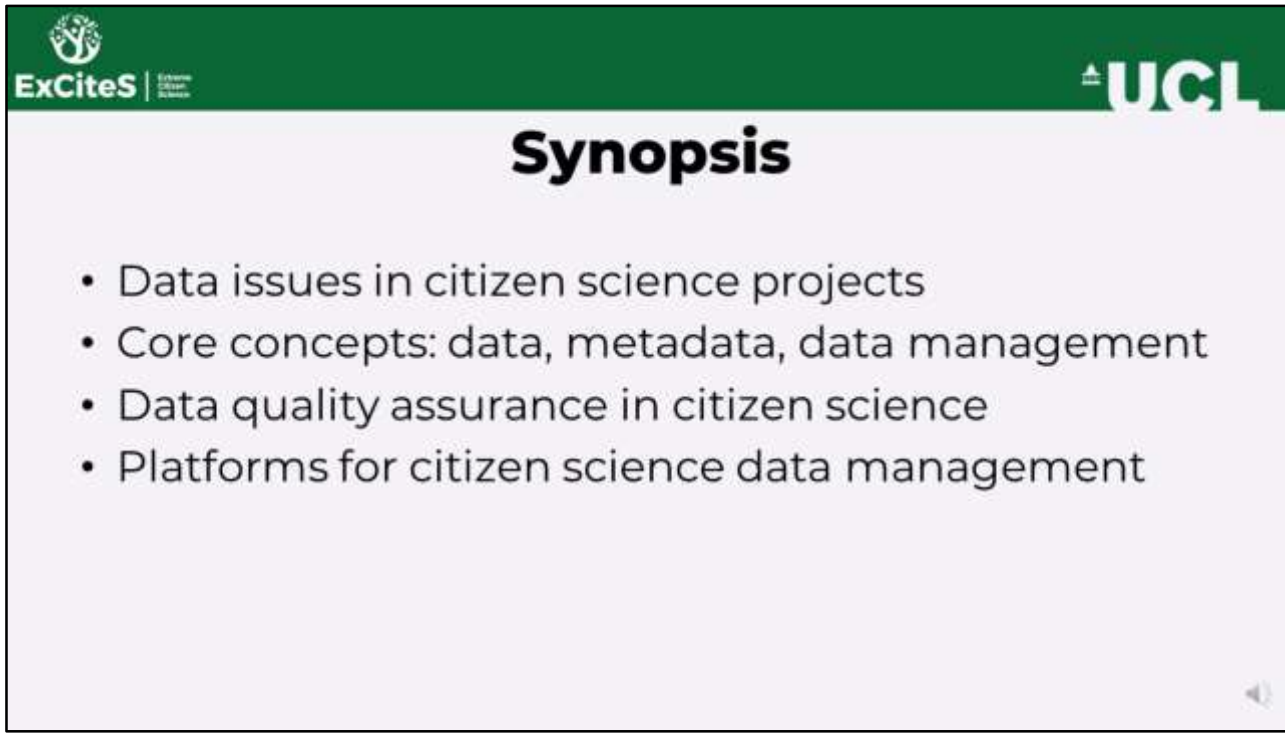
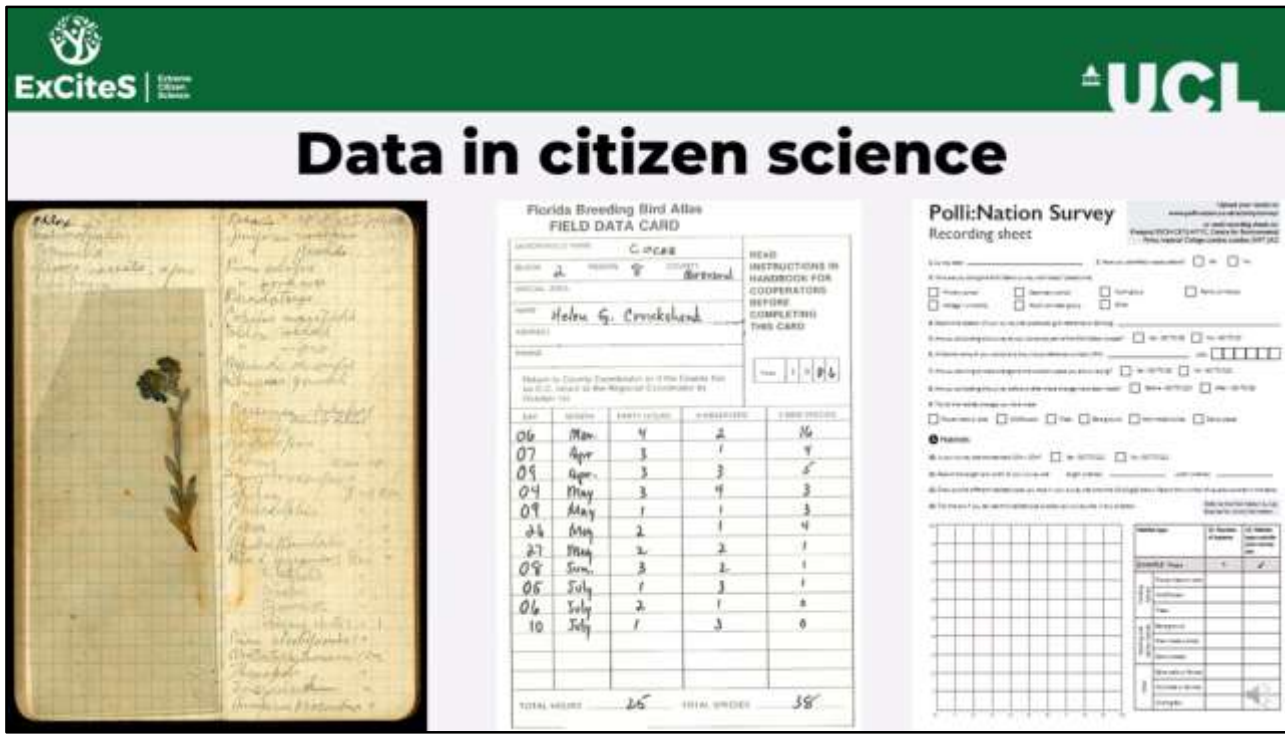

