WEEK 8: CROWDSOURCING AND VOLUNTEERED GEOGRAPHIC INFORMATION

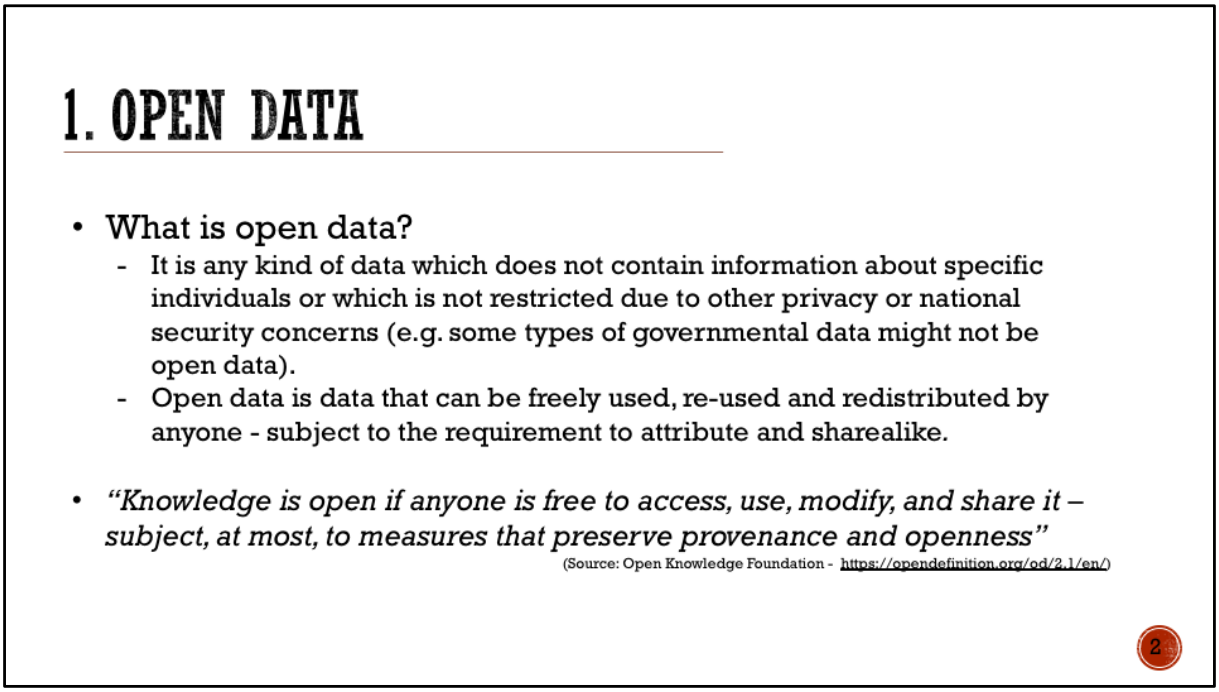
Dr Artemis Skarlatidou
Extreme Citizen Science group (Dept of Geography, UCL)
a.skarlatidou@ucl.ac.uk

LECTURE OUTLINE

- Open Data;
- Crowdsourcing;
- Volunteered Geographic Information;
- Ethical, Legal and Privacy Issues in VGI;
- 4 VGI Examples;
- VGI quality: Data usefulness, usability and trust

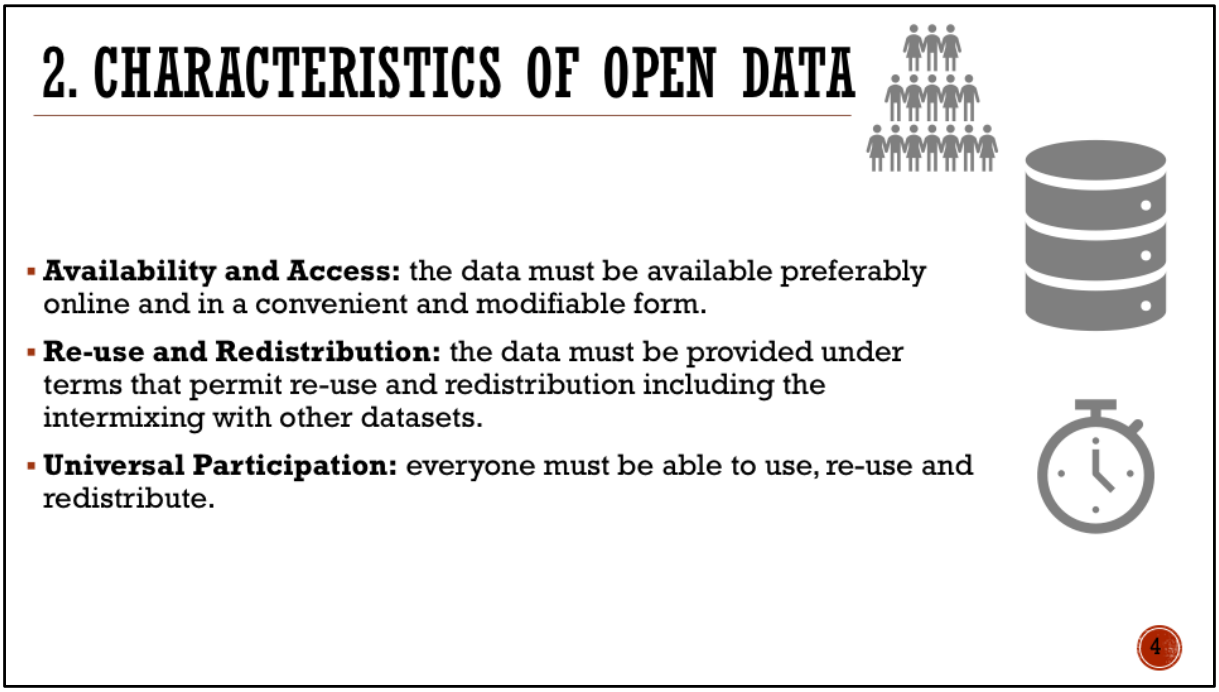
Citizens are increasingly becoming an important source of geographic information, sometimes entering domains that had until recently been the exclusive realm of authoritative agencies. This activity has a very diverse character; it can be active or passive, involve spatial or aspatial data and the data provided can be variable in terms of key attributes such as format, description and quality.

This lecture covers the basic principles and definitions for open data, crowdsourcing and Volunteered Geographic Information (VGI). It shows examples of how VGI is used in various contexts (e.g. disaster relief, conservation, in local authorities and urban planning). It explores issues of legal, privacy and ethical concerns in VGI which are usually ignored or not well understood by the producers and consumers of this relatively new form of data source. The lecture also briefly explores some data quality issues and the design implications for data usability (i.e ease of use), usefulness and trustworthiness which will be further discussed in depth later in this week's seminar.



'Open Data' as a concept originated due to the belief that the enormous amount of information routinely collected by government entities should be available to all citizens. In the late 2000s, governments and other organisations began to allow a greater number of users to have access to these resources. The first government policies on Open Data appeared in 2009. Today, more than 250 governments at national, subnational and city levels and other entities such as the World Bank and the United Nations have launched Open Data initiatives—and more are being launched every year.

Data are considered to be "open" if anyone can freely use, re-use and redistribute them, for any purpose, without restrictions. While a large amount of data are published on government websites, the majority of the published data are intended only to be read as stand-alone documents, and not to be re-used for other purposes. Though in order to be considered as "open," the data must be re-usable, meaning they can be downloaded in open formats and read by software, and users have a legal right to re-use it.



Three are the main characteristics or conditions which define open data.

Availability and Access: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.

Re-use and Redistribution: the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.

Universal Participation: everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes are not allowed.

3. OPEN DATA – WHY IS IT IMPORTANT?

- Transparency and democratic control;
- Increases the degree of autonomy in th lives of people in order to enable them to represent their interests in a self-determined way;
- Promotes public participation;
- Leads to improved and innovative business solutions;
- Leads to the production of knowledge through the combination of data patterns.

5

The European Commission, before launching the European Data Portal, run several studies to obtain further evidence on the impact of Public Data Resources. A study was carried out with the aim to collect, assess and aggregate economic evidence to forecast the benefits of the re-use of Open Data for all 28 European Member States for the period BETWEEN 2016 TO 2020 and highlights the following benefits:

1. Open data improves transparency and democratic control – which at times of lack of public trust is extremely important for our societies.
2. Increases the degree of autonomy in the lives of people in order to enable them to represent their interests in a self-determined way.
3. it Promotes public participation – if people have access to information, theoretically they can fully participate in decision-making processes FOR ISSUES that concern their lives
4. It Leads to improved and innovative business solutions;
5. and it Leads to the production of new knowledge through the combination of data and the identification of new complex data patterns

Due to these benefits the European Union has adopted legislation to foster the re-use of Open (Government) Data.

[Note that the source of the actual study for further information is under the image on the next slide]

4. OPEN DATA – WHY IS IT IMPORTANT?

Further details about the full impact of re-use of Open Data – Source:
https://www.europeandataportal.eu/en/highlights/creating-value-through-open-data

6

Some of the identified benefits for member states in the EU that are related to the re-use of Open Data include:

- 100,000 open data jobs by 2020;
- Public administrations seek to further understand the cost savings they can realise thanks to Open Data, which was estimated to be around 1.7 bn EUR (and we will see later a direct example of this from in the context of VGI);
- Interestingly, the report highlights efficiency gains of Open Data which were translated into achieving environmental benefits, saving 1,425 lives a year from road fatalities. Further use of Open Data in traffic can save 629 million hours of unnecessary waiting time on the roads;
- There is a broader discussion of the Open Data benefits if you visit the link on the slide, which I encourage you to do so, and critically reflect on these benefits as well as thinking of the potential implications of using Open Data.

5. CROWDSOURCING

"Remember outsourcing? Sending jobs to India and China is so 2003. The new pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R & D".

By Jeff Howe

Howe, J. (2006) The rise of crowdsourcing. Wired Magazine, 16 (06).

https://www.wired.com/2006/06/crowds/

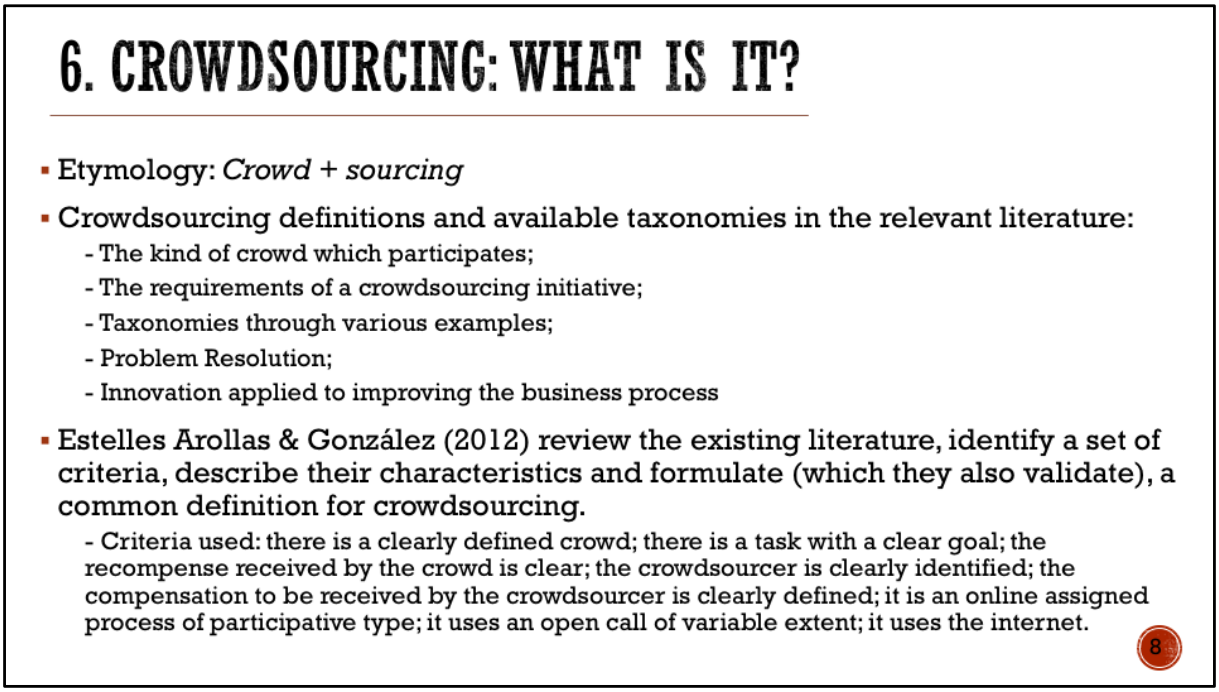
7

Crowdsourced data contributes massively to big data and it has been considered "the future in data collection". That is because because of the numerous successes in crowdsourcing and the simplicity of these schemes. Of course there are many who would argue that crowdsourced data is actually **a necessary tool within big data collection and we cannot really have one without the other**.

One of the very first papers on crowdsourcing which actually coined the term is by Jeff Howe in the Wired Magazine in 2006.

The author explains that the word crowdsourcing is used for a wide group of activities that take on different forms. The adaptability of crowdsourcing allows it to be an effective and powerful practice, but makes it difficult to define and categorize and some times explain what it is.

It is a very interesting paper and the link is provided here for anyone who want to have a look.



From an etymological point of view: The name crowdsourcing is formed from two words: *crowd*, making reference to the people who participate in these initiatives; and *sourcing*, which refers to a number of procurement practices aimed at finding, evaluating and engaging suppliers of goods and services".

