International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: https://www.tandfonline.com/loi/tjde20

Data democracy – increased supply of geospatial information and expanded participatory processes in the production of data

Max Craglia & Lea Shanley

To cite this article: Max Craglia & Lea Shanley (2015) Data democracy – increased supply of geospatial information and expanded participatory processes in the production of data, International Journal of Digital Earth, 8:9, 679-693, DOI: 10.1080/17538947.2015.1008214

To link to this article: https://doi.org/10.1080/17538947.2015.1008214

International Journal of Digital Earth, 2015
Vol. 8, No. 9, 679–693, http://dx.doi.org/10.1080/17538947.2015.1008214

Taylor & Francis
Taylor & Francis Group

Data democracy – increased supply of geospatial information and expanded participatory processes in the production of data

Max Craglia[a]* and Lea Shanley[b]

[a]*European Commission, Joint Research Centre, Ispra, Italy;* [b]*Gaylord Nelson Institute for Environmental Studies, University of Wisconsin, Madison, WI, USA*

(*Received 15 December 2014; accepted 13 January 2015*)

The global landscape in the supply, co-creation and use of geospatial data is changing very rapidly with new satellites, sensors and mobile devices reconfiguring the traditional lines of demand and supply and the number of actors involved. In this paper we chart some of these technology-led developments and then focus on the opportunities they have created for the increased participation of the public in generating and contributing information for a wide range of uses, scientific and non. Not all this information is open or geospatial, but sufficiently large portions of it are to make it one of the most significant phenomena of the last decade. In fact, we argue that while satellite and sensors have exponentially increased the volumes of geospatial information available, the participation of the public is transformative because it expands the range of participants and stakeholders in society using and producing geospatial information, with opportunities for more direct participation in science, politics and social action.

Keywords: citizen science; volunteered geographic information; geospatial data

1. Introduction

The global landscape in the supply, co-creation and use of geospatial data is changing very rapidly with new satellites, sensors and mobile devices reconfiguring the traditional lines of demand and supply and the number of actors involved. In this paper we chart some of these technology-led developments and then focus on the opportunities they have created for the increased participation of the public in generating and contributing information for a wide range of uses, scientific and non. Not all this information is open or geospatial, but sufficiently large portions of it are to make it one of the most significant phenomena of the last decade. In fact, we argue that while satellite and sensors have exponentially increased the volumes of geospatial information available, the participation of the public is transformative because it expands the range of participants and stakeholders in society using and producing geospatial information, with opportunities for more direct participation in science, politics and social action.

The paper explores these opportunities but also the many issues that arise in using data generated by the public: can the 'wisdom of the crowd' improve the quantity and quality of data available? Can it be relied upon? What are the risks? Is it sustainable? What are the quality aspects to consider?

*Corresponding author. Email: Massimo.Craglia@jrc.ec.europa.eu

The paper is organized in six sections, reviewing the changing data landscape, the different categories of citizen-generated content, addressing quality issues, outlining key open issues and concluding with some implications for data democracy and some points for discussion, respectively.

2. The wider data landscape

The production and use of geospatial information has changed dramatically over the last 10 years and promises to change even further in the next 10. The first 30 years of development of digital geospatial data were characterized by transition from analogue paper maps to digital products, the launch of an increasing number of civilian-use satellites for telecommunications and Earth monitoring. Most developments were government-led, with relatively few users in government, academia and the private sector. From the 1990s, we saw a more rapid diffusion of Geographic Information System and image processing as software moved from specialized workstations to PCs (Masser, Campbell, and Craglia 1993). We also saw the emergence of data sharing across distributed spatial data infrastructures in the United States, and then in Europe and elsewhere in the world (Masser 1999). While most developments continued, however, to be government-led, the private sector started playing an increasing role as value-added providers of geospatial products. There was substantial use in some areas of the private sector such as oil, minerals and agriculture commodities. However, these were privately held and not generally available for broader use.

The recent growth of technology in sensors, computers and storage has significantly expanded access to information. Belward and Skøien (2014) discuss the evolution of civilian earth observation [EO] satellites and chart the exponential growth in the number of operational missions since 1970, with a steep increase since the year 2000. Moreover, they report that:

> Since the 1970s the number of missions failing within 3 years of launch has dropped from around 60% to less than 20%, the average operational life of a mission has almost tripled, increasing from 3.3 years in the 1970s to 8.6 years (and still lengthening) the average number of satellites launched per-year/per-decade has increased from 2 to 12 and spatial resolution increased from around 80 meters to less than one meter multispectral and less than half a meter for panchromatic.