In this talk, we look at data management - without good data, we cannot do science…




In terms of coverage, today we start by explaining the data issues in citizen science – what are they and why are we concerned about them. We will also understand some of the core ideas about data – capta, data, information, knowledge, and wisdom. After understanding these basic concept, we turn to one of the most frequent issues that are coming up in citizen science – the concern over data quality and quality assurance. The next part of the lecture will look at the parallel area that has grown in the geographical literature, that of volunteered geographical information and we will see how insights from it can be useful in citizen science. We conclude by looking at some of the platforms that are available to manage citizen science data.

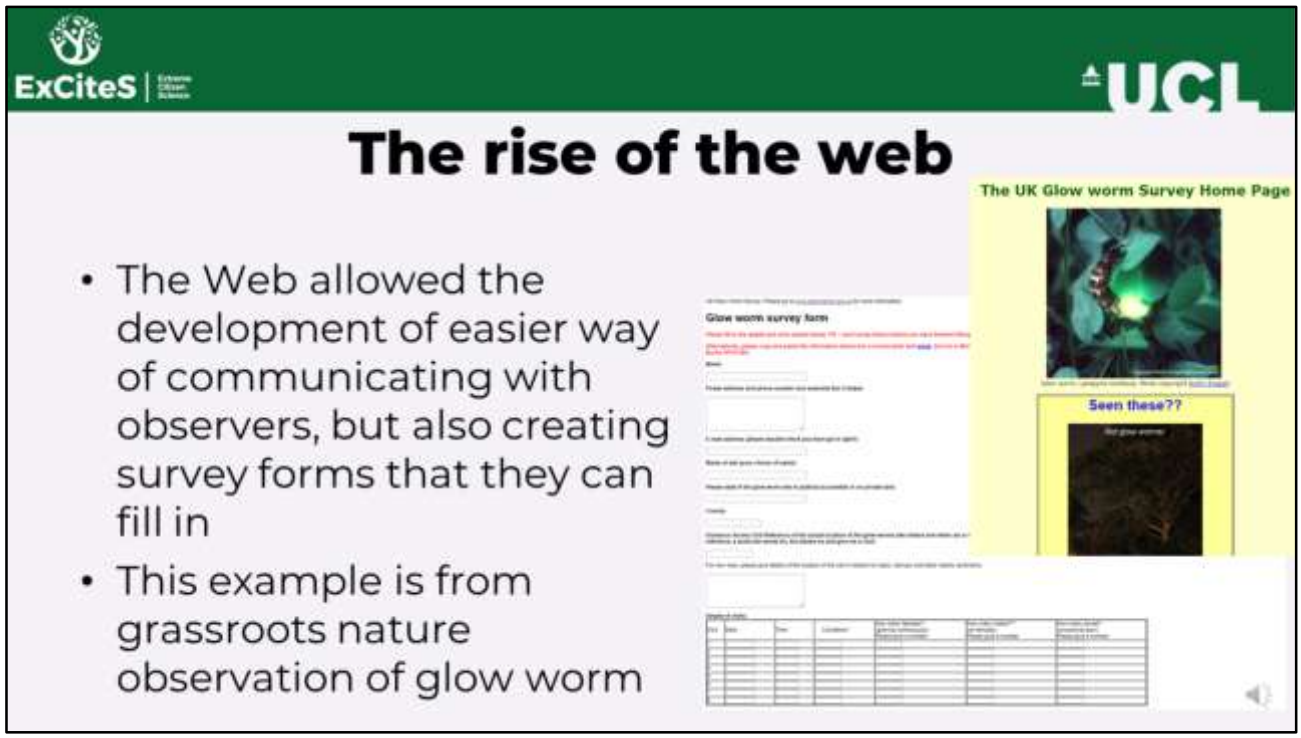



The history of data in citizen science, is similar to the history of data in the wider scientific practice, with some additional complexity because of the range of participants, their skills and their knowledge. Paper and forms continue, to this day, to be a major tool for recording observations and sharing them. Here we see examples from the field notebook of Vernon