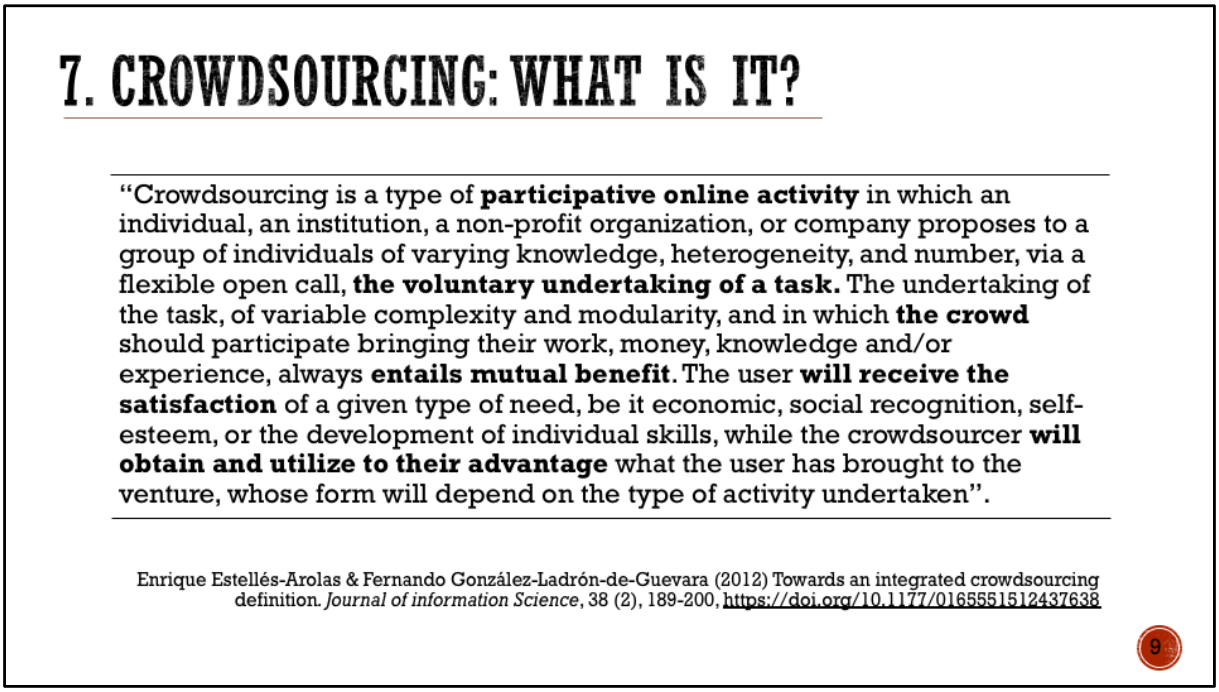
As simple as that sounds and as much as we all already know about crowdsourcing, in the academic literature there is still a lack of consensus and a certain semantic confusion of what crowdsourcing really entails. Some scholars make a general overview of various characteristics of crowdsourcing which resulted in different taxonomies based for example: on the kind of crowd that can participate; the incentive schem; the different variants of crowdsourcing initiatives; the requirements of a crowdsourcing initiative; crowdsourcing concrete examples which subsequently support the development of other taxonomies; and other crowdsourcing characteristics such as problem resolution and the innovation aspect of crowdsourcing in improving business processes.

What is interesting, is that depending on the perspective and the definition being used, certain initiatives classified by some authors as crowdsourcing are not classified

as such by others. For example, for some studies Wikipedia and YT are considered classic examples of crowdsourcing, while others declare the opposite in both cases.

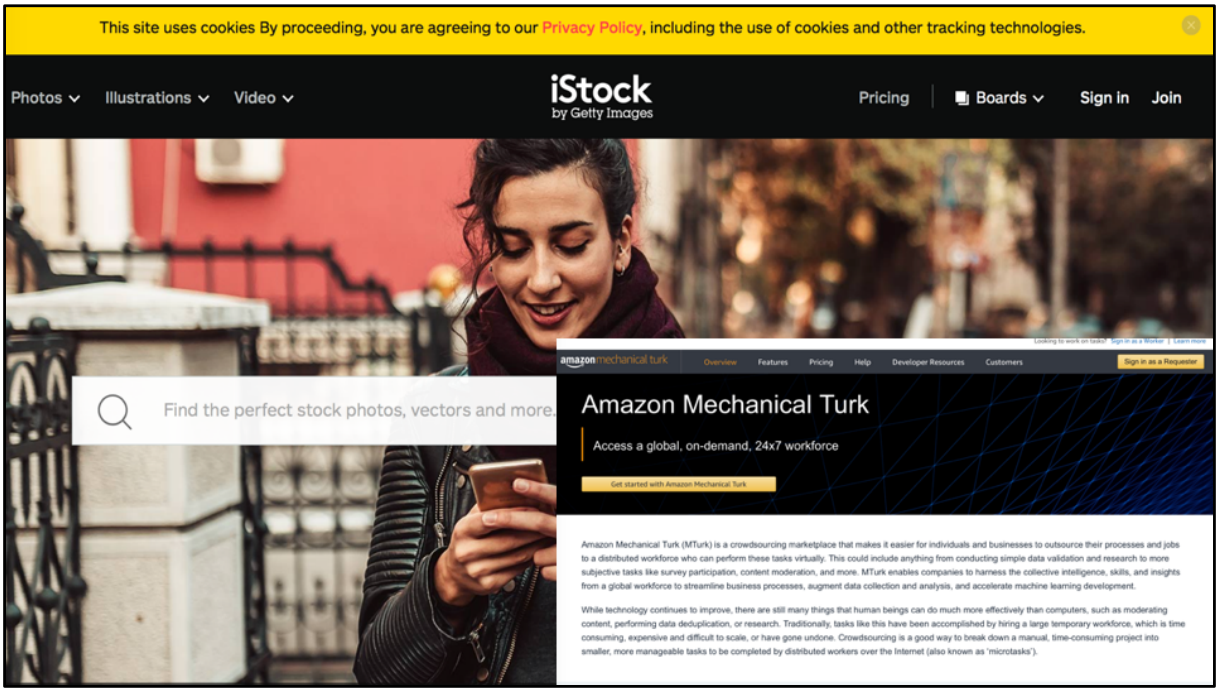
A very highly cited study on seeking clarification and an academic definition for crowdsourcing is by Estelles Arollas and Gonzales; The authors review the broader academic literature and they formulate – which they later also evaluate – a common definition for crowdsourcing which carries the following key characteristics:

- First, there is a clearly defined crowd;
- second, there is a task with a clear goal;
- Third, the recompense received by the crowd is clear;
- Fourth the crowdsourcer is clearly identified;
- As well as, the compensation to be received by the crowdsourcer is clearly defined;
- Sixth it is an online assigned process of participative type;
- Next, it uses an open call of variable extent;
- And finally it uses the internet.

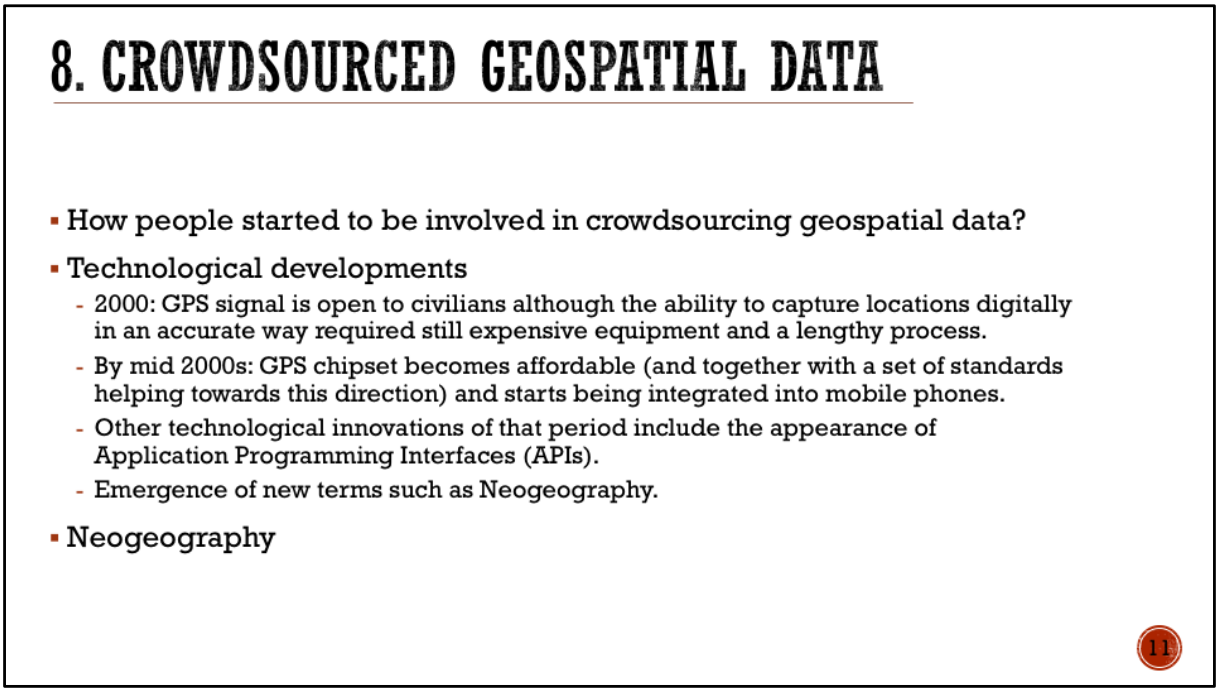


From these characteristics the authors came up with the following definition.

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** what the user has brought to the venture, whose form will depend on the type of the activity undertaken"



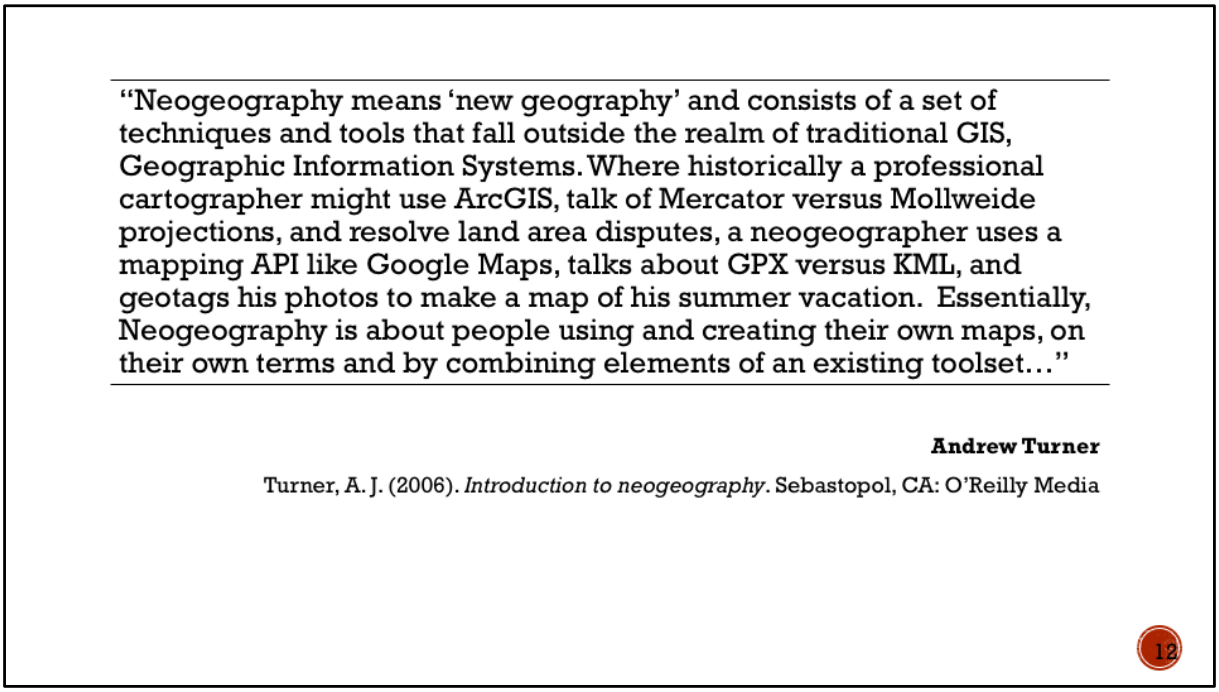
Using that definition, the authors go on to describe some examples of crowdsourcing, which include the very popular Amazon Mechanical Turk (which if you are not aware is a platform where crowdsourcers can propose tasks that are offered in exchange for money) and the also very popular iStock platform (which I am sure you have heard of – an image sale platform).
The authors also explain, why other examples such as Wikipedia and YouTube do not fall under their definition as they do not comply with specific crowdsourcing criteria.



- For centuries mapping and the collection of geospatial information has been accomplished by state organizations, with civilian National Mapping Agencies, mostly responsible for land administration, infrastructure planning or environmental monitoring. As you all know, this work was mostly carried out by trained staff and professionals based on well defined specifications as well as mapping and quality assurance procedures.
- In around 2005 a series of technological developments resulted in citizens themselves being involved in mapping their immediate environment. Perhaps the most important was the removal of selective availability of GPS signal announced by US President Bill Clinton on the 1st of May 2000., which is usually described as the official birthday of neogeography as it opened the way to produce improved accuracy for simple, low-cost GPS receivers
- Other technological developments such as the introduction of Application Programming Interfaces, enabled users starting to have access to centralised pools of very high-resolution background geographic data including maps, satellite imagery, street photography and building outlines. APIs were relatively easy to use and have made the application development much more accessible, thus enabling a far larger community of people to create, share and mash up (geographic)

information.

- As a result the whole field of geography was kind of reshaped and new terms emerged – one of those is the term ’Neogeography’ which incorporates references to the new socio-technical requirements and characteristics of how people interact with maps.



- According to Turner, who coined the term

  "Neogeography means 'new geography' and consists of a set of techniques and tools that fall outside the realm of traditional GIS. Where historically a professional cartographer might use ArcGIS, talk of Mercator versus Mollweide (MOLVAIDE) projections, and resolve land area disputes, a neogeographer uses a mapping API like Google Maps, talks about GPX versus KML, and geotags his photos to make a map of his summer vacation. Essentially, Neogeography is about people using and creating their own maps, on their own terms and by combining elements of an existing toolset…"

9. LECTURE ACTIVITY: CROWDSOURCING SPATIAL DATA

Identify at least **TWO** spatial crowdsourcing applications which meet ALL crowdsourcing criteria introduced by Enrique Estellés-Arolas & Fernando González-Ladrón-de-Guevara (2012).

"Is that road up there passable? . . . Does it really exist?"

13

10. VOLUNTEERED GEOGRAPHIC INFORMATION (VGI)

"the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. [..] I term this volunteered geographic information (VGI), a special case of the more general Web phenomenon of user-generated content"

Michael Goodchild

Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69: 211–221.

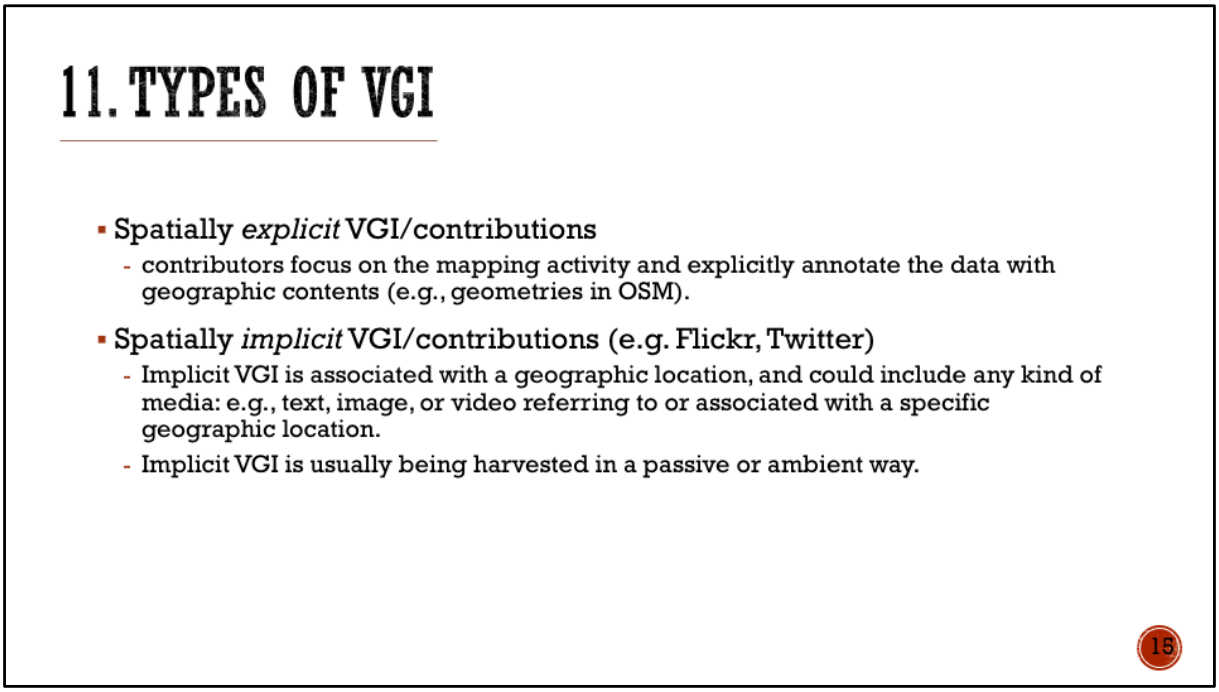
14

VGI is closely related to the concept of crowdsourcing as it is an assertive method of collecting geospatial information from people who are mainly participating in various activities.