Therefore, not only more missions are launched but also they last longer as they fail less, and the resolution continues to increase. As an example, Landsat 7 and 8 collect more than 1000 scenes per day, equivalent to some 32 million sq. km each day, while the 41-year Landsat archive includes now some 160 million sq. km (30 times the surface of the Earth) as reported by Covington (2014).

Europe will increase significantly its EO capacity with the launch of its Copernicus[1] and Galileo[2] programs. The Copernicus program envisages six sets of Sentinels' missions with the first launched in April 2014, which over the next few years will deliver up to 8 TB of data per day. A significant change has occurred with the Sentinel adoption of a full and open data policy. Galileo is the European civilian equivalent to the US Global Positioning System (GPS) program for high-precision positioning, navigation and tracking. It will include some 30 satellites in medium Earth orbit, able to provide full global coverage.

In addition to these public sector initiatives, it is important to note the much-increased role taken by the private sector as a provider of data and services. For example, Digital

Globe provides high-resolution, real-time imagery for 45% of the Earth's land surface and processes over 1 billion sq. km of data per year (Marchisio 2014). New companies have also recently emerged like PlanetLabs that have designed and launched a flock of 28 miniature satellites Cubesats called Doves (approx. 6 kg in weight) providing high-resolution imaging (3–5 metres). The company plans to have 131 satellites in orbit in 2015[3].

Another recent development is UrtheCast[4] that has deployed two high-resolution cameras providing high-definition video and imaging at sub-metre resolution on the International Space Station. The cameras cover the earth surface from 51° to –51° latitude with up to 90 passes per day. It is therefore possible to subscribe to a particular location and watch it change in near to real time and high definition (weather permitting).

Aside from the increasing production of data from space, it is interesting to note the merging of sub-metre resolution data from space with millimetre precision data captured from Lidar total stations such as those provided by Trimble or Faro Technologies that can acquire up to 1 million points per second, i.e. can reconstruct a complex building like Tower Bridge in London, in less than 1 hour. The convergence of Building Information Modelling at planning and construction stage, with Lidar reconstructions of three-dimensional (3D) models, makes it now possible to model individual buildings, blocks, neighbourhoods and merged with high-resolution data from drones, aircrafts or space can provide now full 3D models of entire cities down to the most minute detail.

Cities, buildings and objects are also becoming alive through web-enabled sensors. The Internet of Things (IoT) is expected to connect in 2020 forty billion man-made objects in real time.[5] The potential applications of these networks of sensors combined with precision location and imaging are many from real-time mobility information, environmental monitoring, precision farming, urban management and so on. Many new products and services from the private sector are also expected to develop an entirely new industry based on data from space and IoT. Whether these developments will be entirely beneficial or not depends on the points of view,[6] but there is little doubt that change is happening fast with potentially significant social impacts.

A key development has been the rise in availability of mobile phone and smart phones worldwide. Smart phones users represent 30% of the 5.2 billion mobile phone users and 25% of web views globally (Meeker 2014). Mobile phones have changed the lives of millions of people in Africa where they provide support for mobile banking, checking of market prices and other critical applications (Fox 2011). They have contributed to the spreading of Arab Spring uprising,[7] and in general, we often find ourselves wondering 'how did we do that, before mobiles?'

Mobile phones, Internet, mash-ups and sensors have also contributed to a major shift in paradigm in the production and use of geospatial information, with the diffusion of crowd-sourcing and citizen science. The remainder of this paper focuses on these new sources of data and the issues they raise.

3. A typology of citizen-generated content

The massive diffusion of mobile technologies and social media has altered significantly the traditional relationship between producers and users of information in general, and geographic information more specifically. The success of OpenStreetMap (OSM, www.openstreetmap.org) with millions of registered contributors and users is a striking example in the geospatial domain, but one can think also at the multitude of mash-ups on GoogleMaps, and the role of social media (Facebook, Twitter, YouTube, etc.) as main

providers of geographically-tagged content. Coleman, Georgiadou, and Labonte (2009) used the concept of 'prosumers' to summarize this changing landscape in which the public is both producer and consumer of (geographic) information in ways not seen before.