Bailey from 1903, in New Mexico, a field data card from a bird survey in Florida, and a recent form from the Open Air Laboratories project. We can already see that in all citizen science project, collecting data is important – science includes the systematic collection of empirical observations, so careful documentation of what was observed is central to the activity.

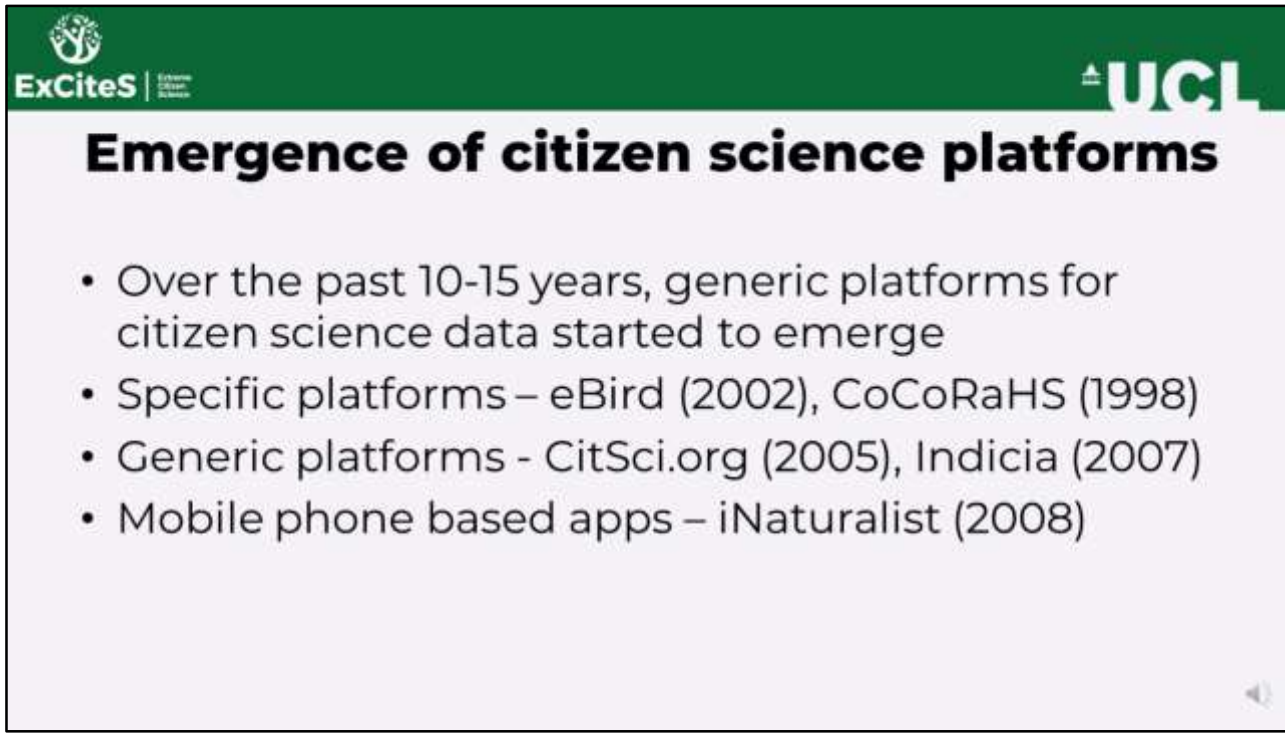
In terms of images, on the far left we see the "Field notes, New Mexico, May 16-October 14, 1903 (Pages 24 and 25) BHL46205853" which are from 1903 by Vernon Bailey. "Vernon Orlando Bailey (1864-1942) was born in Manchester, Michigan. At an early age his family moved to Elk River, Minnesota, where he developed an interest in natural history. Around 1885, Bailey began sending collections of birds and mammals to Clinton Hart Merriam, Chief of the newly created Division of Economic Ornithology and Mammalogy of the United States Department of Agriculture (in 1896 the name was changed to the Bureau of Biological Survey). In 1887, Bailey was appointed Special Field Agent for the Division of Economic Ornithology and Mammalogy. His title was changed to Chief Field Naturalist in 1890, and he remained with the Biological Survey until his