The term volunteered geographic information was coined in 2007 by Michael Goodchild (2007) as a subset of user-generated content which carries specific spatial and temporal components. He described VGI as:

> *"the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. [..] I term this volunteered geographic information (VGI), a special case of the more general Web phenomenon of user-generated content".*

This spatial shift, which was as we saw before, fuelled by the neogeography revolution, which has put mapping within the grasp of almost any user who has internet access and the right equipment and it combines technology, social practices and power relationships.



There are two types of VGI contributions which define and shape some of the issues that we will discuss later, such as the wider ethical and legal landscape.

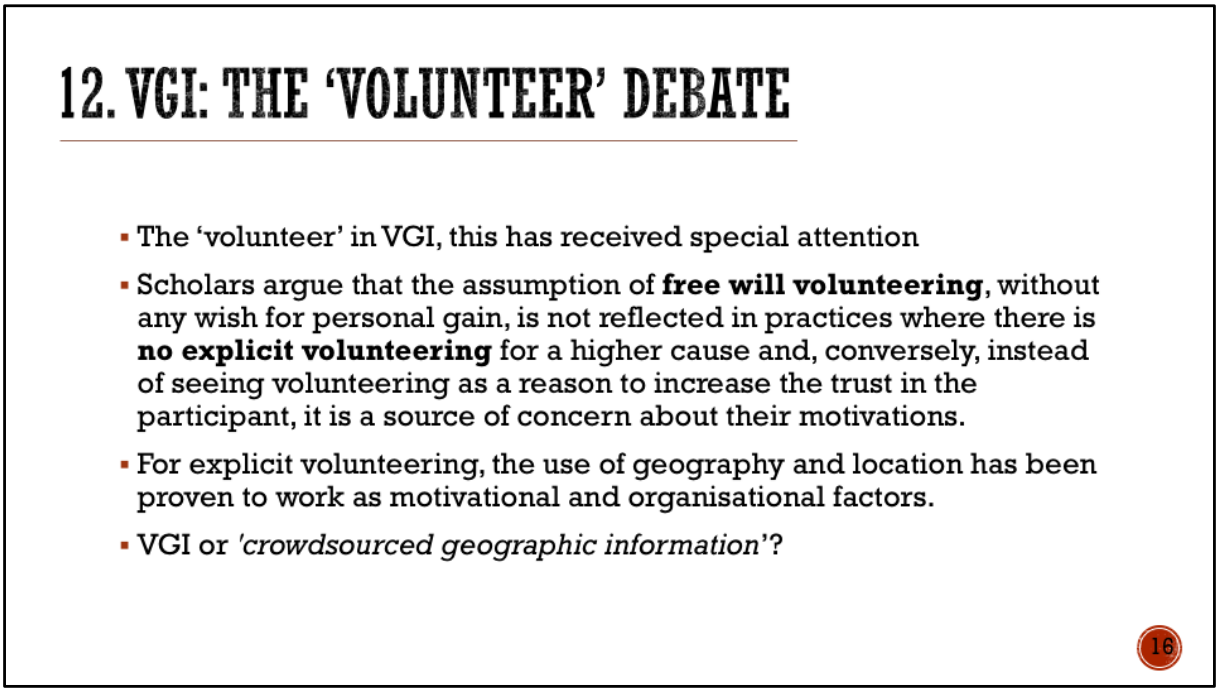
Spatially *explicit* VGI/contributions

-This refers to data collected by contributors who focus on the mapping activity and explicitly annotate the data with geographic contents (e.g., geometries on OpenStreetMap).

Spatially *implicit* VGI/contributions (e.g. those found in Flickr,Twitter)

-These are associated with a geographic location and can be various kinds of media – for example implicit VGI may include text, image, or even a video referring to or which is associated with a specific geographic location. Geotagged microblogs (e.g., Tweets), geotagged images from Flicker, or Wikipedia articles that refer to geographic locations are examples of implicit contributions.

- Implicit geographic contributions may be harvested in a passive way – which is the user might not be actively engaged in the collection of geographic data but the device he uses has the right equipment and can do so while other types of information are being contributed.



An important debate within VGI academic literature is that of the "volunteer. The word volunteering implies some sort of contribution offering, without personal gain or other kind of personal benefit, which may be found in forms of explicit volunteering – for example you will see that in some of the examples in the next slides where people map areas hit by natural disasters and they are motivated by participating in that higher cause or perhaps similarly in collecting environmental monitoring data contributing in that way to environmental sustainability.
But this attribute of free will volunteering according to some scholars is not reflected in other practices where there is **no explicit volunteering** for such higher causes and therefore a series of issues arise around people's motivations. In that respect the geographical component has a phenomenological value.
In the case of geographic information involuntarily produced by individuals, quality might be debated since data are often collected publicly without strict standardization and every user inserts data according to his/her personal background and point of view. This is another reason some scholars argue that the term '*crowdsourced geographic information' is used instead (as opposed to VGI)* to refer to data collected via 'opt-out' agreements which are more open-ended and offer fewer opportunities to control the data collection, and subsequently its quality and assessment.

Again I think I should not that '*crowdsourced geographic information'* can be also a problematic term – as it is in the general crowdsourcing literature as I explained in the previous slides – because it implies this process of consensus production whereby many people will provide and augment information about the same thing which will become more and more accurate thanks to a convergence of information. In that sense, the term 'volunteered' or even 'crowdsourced' as an alternative and more broadly inclusive term for implicit contributions, can be misleading for some types of VGI, but that doesn't mean that these activities do not strictly fall under the VGI umbrella.

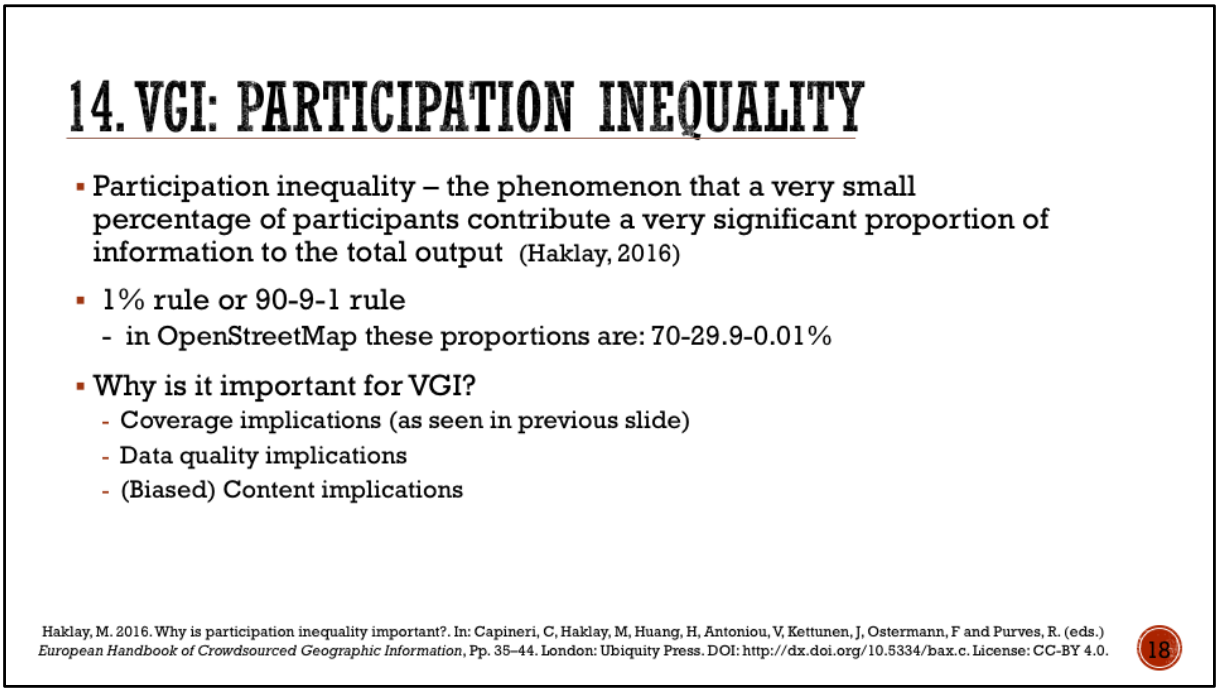
13. VGI: COVERAGE INEQUALITY

- Many case studies show strong uneven geographical patterns of participation in the production of VGI – mainly due to the digital divide, number of contributors etc.

17

Another important aspect of VGI which has been very thoroughly examined in the academic literature is that of the geographic focus of contributions. Many case studies show strong and uneven geographical patterns of participation in the production of VGI – and for any crowdsourced data for that matter - mainly due to the digital divide. This digital divide may be caused by different reasons such as uneven diffusion of the technology but also the ability to work with this technology or benefit from it. Despite the pervasiveness of the internet, recent analysis shows that the practice of producing content is mainly concentrated in the United States, Western Europe, Japan and in some emerging countries in South America and Asia. Also the number of contributions is influenced by where the users are located and their interests that are associated with specific locations. So generally we have much more higher numbers of VGI contributions for urban and populated urban areas - where the technological infrastructure makes it easier to contribute –and less for less populated areas. We should not of course ignore that with VGI platforms such as OpenStreetMap, where people usually map their surrounding environments, these tend to be much more detailed for big urban centres (as the Figures two show) comparted to less populated areas where the contributions are sparse.
Go to OpenstreetMap and try this yourself. Zoom into urban centres and the level of detail and number of contributions is impressive and then check out other parts of

the world as in Africa or even more remote areas (for example the countryside) in other European countries.



One of the most persistent aspects that can be noted in systems which facilitate user-generated content (among them volunteered geographic information and citizen science data) is the inequality in the level of participation that they exhibit.
Haklay (2016) defines participation inequality as a phenomenon where a small percentage of participants contribute to a very significant proportion of information.

- This general behaviour manifests itself in other online contexts such as online discussion forums games as well as applications for ecological observations. In each of these cases, the overwhelming majority of people who use the information or who are registered to the service do not contribute any information to it. The proportion of registered people who do not contribute can reach 90% or even more of the total number of users. Of the remaining participants, the vast majority contribute infrequently or fairly little – and these account for 9% or more of the users. Finally, the last 1% contribute most of the information. This has led to framing the phenomenon as the 90-9-1 rule.
- But why participation inequality is important for VGI?
    - We need to appreciate that data is not only heterogeneous in terms of consistency and coverage; but it is also highly heterogeneous in terms of contribution, which can have far-reaching implications on quality, coverage and content. For example if the majority of the contributions comes from

heavy contributors with significant expertise this can improve confidence in data quality.

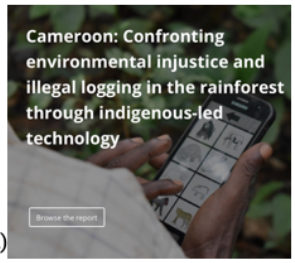
- Also knowing where specific contributors are active (and taking into account what we said in the previous slide about contributors naturally concentrating their effort in mapping their surrounding environments) can also help in quality assurance processes by comparing novice practices to their actions.
- Another aspect of the impact of high contributors is the social evolution of the project. In some projects, high contributors might exhibit abrasive behaviour towards other participants or protect the area in which they operate, by aggressively editing any new information in order to fit their standards.
- Another element to consider is the specific background and interests of high contributors will, by necessity, impact on the type of data that it is is being recorded. This is especially important in VGI projects where the details of what to record are left to the participants. For example, lack of interest in a class of facilities (e.g. wheelchair accessible toilets) will mean that such information will be lacking from the resulting dataset and might shape the activities of other participants.

15. VGI: PRIVACY, ETHICAL AND LEGAL ISSUES

19

16. VGI IMPLICATIONS AND ISSUES TO CONSIDER

- As crowdsourced geographic information becomes more prevalent, spatial data are continuously being collected using spatially aware devices.
- The collection and share of VGI has privacy, legal and ethical implications.
  - VGI can potentially expose sensitive private information (e.g., personal data about users, their living habits and other private information).
  - Or in capturing aerial imagery, geotagged photographs and street-level photography, the faces of users themselves or their houses may be identifiable too (see Street View privacy discussions…)
  - For some projects, (even for explicit) VGI the data collection process may put the users' life at risk (see S. Hoyte's project on the right with Baka people in Cameroon)
- Therefore use, storage, access or integration of these data are often the subject of complex legal and ethical considerations.