That said, we often see terms like crowd sourcing, citizen science, participatory sensing, volunteered geographic information (VGI) and many others used interchangeably, generating confusion and possible misunderstanding about type of activities, objectives methods and issues. An initial typology is provided below. The categories introduced are not intended to be rigid or mutually exclusive. The purpose is to articulate the similarities and differences among the different concepts used in the literature.

3.1. Citizen science

Citizen Science projects are typically those 'in which members of the public engage in authentic scientific investigations: Asking questions, collecting or processing data, and/or interpreting results' (Bonney 2014). Citizen Science is a form of open collaboration where members of the public participate in the scientific process, including asking questions, collecting and analyzing the data, interpreting the results and problem solving.

We can distinguish among four different 'flavours' of citizen science projects, depending on their primary objective:

- Citizen Science to 'advance scientific discovery and knowledge' (e.g. http://ebird.org, https://www.zooniverse.org/, http://fold.it)
- Citizen Science to 'inform policy' and environmental management (e.g. collect environmental data on air and water quality, noise See, e.g. the five citizen observatories funded by the EU R&D programme: http://www.citizen-obs.eu/)
- Citizen Science for 'education and awareness raising' (e.g. Sensebox for schools: http://www.sensebox.de, and the GLOBE project: http://globe.gov)
- Citizen Science for 'community building' (e.g. http://publiclab.org/ has developed kit, methods and resources to support scientific or local communities in setting up their projects; and the US Environmental Protection Agency's Air Sensor Citizen Science Toolkit available at www.epa.gov/heasd/airsensortoolbox/ connects scientists with local communities and provides guidance on new low-cost compact instruments for measuring local air quality where people 'live, work, and play').

Of course, these are not mutually exclusive categories, and many projects will aim at more than one objective. The distinction is, however, useful for the discussion in data quality in Section 4.

One of the common elements of Citizen Science projects is that they frequently, although not always, involve a strong interaction between the academic community and the public, with the methodology of the projects usually designed by the researchers with a greater or lesser extent of involvement by the participants from the public (Shirk et al. 2012; Haklay 2011; Newman et al. 2012).

3.2. Crowdsourcing

This is a process 'where individuals or organizations solicit contributions from a large group of unknown individuals (the crowd) or, in some cases, a bounded group of trusted individuals or experts' (Bowser and Shanley 2013, 45).

Contributors may be paid or not, and the range of contributions may include specific tasks of smaller or greater complexity, e.g. mapping as in www.openstreetmap.org,

pattern recognition (as in the search for the missing flight MH 370 in which more than 8 million people scrutinized over 1 million sq. km of ocean[8]), or innovative ideas (e.g. https://www.atizo.com/), but also other forms of contributions like money (e.g. https://www.kickstarter.com/), time (e.g. http://www.timebanking.org/) or computing resources (e.g. http://setiathome.ssl.berkeley.edu/, or http://folding.stanford.edu/).

In many of these projects, the methodology for data collection and analysis is also centrally designed by researchers, and quality assurance (QA/QC) methods are often put in place. The boundary between citizen science and crowdsourcing is a fuzzy one. So platforms like Zooniverse include citizen science projects based on crowdsourcing, but not all citizen science projects are using this approach, nor are all crowdsourcing projects about citizen science, as for example the case of projects related to emergencies and natural disasters.

3.3 Data mining of citizen-generated content

This category of projects is different from the others in that they do not, by and large, use data that are specifically volunteered for a project, but reuse data published on the Internet (via social media, mobile phone traces, of photo-sharing sites) for other purposes (communication among friends or 'like'-minded communities, photo sharing, digital activists, etc.). Examples include the detection of forest fires via Twitter (Craglia, Osterman, and Spinsanti 2012), and several instances in relation to crisis mapping (e.g. Shanley et al. 2013).

There are different 'flavours' in this category between the data mining of social media 'that is out there' purely on the base of keywords or location (e.g. Dittrich and Lucas 2014) and the directed data mining in which people are encouraged to use specific tags to signal an event. This latter case is in effect another form of crowdsourcing such as the Tweet Earthquake Dispatch service by the US Geological Survey to monitor and alert about earthquakes (@USGSted and @USGSBigQuakes), or applications such as #snow.

Projects in this category use a range of methods for data extraction, integration and assessment, which often go under the label of 'data analytics'.

3.4. Other dimensions

Orthogonal to the categories outlined above, there are other dimensions that are relevant to this discussion: the extent of engagement and the geographic nature of the data.