retirement in 1933. Bailey's chief biological interest was the study of the life history and distribution of mammals. During his career with the Biological Survey, he made field investigations throughout the United States, Canada, and Mexico, including intensive biological surveys of Texas, New Mexico, North Dakota, and Oregon."
(https://transcription.si.edu/project/7288)
Next to it, is a 1986 observation card from Florida Breeding Bird Atlas volunteers
(copyright state of Florida http://legacy.myfwc.com/bba/chapt1.asp)
The final example is from OPAL pollinator survet

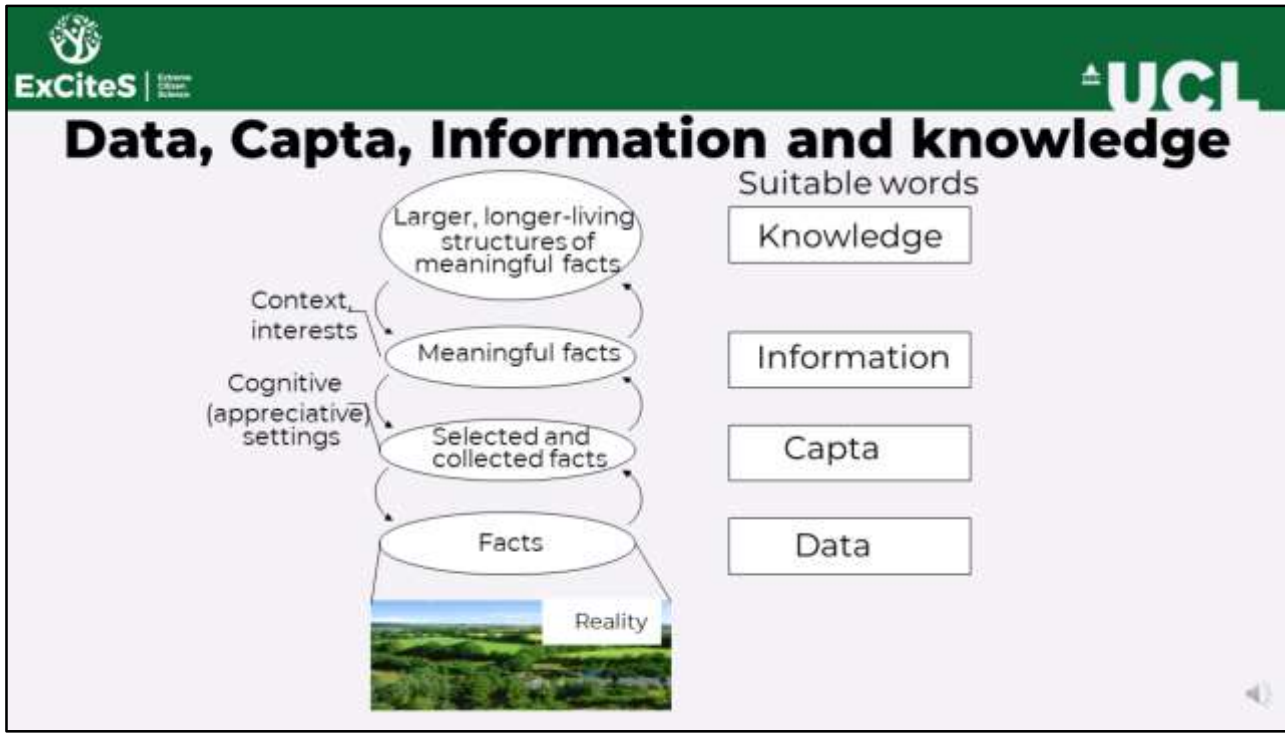

In addition to direct observations and taking notes, science and citizen science rely on the use of samples and sensing of the environment – this is another way of collecting and generating data. Sensing is quite central to environmental science. Here we see an example of sampling air using Palmes diffusion tubes which measure the concentration of pollutants such as NO2 in the air, or the use of cameras during a bioblitz to record environmental observations, or a sound meter to record the level of noise that was measured at a location. In all these cases, we see a method to sample information from the environment and turn it into data.




Within this period, and in accordance with the trends that we have covered in week 1, we can see the evolution of the way citizen science data got digitized, but also standardised (meaning that all the observations or the details are coming in a common format which makes it easier to analyse and share). By the mid 1990s, it was possible to create forms online and to allow people to fill them in on the web. In this example, we see a typical form for data collection that is coming from the UK Glow Worm Survey – this is run by amateur naturalists (citizen scientists) without an official affiliation to a research institute. Online forms, unlike observation cards or other paper versions, force the participants to enter very specific information – which mean that the process got standardised