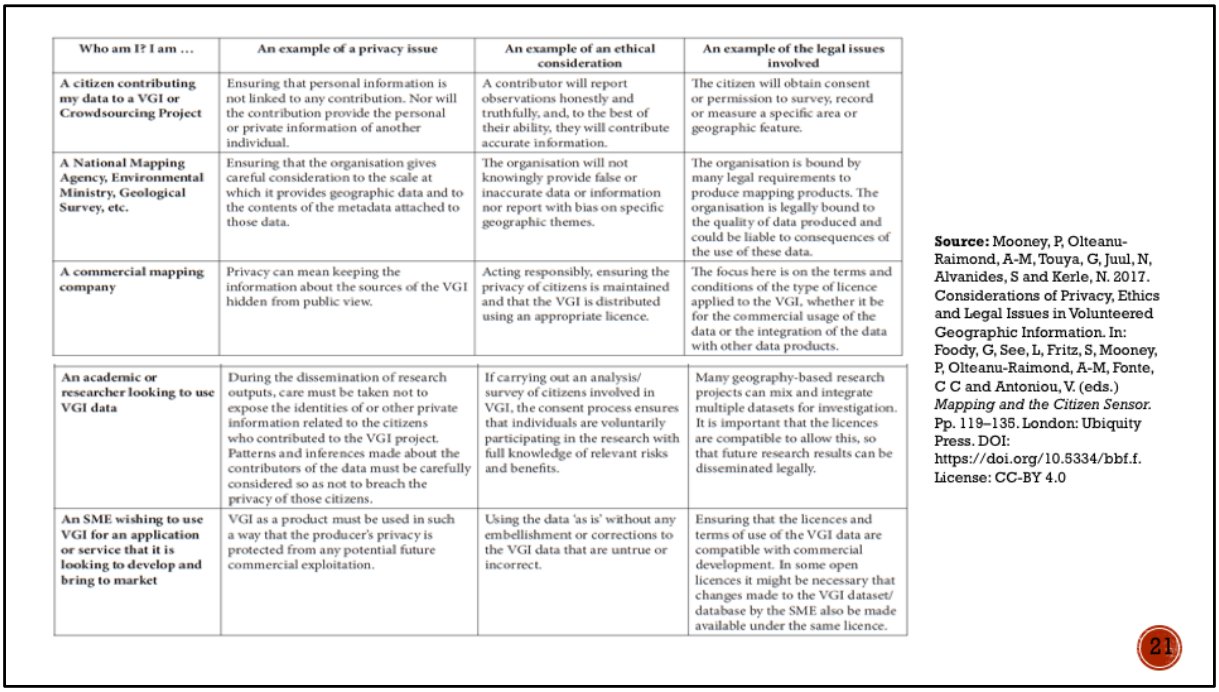
https://minorityrights.org/trends2020/cameroon/

20

- As crowdsourced geographic information becomes more prevalent, spatial data (explicit and implicit) are constantly being collected from citizens, through spatially aware devices such as smartphones, smart devices and other types of sensors.
- The public collection and exchange of geospatial data and information as Volunteered Geographic Information (VGI) involve many privacy, legal and ethical issues. The significance of these issues is magnified further due to the re-distribution and dissemination of these data by third parties such as libraries, online data services, etc.
- There are several examples of these issues to note here. First, VGI can potentially expose sensitive private information (e.g., personal data about living habits or even health conditions of the contributor). Similarly in collecting geotagged photographs privacy information of individuals may be accidentally shared. In this context there is a whole debate around the use of street view and several countries like Germany banned street view from launching in particular countries when it first became available. This is because still many people have substantial concerns with regard to their online privacy while they are often engaged in behaviours that do not adequately reflect their concerns, frequently without them being aware.
- In other examples of VGI, even when these refer to explicit contributions, may put the users' life at risk. For example, In Cameroon Simon Hoyte runs a project with Baka

hunter gatherers who are trying to prevent poaching and illegal logging in their territories. Confronting environmental injustice through the collection of this type of VGI may put the lives of the contributors at a very high risk. I will explain a little bit more about this project in the next few slides.

- From these examples it is clear that the use, storage, access or integration of these data with other data sources are often the subject of complex legal and ethical considerations.

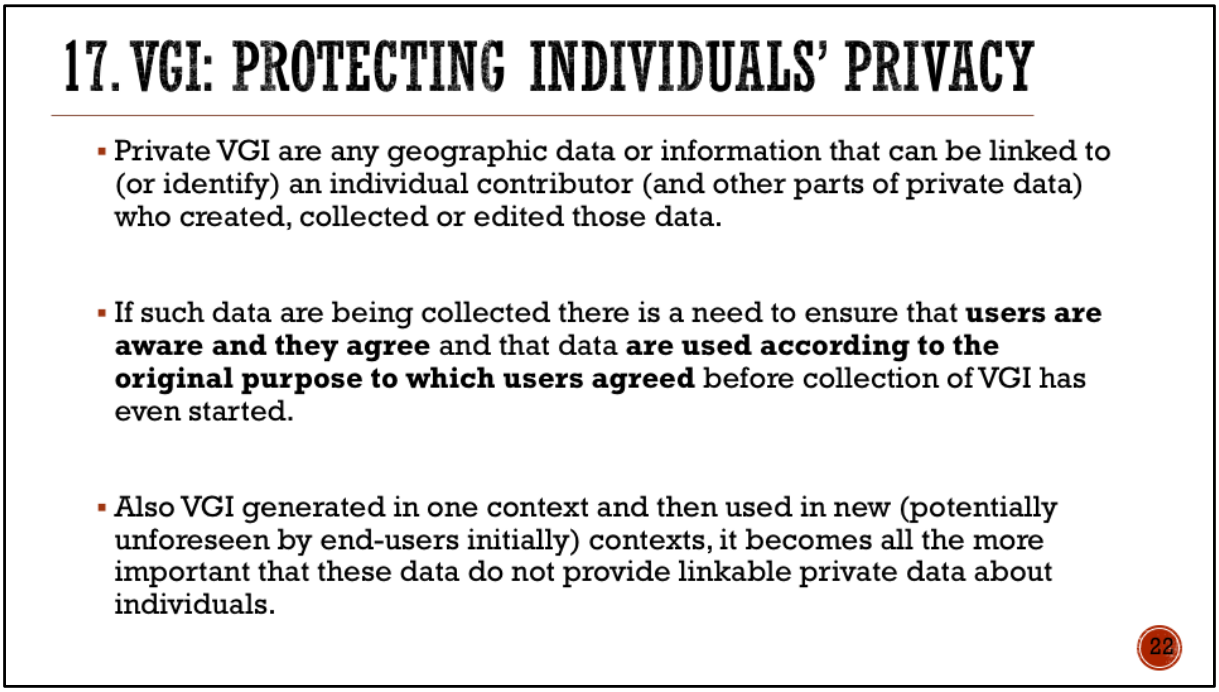


Who am I? I am …	An example of a privacy issue	An example of an ethical consideration	An example of the legal issues involved
A citizen contributing my data to a VGI or Crowdsourcing Project	Ensuring that personal information is not linked to any contribution. Nor will the contribution provide the personal or private information of another individual.	A contributor will report observations honestly and truthfully, and, to the best of their ability, they will contribute accurate information.	The citizen will obtain consent or permission to survey, record or measure a specific area or geographic feature.
A National Mapping Agency, Environmental Ministry, Geological Survey, etc.	Ensuring that the organisation gives careful consideration to the scale at which it provides geographic data and to the contents of the metadata attached to those data.	The organisation will not knowingly provide false or inaccurate data or information nor report with bias on specific geographic themes.	The organisation is bound by many legal requirements to produce mapping products. The organisation is legally bound to the quality of data produced and could be liable to consequences of the use of these data.
A commercial mapping company	Privacy can mean keeping the information about the sources of the VGI hidden from public view.	Acting responsibly, ensuring the privacy of citizens is maintained and that the VGI is distributed using an appropriate licence.	The focus here is on the terms and conditions of the type of licence applied to the VGI, whether it be for the commercial usage of the data or the integration of the data with other data products.
An academic or researcher looking to use VGI data	During the dissemination of research outputs, care must be taken not to expose the identities of or other private information related to the citizens who contributed to the VGI project. Patterns and inferences made about the contributors of the data must be carefully considered so as not to breach the privacy of those citizens.	If carrying out an analysis/ survey of citizens involved in VGI, the consent process ensures that individuals are voluntarily participating in the research with full knowledge of relevant risks and benefits.	Many geography-based research projects can mix and integrate multiple datasets for investigation. It is important that the licences are compatible to allow this, so that future research results can be disseminated legally.
An SME wishing to use VGI for an application or service that it is looking to develop and bring to market	VGI as a product must be used in such a way that the producer's privacy is protected from any potential future commercial exploitation.	Using the data 'as is' without any embellishment or corrections to the VGI data that are untrue or incorrect.	Ensuring that the licences and terms of use of the VGI data are compatible with commercial development. In some open licences it might be necessary that changes made to the VGI dataset/ database by the SME also be made available under the same licence.

Source: Mooney, P, Olteanu-Raimond, A-M, Touya, G, Juul, N, Alvanides, S and Kerle, N. 2017. Considerations of Privacy, Ethics and Legal Issues in Volunteered Geographic Information. In: Foody, G, See, L, Fritz, S, Mooney, P, Olteanu-Raimond, A-M, Fonte, C C and Antoniou, V. (eds.) *Mapping and the Citizen Sensor.* Pp. 119–135. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbf.f. License: CC-BY 4.0

21

Mooney et al (2017) examine legal, privacy and ethical concerns for VGI. In doing so, they provide this table that situates privacy, ethics and legal issues for the principal actors involved in the collection, production and dissemination of VGI, namely citizens, national mapping agencies (NMAs), commercial companies, researchers and other entities such as small and medium-sized enterprises (SMEs). For each category of stakeholders they provide an example of privacy, ethical and legal issues involved.

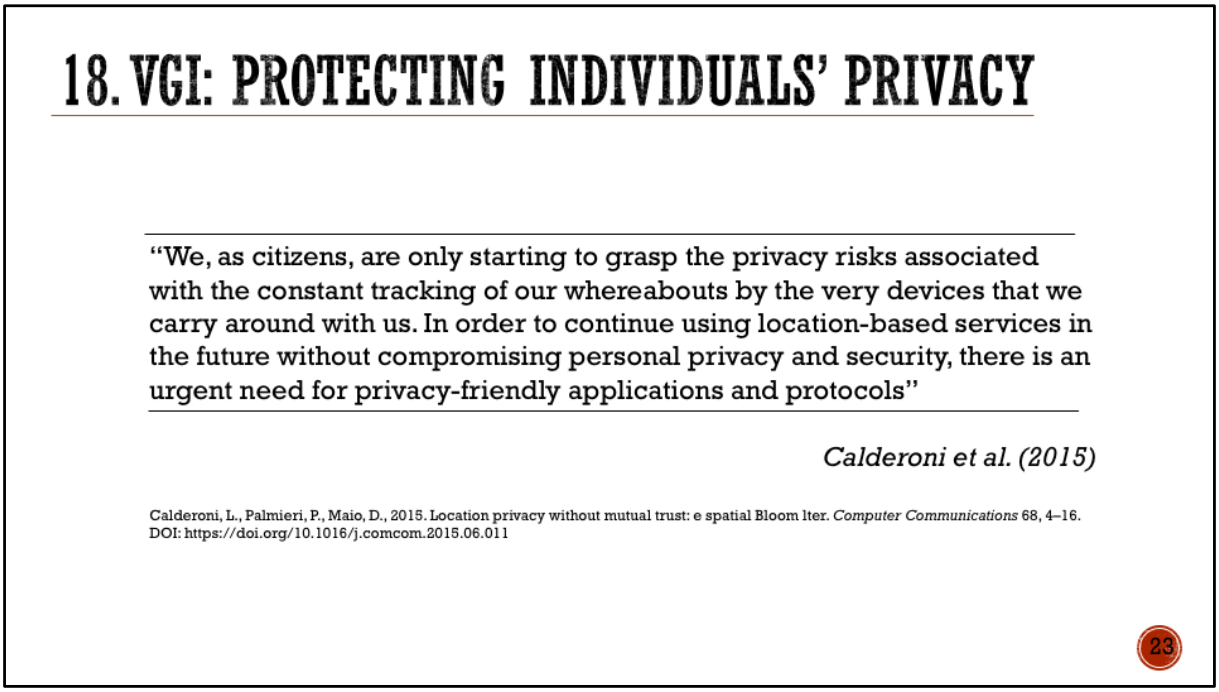


In terms of protecting individual user privacy we need to start by thinking about data privacy.

- Private VGI are any geographic data or information that can be linked to (or which can be used to identify) an individual who created, collected or edited those data, and data which may also reveal other private data of this particular contributor. For example many location data are accurate enough to be bound to one individual or to a small group of individuals, e.g. an office or home, and sometimes these data are even combined with precise times and dates.

- To prevent VGI data being used to violate the privacy of individuals, we need to look at the character of the data and investigate the entire process from the collection of data to the submission of the VGI to data repositories, and then onwards to the use of the data.

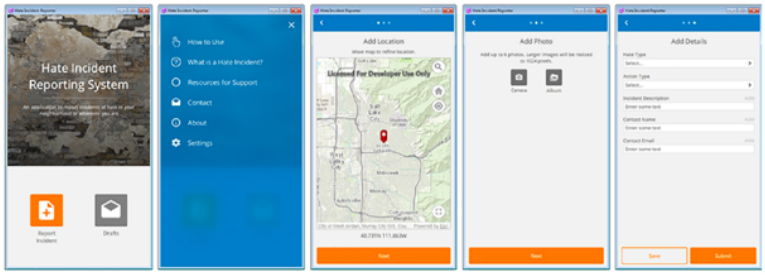
- **The guiding principle of privacy protection is to collect as little private data as possible during all this stages. But** If such data are being collected there is a need to ensure that **users are aware** and that data **are used according to the original purpose to which users agreed** before collecting it.
- We saw when defining open data in the beginning that open data can be re-used and VGI can be also open. It is extremely important to consider that while in one context the data may not expose users' privacy, the same data used in a different context may do so. Therefore it becomes very important that these data do not provide linkable private data about individuals.



In 2015 Calderoni noted that "*We, as citizens, are only starting to grasp the privacy risks associated with the constant tracking of our whereabouts by the very devices that we carry around with us. In order to continue using location-based services in the future without compromising personal privacy and security, there is an urgent need for privacy-friendly applications and protocols".*

Therefore, in order to continue using location-based services in the future without compromising personal privacy and security, we need to be much more critical and identify ways to collect and share VGI without compromising users privacy.

19. VGI: PROTECTING 'SENSITIVE' VGI CONTRIBUTIONS



Source: Nicolosi et al., 2020. Crowdsourcing sensitive VGI: Constructing the hate incident reporting system. Digital Geography and Society, 1, https://doi.org/10.1016/j.diggeo.2020.100003

24

-This is an example from a research paper by Nicolosi et al (2020) where VGI is used to report hate crime incidents in the USA and which highlights another key element of **protecting sensitive VGI contributions**, similar to the Cameroon project I mentioned before.

-The idea comes from the fact that "general hate and hate incident data would be very valuable to governments, as knowledge of hate likely falls short. Understanding the nature of hate within a community would allow for more effective planning and outreach efforts, and might ultimately reduce the number of hate crimes.

-So the authors launched the Hate Incident Reporting System (HIRS in short) to report these incidents eliminating at the same time any political and police-level (institutional) barriers to hate crime reporting. Yet by eliminating these barriers what happens with several VGI projects is that although they may have a tremendous value and great impact, yet if they are not properly designed in terms of privacy considerations they could make contributors extremely vulnerable.

