coming into force from 25 May 2018 it is essential for anyone working in citizen science projects to understand the General Data Protection Regulations. This refers to the regulation intended to primarily give control back to citizens and residents of the European Union over their personal data. An example of aspects you need to take into account include a document of what personal data you hold, where it came from and who you share it with. A very important aspect to take into account regarding this process is when using online platforms such as citsci.org or ispot from an European country especially as there are based in America and their data protection regulations are different from those in Europe and you might be in breach of a regulation. What you are looking to understand here is where the data centres are based – these are the physical places where your data is stored i.e the building with the big computers, and example will be Microsoft which has data centres in Ireland, UK, Netherlands, so a good couple of questions to yourself or to any technical staff giving support with the data management in the project will be : where is the data centre

based? And Do they comply with the European Data Protection Regulation?
Several issues in data management can arise if not enough care is taken when going through this stage, and some of the issues can be of a legal aspect. For further information regarding this topic please follow the link in the slide.

At this point I would like to stop so that you have the chance of thinking about all the information we have just been going through before we move forward with the other stages of the life cycle.



The DataONE methodology has described two faces for this stage of the data management plan:

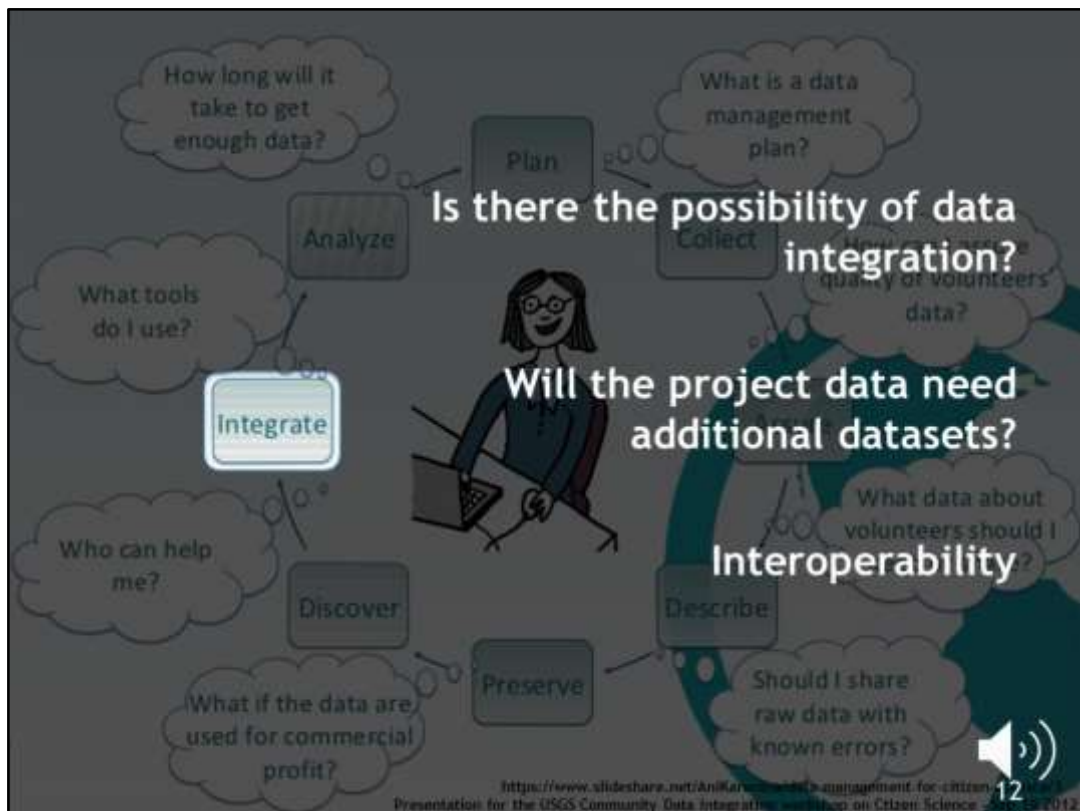
1. Finding Existing data: at this point you will be asking yourself is there any data that can be used in conjunction with the project data?
2. Making the project data available.

This is a straight forward part of the process and it requires the definition of the methods through which your project's data will be shared, as the project aims to be reachable by different audiences. When working in this section you are trying to find the answer to the question: where is the data going to go so that it can be shared?

It is important to clarify that when your data becomes available this can take place in two ways, the first one data discovery: which means that anyone will be able to find your data but not everyone will be able to see it or access it, and the second one is a term that is highly linked with this stage and citizen science projects is 'Open Data', which is data that anyone can access, use or share with out restrictions. The most common example of these type of open data can be found at data.gov.uk and National Biodiversity Network – NBN - I will recommend you visit this site to check what type of data is available, some of the examples include environmental, government and education data

Other questions are:

- Is there a deadline or schedule for sharing your data as required by the funding agency?
- Does your project or program have a specific repository for your data?



It is important to define within the data management plan what type of integration the project will be aiming to have for the data collected by citizen scientist, some of the possibilities include integration with other projects or citizen science observatories. The success of this stage is linked to the clarity and robustness of the data documentation (metadata) particularly as the possibility of reusing the data by other users outside of the initial scope of the project depends on how easily they can understand and gain access to the data. Some of the questions include:

Is there the possibility of data integration?

Will the project data need additional datasets?

In this stage of the data management plan you are dealing with an important concept that is currently in the spotlight - Interoperability, which is the ability to coordinate different citizen science initiative towards the same objective. This is a very interesting topic and it is important to people working in cit sci to have an overview of this concept especially as it is continuously evolving. I have added a link in the transcript to the European Joint Research Centre, which contributes to scientific and technical debates about Citizen Science data and service interoperability

<http://digitalearthlab.jrc.ec.europa.eu/activities/international-collaboration-advance-citizen-science-interoperability/57576>



This part of the data management plan is normally well documented by people working in citizen science as this is guided by the initial aims set by the project. It has been documented that a clear understanding of potential analysis and visualizations has a positive influence in the methodology use for collection of the data. This stage of the life cycle will benefit from an inter-disciplinary approach as the combination of expertise can result in a more wholesome analysis of data – For example the involvement of social sciences can give a greater understanding of engagement dynamics whilst an expert in gamification (the application of typical elements of game playing e.g. point scoring, competition with others) could help with the levels of engagement during the life span of the project, having the involvement of people with knowledge of website protocols i.e. web designers could help the analysis and visualization of data which can be aim to increase public outreach.

Some of the main questions when thinking of this stage are:

What are the project goals?

Which are the expectations of intended audiences for project results?

Another important aspect of this stage is monitoring and evaluation of your project particularly as you will be aiming to demonstrate project output to the projects funders as well as any partner, in addition of providing overall understanding of citizen science and its benefits. It is important to remember that monitoring and evaluation play an important role in demonstrating how successful you and/or your organisation are at delivering Cit-Sci projects as well as possible areas of development.



When setting a new citizen science project the incorporation of a data management plan should be a straight forward process as this can evolve along the initial project. The stages of the data management plan can also take place at any time during the project thus existent initiatives can also benefit of creating a data management plan when the need arises.