While specific projects collected data using forms and started to understand the utilisation of the web to get information from participants, the changes towards fully integrated systems for working with volunteers and managing the data. The early systems – such as eBird for bird observations, of CoCoRaHS for Rain, Hail and Snow created specific platforms that are addressing data collection and management for their specific domain. Such applications are common and valuable in a specific study area or a topic. Then in the past decade, we start to see more generic platforms which provides support for more than one type of project. CitSci.org, which we've seen in the reading for class, or Indicia (indicia.org.uk) are systems that provide support for biological recording and observations. Finally, with the rise of the smartphone, we see applications and platform that rely on the availability of the phone – such as iNaturalist. We will see how these platform work later. Now let's turn to critical concepts that we need to pay attention to in the context of data – the definition of data, and the need to have data about data – metadata.

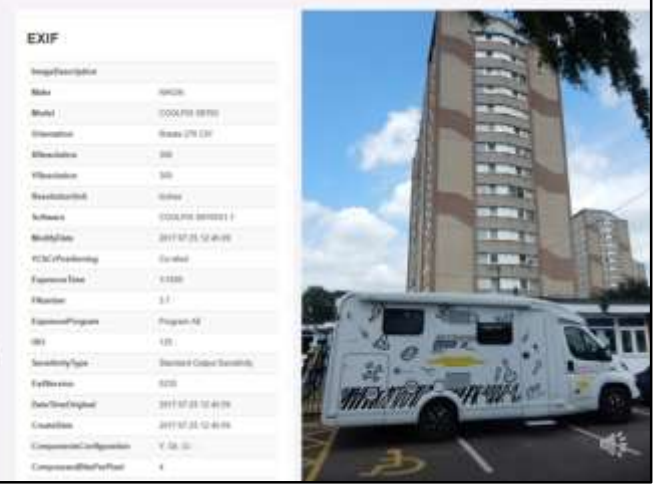



Here we come across the first set of concepts: data, capta, information, and knowledge. First, we have reality, and it is rich and complex, and not all of it can be captured with the identification of specific facts that can be clearly define. The facts that are well defined and can be captures are called data. Above data, we will use he the idea of capta – facts that are actually collected from the real world – which as was developed by Checkland and Holwell (1998) book "Information, Systems, and Information Systems". It is used to differentiate from data – all the facts about the world and point that in any given situation we are making a decision on which facts we want to collect out of all the possible facts. This mean that even at the data collection stage we've already made some important decision on what questions we can ask. Information is frequently understood as an organisation of data in a way that makes it possible to organise the data and make sense of it. Information can be turning the data that we have collected (cpata) into a table that allow us to examine the changes throughout a day by marking the hours as one organising aspect. Above information, we have knowledge, which is the identification of common structures of facts and identification of patterns, rules or other structures within the capta that we've collected. Despite the introduction of the idea of capta, for all intent and purposes what we usually call data is actually capta, so in order to avoid confusion, we will stick to the term data.