20. VGI: PROTECTING 'SENSITIVE' VGI CONTRIBUTIONS

https://www.theguardian.com/education/2020/nov/16/the-rise-of-citizen-science-can-the-public-help-solve-our-biggest-problems

25

- Another example to demonstrate the importance of protecting sensitive VG Information comes from one of the citizen science projects in the extreme citizen science group, which actually features in an article by The Guardian from a few days ago (link on the slide).
- So, in Kenya, Sapelli is being utilized since early 2019, with Maasai warrior communities. One of the greatest threats the local communities face is the loss of their Traditional Ecological Kknowledge and the increased deforestation in the Maasai Mara National Reserve. Sapelli – which a data collection tool - is used to assist them in collecting and recording TEK related to indigenous plants. Within a few hours after the initiative was launched individuals gathered over a hundred data items and since then they've collected thousands of points with information about the medicinal and other properties of local indigenous flora.
- Here an important elements to consider is that TEK may contain very sensitive community knowledge and it may include data that should NOT in any way be shared with anyone outside the project or a threat may be posed to the community itself. These privacy issues needs to be fully understood and local people together with the data they collect should have complete control of how their data are being utilized while taking extreme care of using technologies that will not expose these communities by accidentally opening up these data sources in the future so that outsiders can gain access to them.
- There might be also additional implications in similar VGI contexts. For example, suppose there is a crowdsourcing or VGI campaign in the area of biodiversity and a very rare or precious plant species is found and geolocated. To protect this species (and potentially its habitat), this information might need to be kept private.

More information about this project can be found here https://www.uclpress.co.uk/products/125702

- Chapter 12

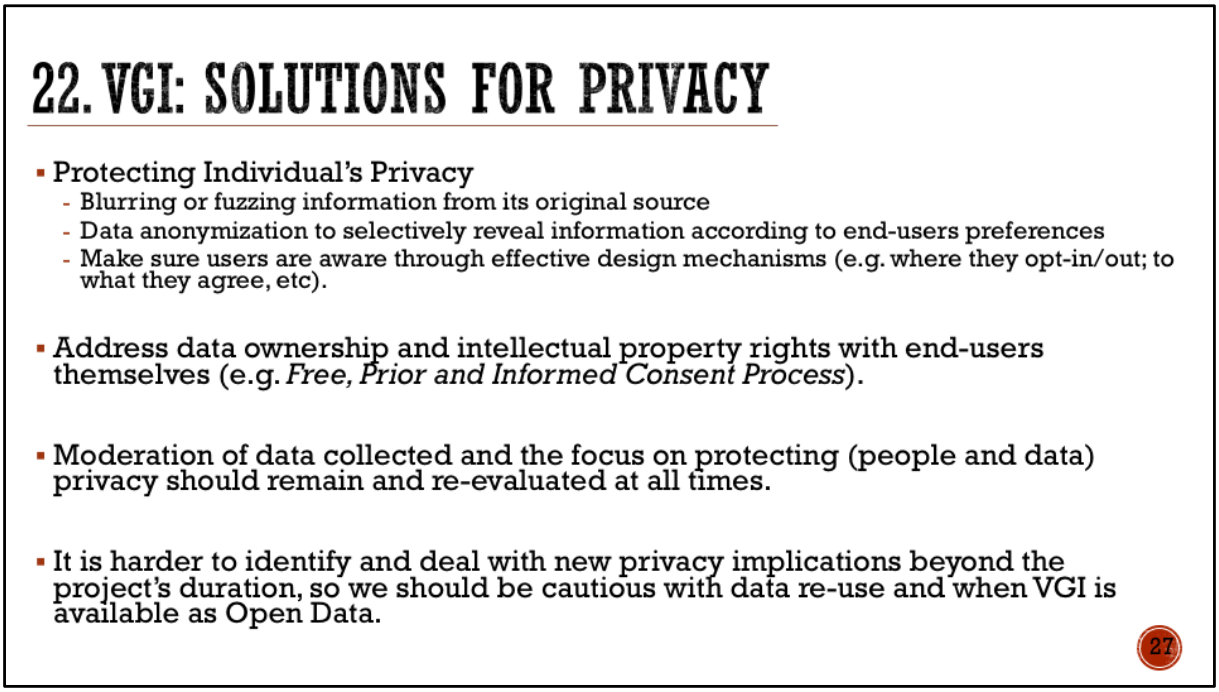
21. VGI: PROTECTING 'SENSITIVE' VGI CONTRIBUTIONS

26

- In Africa, local perceptions and knowledge of environments are still too often considered irrelevant in conservation, despite these communities inhabiting key conservation areas and interacting daily with the fauna and flora so highly prized by the international community. Led by ideas of socio-ecological systems, and damaged by hard-hitting allegations of human rights abuses, organisations and governments are under increasing pressure to adopt community-centred approaches to protect biodiversity. Accordingly, community-based natural resource management projects are increasing around the world, with the application of environmental monitoring by local and indigenous peoples becoming more accepted in mainstream science and policy.
- So VGI in this context, as a form of community-based monitoring, has emerged as a key methodology to both empower community members to collect locally relevant, geo-located data themselves and provide the quantity of local data required to address heightening environmental change.
- Since August 2016, the extreme citizen science group (UCL, Geography) works with Baka communities in Cameroon. Cameroon harbours extraordinary biological ; It is one of the last places on earth where such a diversity of megafauna exists in the wild, but it is being rapidly depleted by the illegal wildlife trade and extractive industries. Many Baka communities feel a great sense of injustice towards external

wildlife traffickers pillaging forest resources, and consistently express a desire to be involved in tackling such activity and at the moment they collect VGI in a geographical area of roughly 953 km².

- In the dense rainforests of southern Cameroon, the collection of illegal wildlife trade and animal monitoring data by indigenous and local communities may be the only viable mechanism to obtain the illegal wildlife trade and ecological information necessary for effective forest management in the future.
- This can be extremely sensitive data and also has significant implications in terms of potentially exposing the contributor's details which could be perhaps used to identify them and in that case put their lives at risk.
- Communities expressed their concerns about this and they asked us to ensure that the application and the data cannot be accessed by outsiders, or if it accidentally does, individuals who collected these data cannot be identified. For example, one person said :'We can be involved in such an activity, but anonymity is of intense importance.' Together with an anonymous ID system additional password protection features are used to lock the devices which are used to collect the data and which cannot be therefore accessed by community outsiders.



So what are some solutions at the moment for protecting privacy in the VGI context?

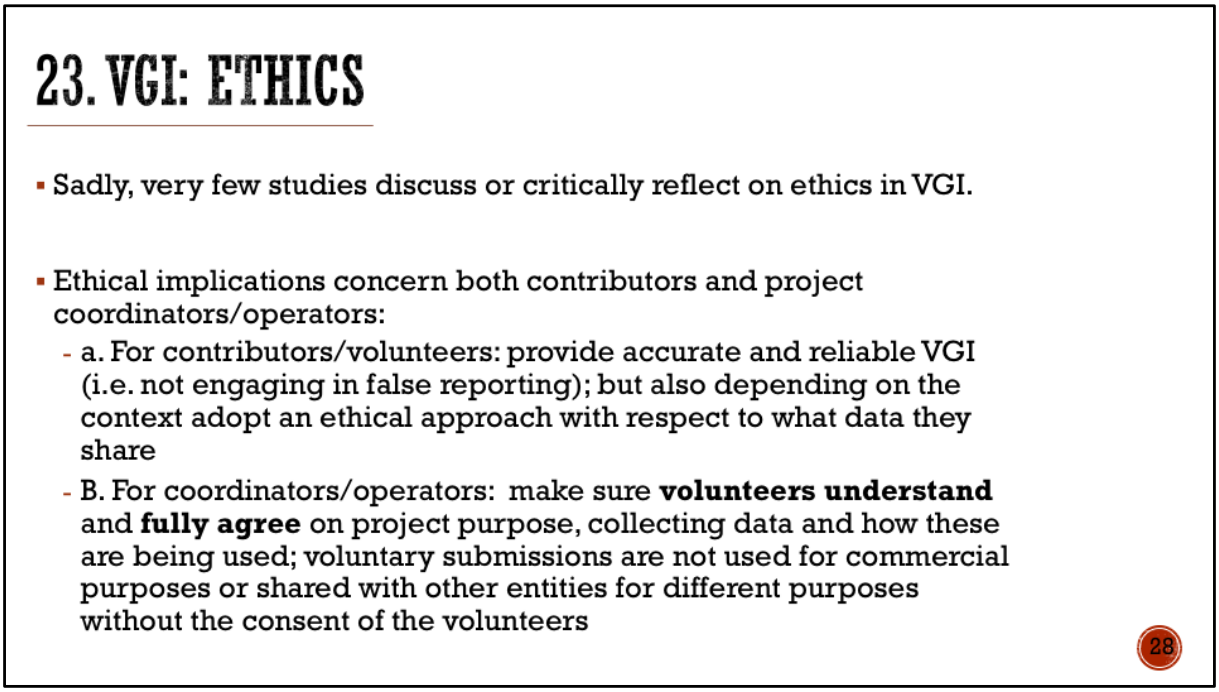
Increasingly sophisticated privacy-preserving data mining techniques are being studied and need to be further developed. These approaches aim to achieve appropriate levels of anonymity by means of controlled transformation of data and/or patterns with limited distortion, to avoid the undesired side effects on privacy while preserving the possibility of discovering useful patterns and trends.
Some solutions for protecting individuals' privacy that you are already aware of include:

- Blurring or fuzzing information from its original source;
- Data anonymization to selectively reveal information according to end-users preferences (something which we also used in the project in Cameroon;
- It is of equal importance in terms of protecting privacy to design tools and mechanisms that make sure that users are aware of any privacy implications (for example where they can opt-in/out; to what they have agreed, etc).

- Another way- and especially in terms of protecting sensitive VGI data - is to address data ownership and intellectual property rights with end-users themselves

(we do this in the extreme citizen science group using the Free, Prior and Informed Consent Process – I will talk more about it in the next lecture).

- Another way to protect individuals privacy is through moderation. Moderation of the data collected and the focus on protecting people and data privacy should remain and re-evaluated at all times during the project's duration as needs change some times during a project's lifecycle.
- But what happens beyond the project duration? It is harder to identify and deal with new privacy implications that exist only after the project's duration, so we should be extremely cautious to critically reflect on potential future issues and in terms of promoting data re-use and when VGI becomes available as Open Data.



Ethical implications concern both contributors and project coordinators/operators:

-a. For contributors/volunteers; an ethical consideration is to provide accurate and reliable VGI (i.e. not engaging in false reporting); but also depending on the context adopt an ethical approach with respect to what data they share;

-B. For coordinators/operators: other ethical concerns include making sure **volunteers understand** and **fully agree** on project purpose, collecting data and how these are being used; voluntary submissions are not used for commercial purposes or shared with other entities for different purposes without the consent of the volunteers

24. VGI: ETHICS AND SOME RECOMMENDATIONS

- How can we improve our understandings of ethical issues in VGI and therefore create the guidelines or frameworks to account for those and ensure that VGI is ethically responsible?

- "Respect individuals, commit to nondisclosure of participants' identities, minimize potential harm, ensure the benefits and burdens are fairly distributed, volunteers are aware of purpose of project and they can withdraw at any point"

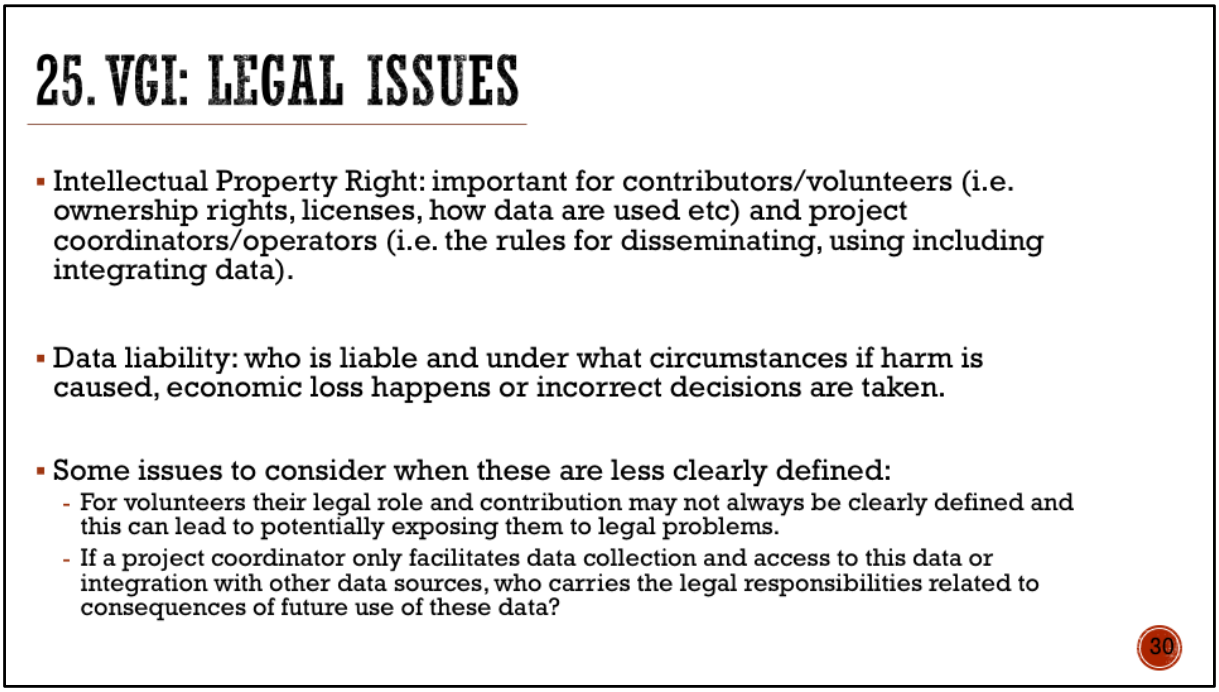
Hartter et al., 2013

Hartter, J., Ryan, S.J., MacKenzie, A.C., Parker, J.N., Strasser, C.A., 2013. Spatially Explicit Data: Stewardship and Ethical Challenges in Science. *PLoS Biology* 11(9). DOI: https://doi.org/10.1371/journal.pbio.1001634

- Follow ethics approval funding review boards and research authorities standards and processes.

29

- A logical question from previous slides: is how can we improve our understandings of ethical issues in VGI and therefore create the guidelines or frameworks to account for those and ensure that VGI is used in ethically responsible ways?
- Hartter et al. (2013) mention: *Respect individuals, commit to nondisclosure of participants' identities, minimize potential harm, ensure the benefits and burdens are fairly distributed, volunteers are aware of purpose of project and they can withdraw at any point.*
- Also, it is extremely important within the context of VGI to follow ethics approval funding review boards and research authorities standards and processes.



What are the most common legal issues that surround VGI?

The first one, refers to Intellectual Property rights. These are important and need to be clearly defined for both volunteers and the project operators. For the former issues that need to be understood and subsequently addressed include:

- who has the ownership data rights and how data are used or will be used in the future? For some VGI ownership data rights are passed to the project operator but for others like for example in the Maasai project from one of the previous slides, these remain with the people who collect the data.
- Particularly important for the project coordinators are the rules for disseminating the data including their integration with other data sources. When VGI is collected and subsequently disseminated, it can be reused, displayed, integrated and transformed in a myriad of ways.

We discussed how open data, public bodies and other organisations are adapting a policy of open data. Generally there are two concepts of open data: one concept means that 'data and content can be freely used, modified, and shared by anyone for any purpose' and the other involves open source licensing applied on software. These need to be considered with cautiousness but they are also likely options for VGI data.

- Data liability is another legal concern. Concerning liability, the main question is that of who is liable and under what circumstances if harm is caused, economic loss happens or incorrect decisions are taken. This is very important for VGI as it is also linked to concerns around data quality. Therefore it also refers to both volunteers and project coordinators. Some issues to consider within this context include:
- - the volunteers' legal role which is rarely clearly defined and hence there is a danger to expose them to legal problems;
- Usually projects define legal issues prior to and during the project duration, but the data collected exists after the project completion. With an increasing number of data being re-used we also need to think of the responsibilities and consequences of future use of these datasets in other contexts.

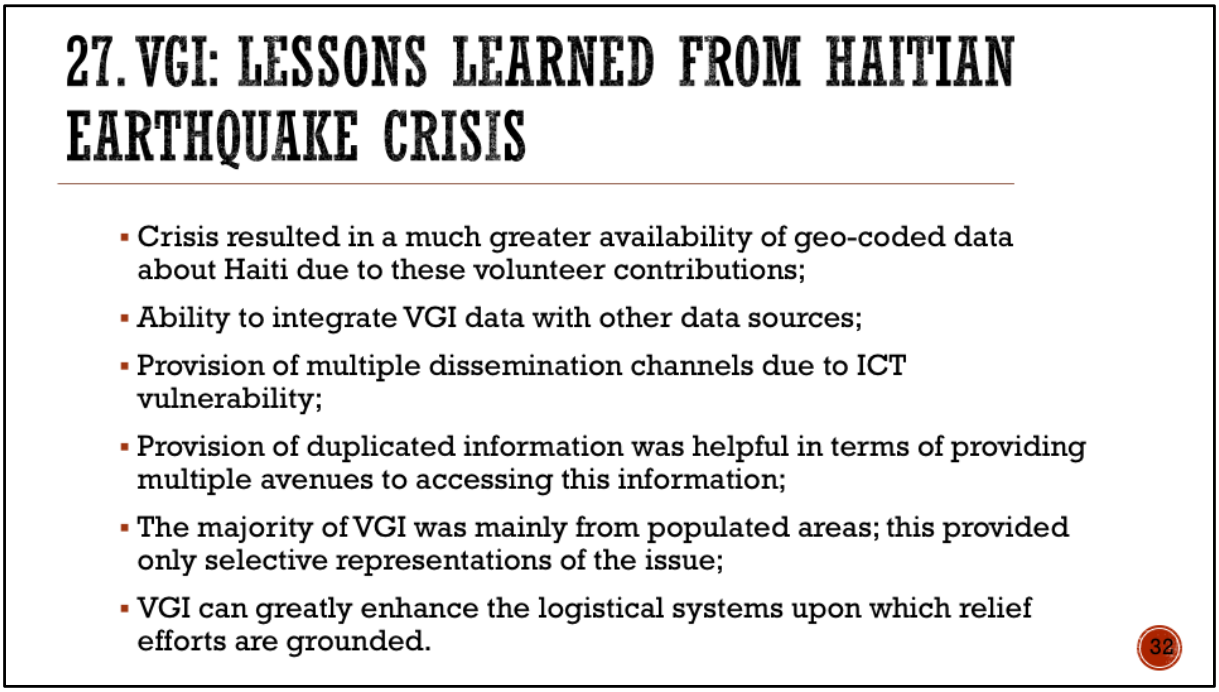
a. Google Satellite Imager Before and After the earthquake; **b.** Twitter feed; **c.** Ushahidi platform reporting casualties; **d.** OSM before and after the earthquake

Source: Zook et al., 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. World Medical & Health Policy, 1, 2.

31

The nest slides review some examples of Volunteered Geographic Information and the notes provide background information and lessons learned - as highlighted by the people who run the projects. Relevant references are provided on all slides.
So the first one is the use of VGI for disaster relief, where VGI has been successfully used in the last few years - perhaps the Haitian earthquake is the most highly cited case study for this type of VGI.
When the magnitude 7.0 earthquake struck Haiti on January 12, 2010, there was an immediate need for maps. Emergency responders had to know where the people most in need were located and how to get assistance and relief to them. Large parts of Haiti and its capital lacked adequate coverage in the standard web mapping services (e.g., Bing Maps and Google Maps) that people in most of the developed world have grown accustomed to using. It should be noted that even some of the most fundamental informational needs, like detailed roadmaps and locations of critical assets, these were not available. Post-earthquake, the demand for spatial information and online maps increased tremendously and, given the urgency of relief operations, the ability to crowdsource the data collection process became particularly important. In this case Google, DigitalGlobe, and GeoEye worked together to get high-quality satellite imagery of post-earthquake Haiti collected, processed, and made freely available within 24 hours of the disaster. This significantly helped with

the coordination of emergency relief and aid services in Haiti. Social networks, as shown on the slide, have been also put in place to mobilize the tech community.
> The third image from the left on the slide shows Ushahidi which followed a very different model of crowdsourcing.
(Note: For those who are not aware of Ushahidi, it allows its users to submit reports through SMS, MMS, or an online interface. Text-based reports are then geo-tagged to a particular location within an interactive map.).
The U.S. State Department further assisted with the geo-location of some messages so that more reliable information could be routed to the Red Cross and U.S. Coast Guard. And finally OpenStreetMap (OSM) volunteers from around the world downloaded satellite images (some already freely available and some donated by Yahoo and Google) in order to trace and record the outlines of streets, buildings, and other places of interest. In the last picture you can see pre and post earthquake digital information available through OSM.



- The first observation was that the crisis clearly resulted in a much greater availability of geo-coded data about Haiti
- Another lesson was the the key role of aggregation of crowdsourced data within disaster response. Otherwise data cannot be leveraged or combined with other data sets to provide the maximum benefit for relief efforts.

- The importance of providing multiple channels by which data can be shared. As ICT infrastructure is a vulnerable and, in many cases, sparse, disasters can destroy the capabilities (particularly Internet connections) within the affected region. Ushahidi's use of SMS highlights the benefits that come from using such a multi-platform approach.

- Although in other cases duplicates are usually problematic here the Provision of duplicated information was helpful in terms of providing multiple avenues to accessing this information
- There was a Geographic inequality or participation inequality issue; The majority of VGI was mainly from populated areas; this provided only selective representations of the issue
- Last, Zook et al. (2010) conclude that VGI can greatly enhance the logistical systems upon which relief efforts are grounded.

28. VGI EXAMPLES: 2 SOIL AND WATER CONSERVATION

Source: Werts et al. 2012 An Integrated WebGIS Framework for Volunteered Geographic Information and Social Media in Soil and Water Conservation. Environmental Management, 49, 816–832.

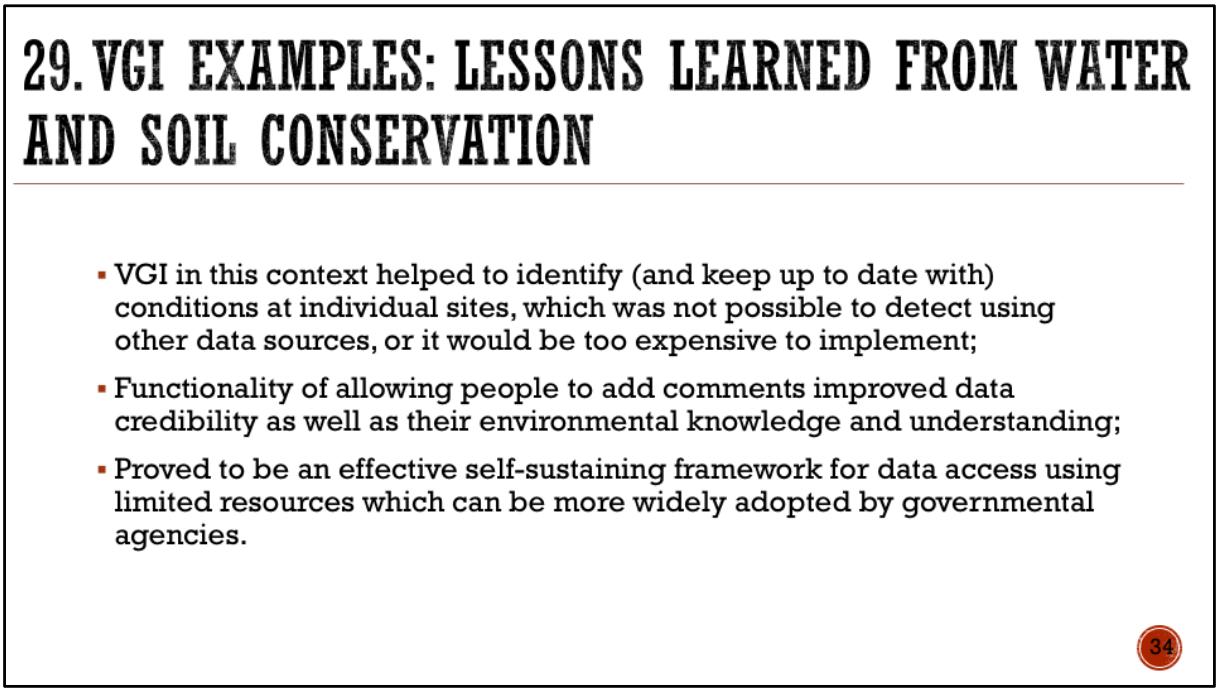
33

There are so many examples of using VGI in the contexts of biodiversity, conservation and environmental management. The idea comes from evidence that local communities and individuals are generally most interested in mapping their immediate surroundings and often make fewer mistakes and can derive higher quality and more frequently updated results than professionally mapped public and private map data layers. This local knowledge can provide timely and relevant information that may not be available through other forms of data collection.

In this example, the focus is on soil and water conservation.

A little bit of background information:
following the burst of the housing bubble, an unprecedented number of residential construction sites in upstate South Carolina have been left unfinished, causing soil erosion and sediment pollution into the streams. These abandoned construction sites are likely found throughout the south eastern United States, but this issue remains largely undocumented. This is an example of an urgent topical environmental issue where greater public involvement may impact the management of these sites to limit future sediment pollution through greater public awareness and changes in laws and policy.

The screenshots we are looking at are from the web-based geographical information system called AbandonedDevelopments.com which was deployed in March 2011 and which focuses on collecting and disseminating data pertaining to abandoned residential construction sites and their potential for sediment pollution in the region of South Carolina and its surrounding areas.
The system displays high quality satellite imagery, sample photos and sites from the original area; and it allows visitors to submit the location as well as pictures and comments about their own subdivisions or sites they have identified. It also allows users to view and comment on other users' submissions.



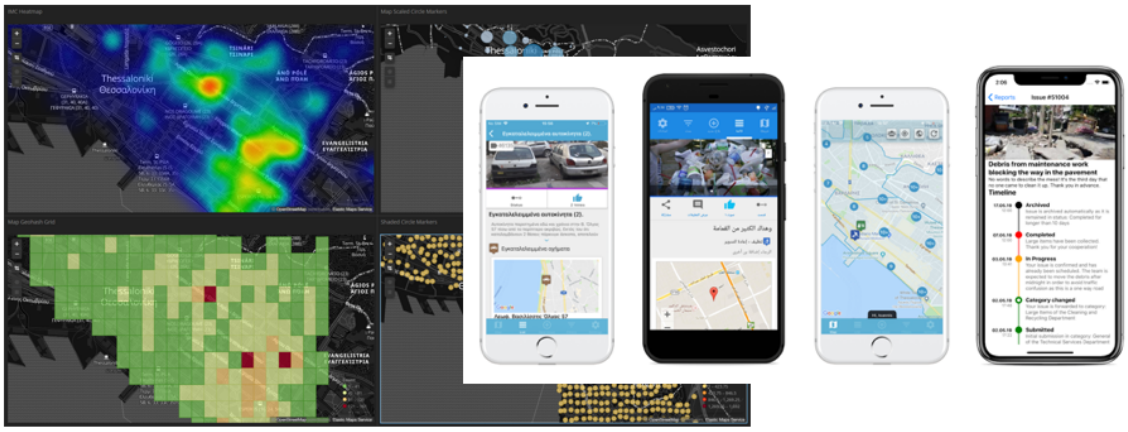
That was a much smaller in scale VGI app and very much focused on a very specific geographical area. Yet the authors still highlight some important lessons learned.

First, they proved that Volunteered geographic information provides a solution for tackling soil and water conservation issues. Locations of abandoned developments were able to be identified through traditional GIS analysis **however,** conditions at individual sites are constantly changing creating the need for a constantly updating sensor network that is too expensive and at too fine of a scale to be implemented with current remote sensing technology. VGI provided an excellent solution to this problem.

By enabling comments on VGI submissions, the credibility of the data increased along with public understanding and retention of knowledge, as the authors explain.

With reduced government funded personnel, the burden of data collection and environmental monitoring may shift more to volunteers. Hence the authors predict that similar approaches will need to be soon adopted by governmental agencies –as we will see also in the next example!

30. VGI EXAMPLES: 3 LOCAL GOVERNMENT FOR REPORTING LOW-EMERGENCY ISSUES (IMC APP)



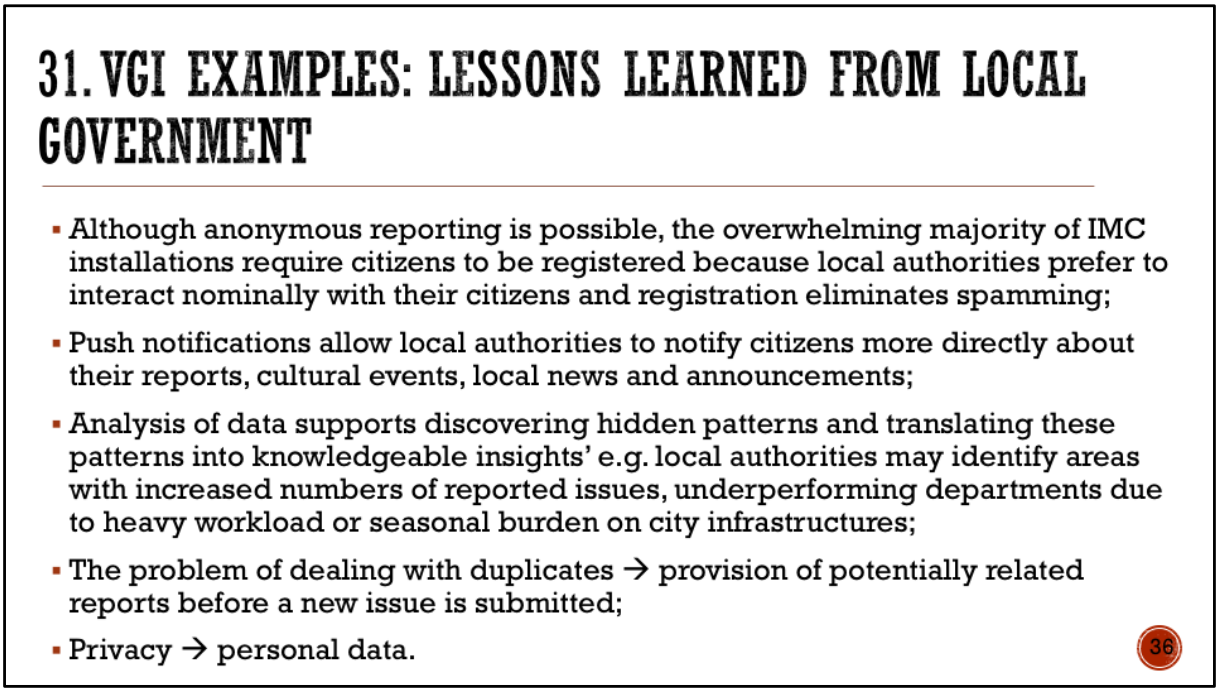
Source: Tsampoulatidis, I. et al (2021) Geographic citizen science in citizen–government communication and collaboration: Lessons from the ImproveMyCity application. In: Skarlatidou, A. and Haklay, M. (eds) Geographic Citizen Science Design: No one left behind. London: UCL Press (https://www.uclpress.co.uk/products/125702).