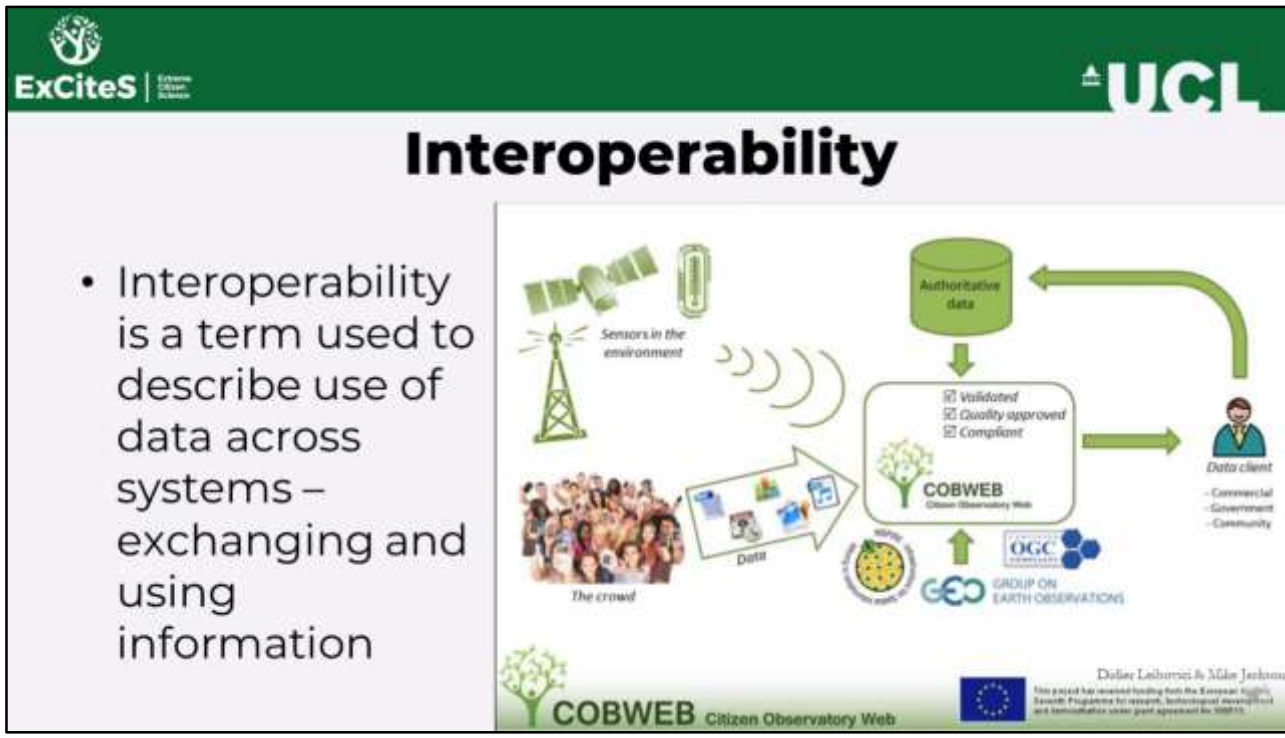
ExCiteS | UCL

# Data and metadata

- When we capture data, often we want to capture some facts about the data itself – this is called metadata. In other words "a set of data that describes and gives information about other data."




Another core concept that we want to remember is the concept of data and metadata. Here is an example to make this easier to understand. Our data is a picture of a Science Bus (which visited a community centre in Birmingham in the Doing It Together Science). The data in this case is an array of colour values that together make the image that we can show here – the specific values for each pixel that make the image. However, we want to capture further information about the image – which camera was used, the resolution, date, time, location etc… Image formats such as JPG include an element called EXIF, which is a piece of information (or a collection of facts) about the image, which is the main data. It all come together in one file, but actually we have data, and data about data – metadata. Metadata is very valuable in analysis processes and can be used in many ways.




The final concept that we need to cover is interoperability. These are standards – this time in the meaning of official agreement about structure, content, and all the other technical details, which facilitate the ability of one computer system to exchange data and information with another system. Here we see the general conceptual framework of COBWEB, and you'll note that there are relationships to the Open Geospatial Consortium (OGC), Group on Earth Observation, and the European INSPIRE Directive. All of them are setting standards on how information about the environment should be shared, and this need to be taken into account with respect to citizen science data.