35

Improve my City aims to promote a participatory culture in local communities. It is an open-source scalable software solution which enables citizens to report non-emergency local issues, such as potholes, blocked bike lanes, broken sidewalks, and so on within their community. The reported issues are automatically routed to the appropriate local authority department which monitors, manages and schedules immediate action.
This voluntary engagement of citizens and the collection of digital geographic information first of all data opens up new forms of interaction between citizens and their government. The adoption of a non-emergency reporting tool such as IMC allows governments to improve their service delivery and accountability and also encourages citizens to be more engaged and play a more explicit role in becoming 'the eyes and ears' of their local authorities.

The adoption of an open and transparent VGI solutions like this one in the governmental context, brings usually some challenges for governments and comes with some political concerns. This is something the application operators comment on but also something I personally experienced in a different project and while working on a VGI tool with three local authorities in Southwark, London, and Torino

and San Dona di Piave in Italy. Local authority politicians, during personal communication with the IMC operators during the early stages of its adoption, expressed concerns about opening up the process to public scrutiny in case they were not able to satisfy citizen requests. However, evidence from other studies suggests that citizens' perceptions about their governments and their willingness to engage can be reshaped and enhanced if the government's operational transparency is promoted using similar VGI tools.

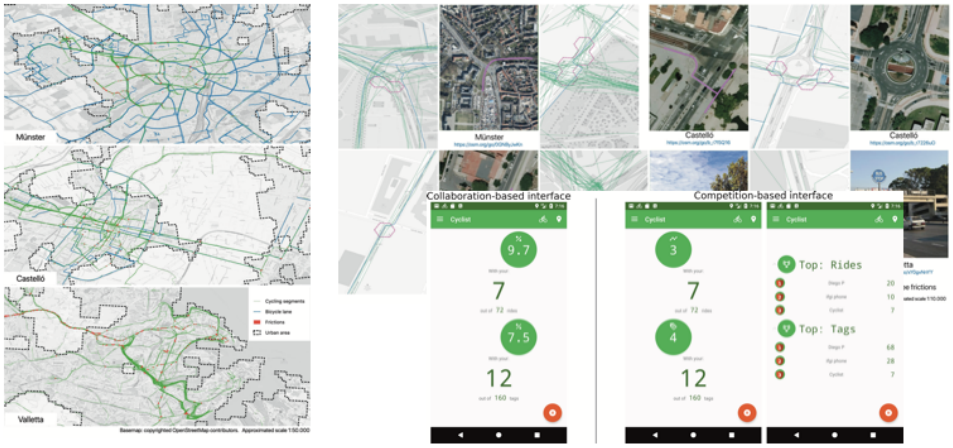


The first lesson learned here is concerned with registration. Although anonymous reporting is possible, the overwhelming majority of IMC installations require citizens to be registered because local authorities prefer to interact nominally with their citizens and registration eliminates spamming; The second lesson learned with a design feature/functionality: i.e. the authors mention that "Push notifications", which were used for local authorities to notify citizens more directly about their reports (which is a brilliant example of two-way communication in the e-government context), but also about other things such as cultural events, local news and announcements; The third concerns the analysis of VGI data. This analysis here provided a great tool for discovering hidden patters in the overall infrastructure of the city and assessing their performance (for example in terms of identifying underperforming departments).

Duplicates can be a significant issue for some VGI projects, but this is not always the case. In ImproveMyCity every reported problem is used to directly notify the relevant authority department - a report is created until the problem is resolved and therefore avoiding duplicated submissions is fundamental. The developers therefore in order to deal with this issue designed the app in a way that during the submission of a new

problem, users are prompted to fetch from the server all issues in the same category within the last 10 days from a radius of about 20 meters and return a list of potential problems reported at the same location; the list contains the title, part of the description and a photo. Greyed-out markers of the suggested issues appear on the input map to denote potentially related reports. The citizen can then follow the link to an existing issue and instead of re-submitting it, simply give a positive vote to show their support.
Last but not least the issue of privacy and protecting people's personal information is very important in this example. Personal data are secured and kept confidential and also photos containing sensitive information are censored through moderation. `There is also moderation for issues that directly or indirectly refer to physical persons or legal entities.

32. VGI EXAMPLES: 3 TRANSPORT PLANNING



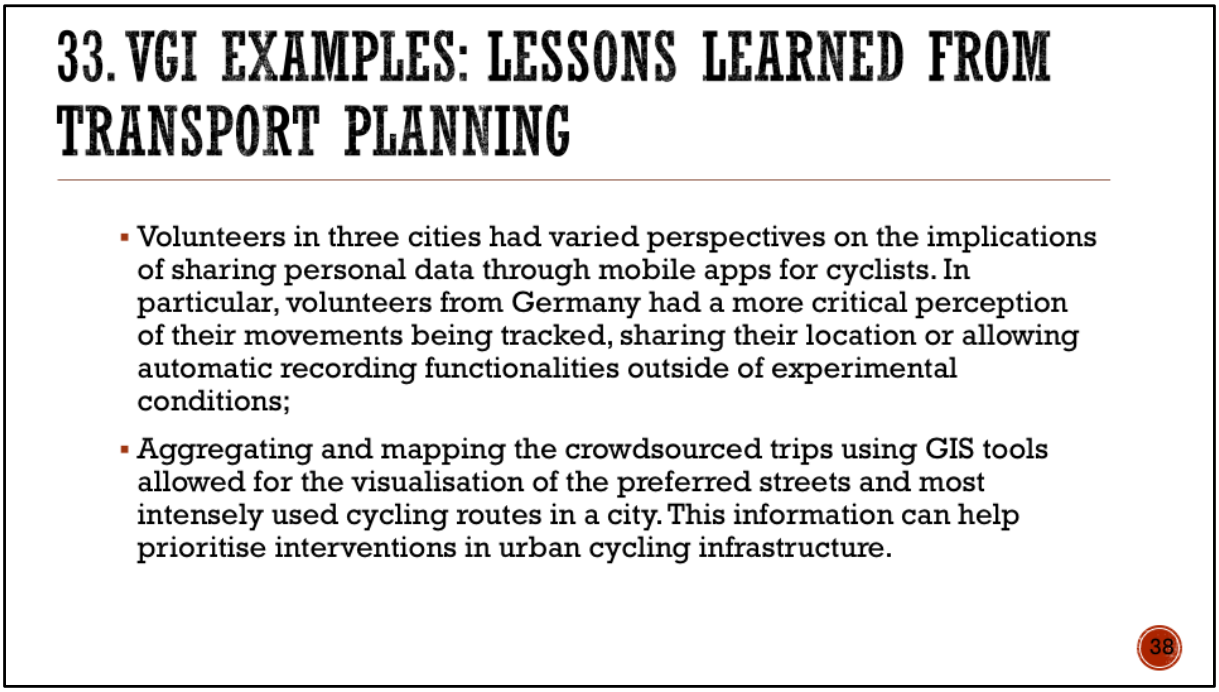
Source : Pajarito et al., (2021). Path of least resistance: Using geo-games and crowdsourced data to map cycling frictions .In: Skarlatidou, A. and Haklay, M. (eds) Geographic Citizen Science Design: No one left behind. London: UCL Press (https://www.uclpress.co.uk/products/125702).

37

Urban cycling is an alternative mode of transport which is promoted by cities worldwide in order to reduce congestion, pollution and increase citizens' physical activity. Cycling data, such as information about the cycling modal share, preferred routes and the main constraints or frictions faced during cycling, can be used as an evidence base for urban planning, cycling infrastructure design, cycling advocacy campaigns, promotion of alternative commuting, and the assessment of impacts and benefits of cycling planning and promotion. Despite the benefits, city managers and citizens have limited access to cycling data usually, and hence different strategies to produce better cycling data have emerged in recent years, which includes the use of smartphones to collect and share geospatial cycling data involving citizens directly. Generally in the context of mobility and transport, volunteered geographic information (VGI) has become a valuable resource of information.

The screenshots arefrom the Cyclist GEO-C mobile app where volunteers share their location data and trip movements to support cycling research. Cyclists use the app to collect information about their trip and through semantic tags they describe their experience. Users can choose up to 30 predefined tags to describe each trip and, by analysing these tags, researchers can also understand how users perceive the cycling environment.

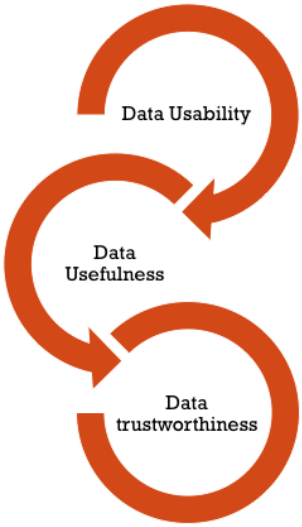
This particular example comes from a testing experiment of the Cyclist Geo-C app in Münster in Germany, Castelló in Spain and the urban area around Valletta in Malta; what it is interesting is that on top of understanding how people use the app the three cities have different cycling cultures and social and geographic contexts.



A particularly interesting lesson learned here is that privacy concerns associated with the collection of implicit VGI were very different in the three cultures where data collection took place.
For example volunteers in Germany reported a strong concern for sharing this data (i.e. personal trips and locations), and they agreed to do that only for the purposes of this experiment.
On the other hand, volunteers in Spain and Malta expressed their intentions to share their trips recorded through third-party cycling apps to either cover a more extended period of time or provide additional insights for the analysis. Some volunteers even suggested an app feature for automatically recording users' activity instead of manually starting and stopping recording every trip (which was a feature initially introduce to make users more privacy-aware and in control of it), and which was something that other users considered intrusive.

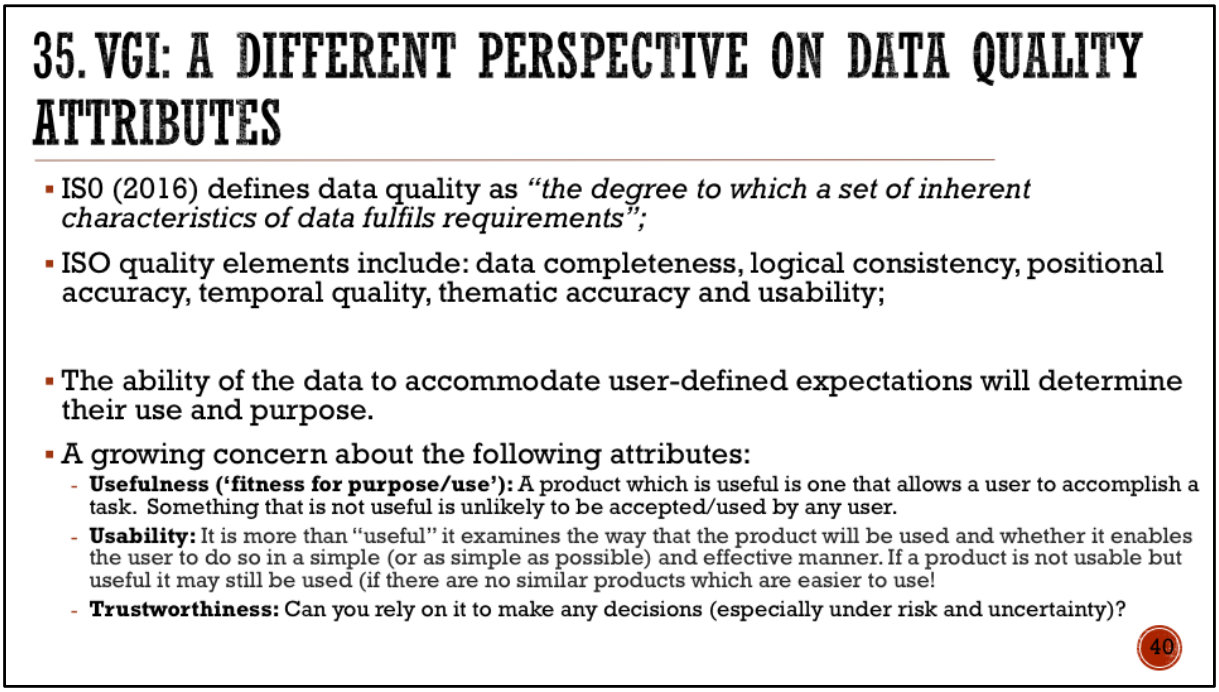
Also, aggregating and mapping the crowdsourced trips using GIS tools allowed for the visualisation of the preferred streets and most intensely used cycling routes in a city. The final maps also showed extra features such as trip origin, destination and the bicycle lane network, the urban area limits, surrounding city infrastructure and landmarks. This analysis was a very useful tool for participants and local authorities to

better understand the city’s cycling environment.



34. VGI: A DIFFERENT PERSPECTIVE ON DATA QUALITY ATTRIBUTES

39



Data quality is extensively discussed in the academic literature with respect to VGI as it is widely discussed for spatial data more broadly.
ISO defines data quality as *"the degree to which a set of inherent characteristics of data fulfils requirements".* The most typical quality elements include: data completeness, logical consistency, positional accuracy, temporal quality, thematic accuracy and usability.

What it is important to understand here is that these data quality elements are evaluated in different ways by different people and within different contexts. Also with the whole context of neogeography (i.e. non-professionals being able to create, share and disseminate but also USE geographic information in so many different ways) the issues of quality become even more important. One can argue that information about thematic accuracy or temporal quality etc can be of little value to a non-professional if that person does not really know or understand what these data quality attributes really mean and how to assess them.

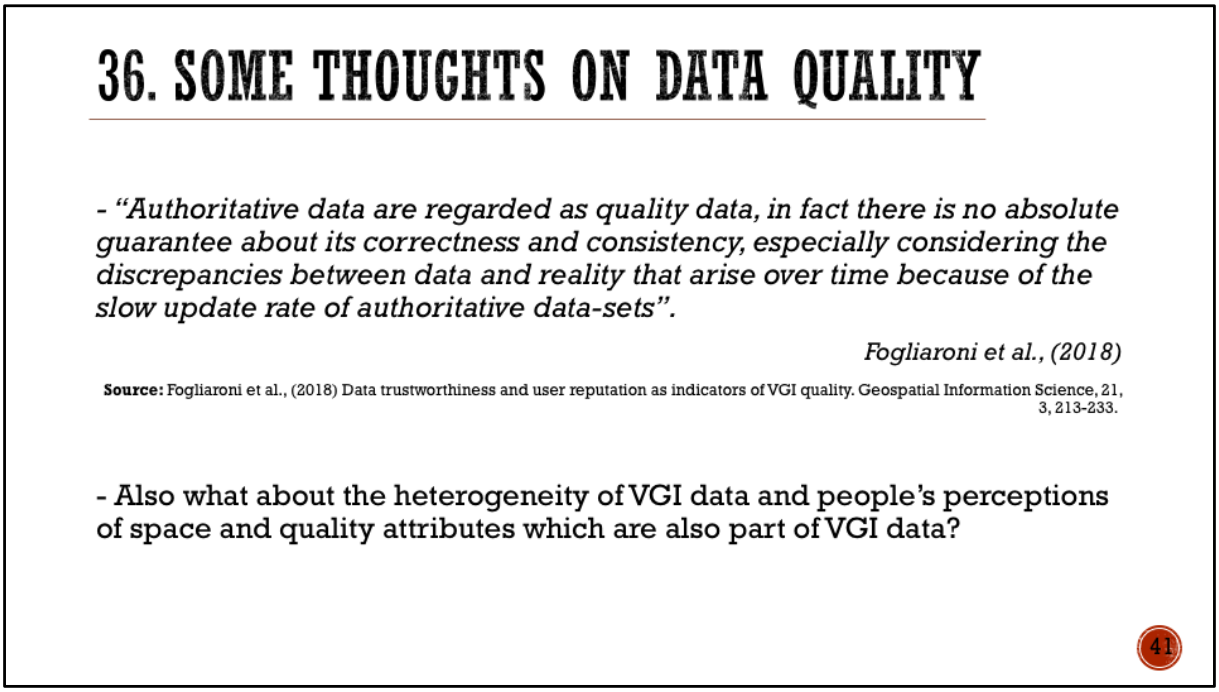
A different way to think about data quality attributes and reflect on the quality of VGI data both from the way that it is being produced but also from the point of how it is being used, includes thinking about data "usefulness", "usability" and

"trustworthiness". This is also very relevant - as you will see in the seminar this week - for designing VGI interfaces so that they directly help people make a conscious decision about these attributes. Before we look into these attributes further let's describe what each one means:

-**Usefulness ('fitness for purpose/use'):** A product which is useful is one that allows a user to accomplish a task. Something that is not useful is unlikely to be accepted/used by any user.

-**Usability:** It is more than "useful" it examines the way that the product will be used and whether it enables the user to do so in a simple (or as simple as possible) and effective manner. If a product is not usable but useful it may still be used (if there are no similar products which are easier to use!

-**Trustworthiness:** Can you rely on it to make any decisions (especially under risk and uncertainty)?

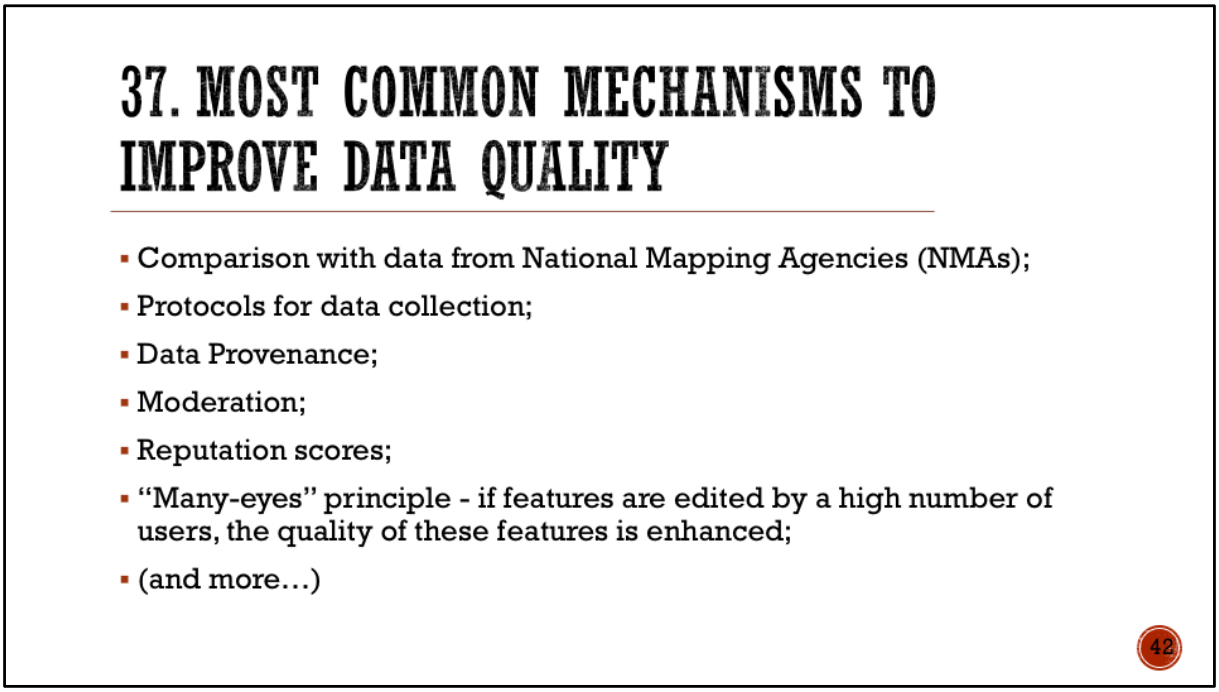


We saw that people are quite skeptical when it comes to the quality of VGI. There are several scholars who investigate VGI data quality mostly by comparing these data sources with authoritative datasets. National Mapping Agencies follow specific protocols and procedures in order to produce high-quality products, which of course were not designed with VGI in mind. Nevertheless, the problem with this approach is that while authoritative data are regarded as quality data, in fact there is no absolute guarantee about its correctness and consistency, especially considering the discrepancies between data and reality that arise over time because of the slow update rate of authoritative data-sets.

Applying the same principles to VGI data cannot always be effective. For example, a significant part of collecting VGI, is capturing people's perceptions about space or quality attributes which are associated with a community's geographic space (for example in VGI applications collecting information about noise or feelings of security and fear of crime, where there is usually some "subjectivity" in how these attributes are perceived with respect to their geographic locatoon). In these cases data quality cannot be really assessed in comparison with authoritative data sources.

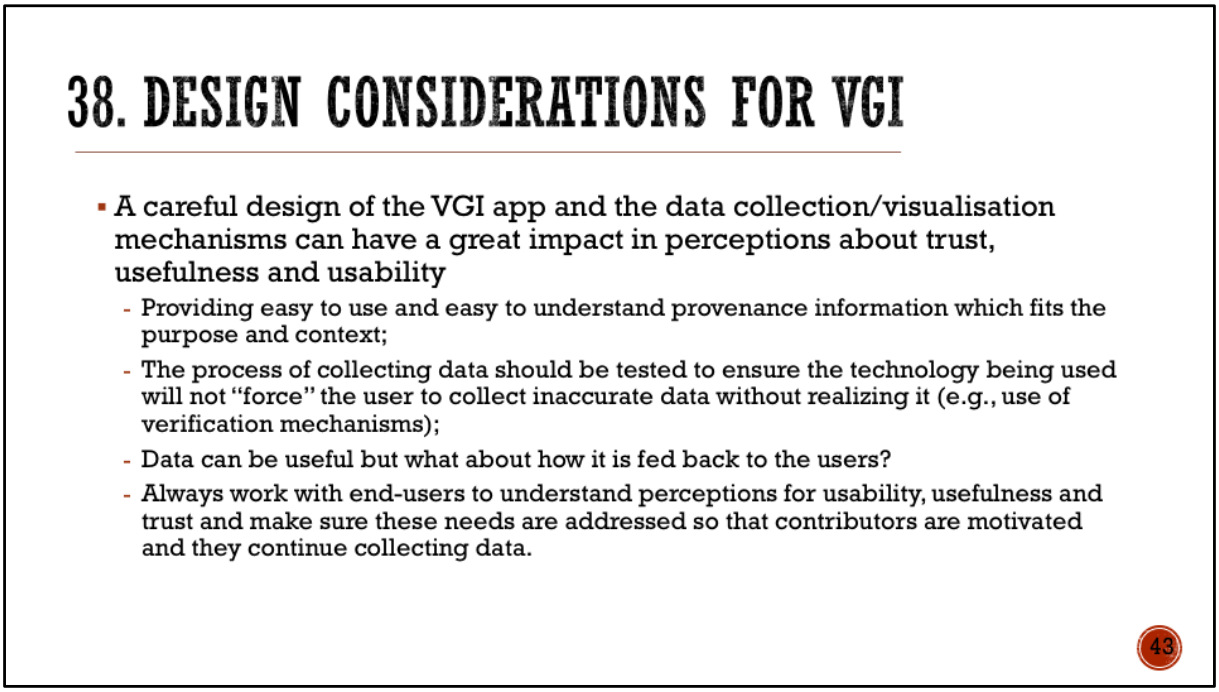
Therefore, another approach of assessing data quality is by analysing the evolution of

the data itself, i.e. its history or provenance and through the provision of metadata. What it is important here is to think of not only what information needs to be provided so that a user can make a conscious decision about whether the data is trustworthy or useful etc, but also the information needs to be in the right format so that can be properly accessed and understood by users with varying levels of geospatial expertise. Nikos Papapesios will talk more about this in the seminar this week.



At the moment some of the most common ways to improve data quality include:

Comparison with data from National Mapping Agencies;
Establishing very rigorous protocols for data collection – although this is not always possible to achieve;
Providing information about data provenance – or metadata. The big issue here is that metadata might not be usable or even useful and do not fit to what users need in a particular context;
Moderation of VGI data to improve quality is also another route to take, but again this is not always possible and really depends on the size of the project;
Reputation scores are quite commonly used in crowdsourcing. Reputation can be seen here as the perception of trustworthiness of a person by a community. A higher reputation score can increase confidence in the quality of the data submitted by that person;
Perhaps the single most important and most powerful mechanism in VGI is the may-eyes principle. – This is that if features are edited by a high number of users, the quality of these features is enhanced and also the higher the number of users the easier it becomes to spot and correct mistakes; Other was exist which are not possible to capture all here.



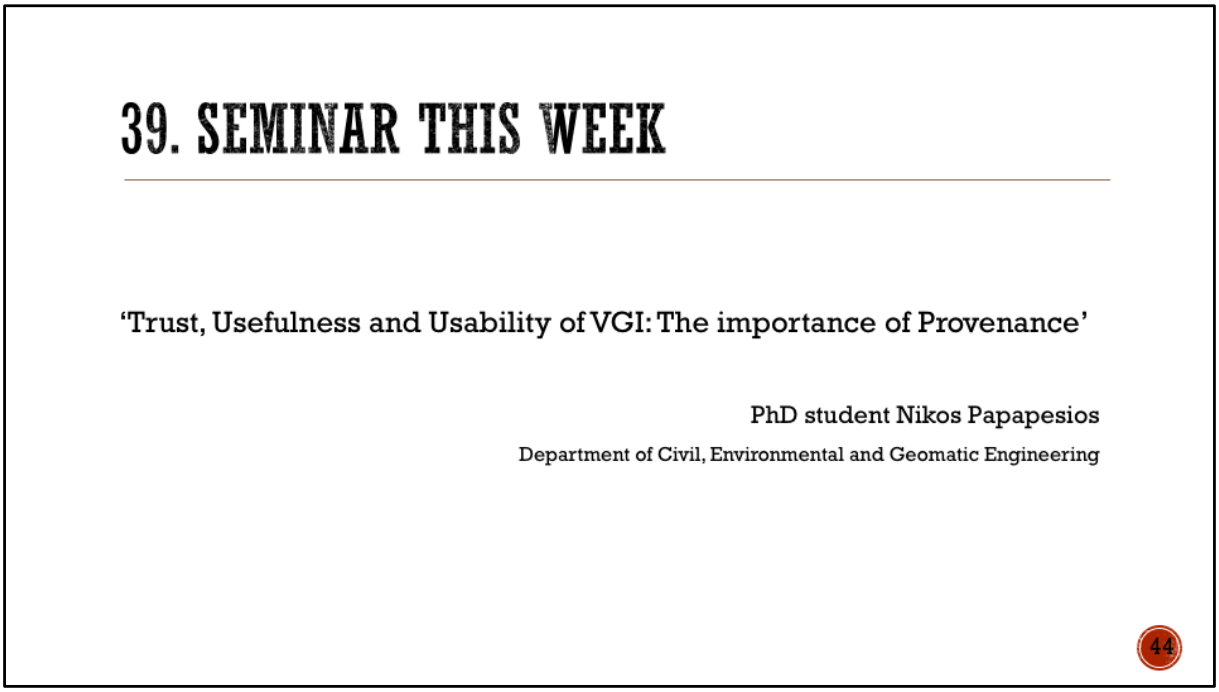
Last but not least, it is not just about what mechanism you use to improve data quality, because it is not always about the actual data quality of your data, but also how the users perceive data quality, especially in cases where appropriate information to make informed judgements does not exist, or they are not well understood.

So ensure that users – including those non-geospatial experts – can make a conscious decision about the trustworthiness of the data; here usability and usefulness are also essential and this can be achieved through a careful design of the VGI app and which may include:

-Providing easy to use and easy to understand provenance information which fits the purpose and context;
-The process of collecting data should be tested to ensure the technology being used will not "force" the user to collect inaccurate data without realising it (e.g., use of verification mechanisms). We have seen countless examples of technological solutions which due to usability barriers they are used incorrectly with some times the users not even being aware of their mistakes;

-Also think about contexts where data could be useful but due to the way it is visualised it is actually not any useful anymore. Or in the context of several VGI examples, where collected data are NEVER shown back to the user who makes the contribution..

-There is no single solution that fits all VGI contexts. It is always essential to work with end-users to understand perceptions for usability, usefulness and trust in specific contexts of use and make sure these needs are addressed so that contributors are motivated and they continue collecting data.



On Friday, the seminar will be delivered by PhD student Nikos Papapesios with a case study from the UK defence context where VGI plays an important role.