Extent of engagement: projects are often classified on a ladder that includes contributory projects (mostly data collection); collaborative projects (data collection and refining project design, analysing data, disseminating results); and co-created projects (designed together by scientists and public where the public shares most or all the steps in a scientific project/process) (Bonney et al. 2009). Shirk et al. (2012) introduce two other classes at either end of this range: contractual projects, where communities ask professional researchers to conduct a specific scientific investigation and report on the results (i.e. no direct community participation beyond commissioning), and collegial projects, where non-credentialed individuals conduct research independently with varying degrees of expected recognition by institutionalized science and/or professionals (i.e. full control, with no direct involvement of scientists). In all these classifications, there is a hidden assumption that scientists are leading the project (even in co-created), but increasingly we should find projects with a reverse relationship where participants lead, and scientist may or may not become involved.

While the collaborative and co-created projects imply an active volunteering of data/information by the participants, the contributory projects may include both active participation (e.g. collecting a measurement or photo and sending via a mobile app.) and a passive form of contribution when data generated by the public (via communication in social media, GPS or mobile network traces) are harvested and used by projects with limited or no knowledge from the original contributor. Other possible terms would be 'participatory' sensing and 'opportunistic' sensing (Jiang and McGill 2010).

With respect to the geographic dimension, one can distinguish between explicitly geographic or implicitly geographic data, with explicit denoting that the geographic dimension is of primary concern to the information provided, while implicit denotes that the dimension was not originally an integral part and is only of secondary concern or derived. So if a piece of information is about the characteristics of a place, it is explicitly geographic (e.g. OSM, geowiki). On the other hand, information that is not specifically about a place (e.g. the picture of a bird, or a measurement of noise levels) but can still be geocoded is implicitly geographic. For a more extensive discussion of the geographic properties of User-Generated Content, see also Antoniou, Morley, and Haklay (2010) and Purves, Edwardes, and Wood (2011). Figure 1 summarizes the different classes introduced in this paper.

As argued earlier, the categories do not have to be seen as mutually exclusive, particularly between citizen science and crowdsourcing, the latter being one of the possible approaches to support citizen science. A clearer distinction in paradigm is suggested between citizen science/crowdsourcing, and data mining/big data. In citizen science/crowdsourcing projects, questions are developed first, then methods and then data are generated to address the question posed. This is the traditional scientific paradigm. On the contrary, in the data mining/big data approach, the data are generated first independently of the questions. These are formulated afterwards.

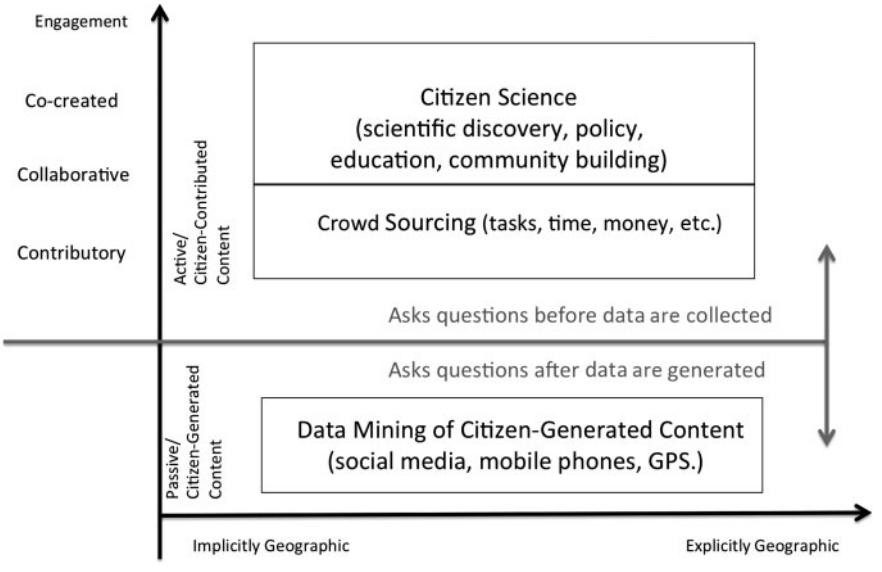

Figure 1. A typology of citizen-generated content.

Croll (2012) made this argument:

> In the old, data-is-scarce model, companies had to decide what to collect first, and then collect it. A traditional enterprise data warehouse might have tracked sales of widgets by color, region, and size. This act of deciding what to store and how to store it is called designing the schema, and in many ways, it's the moment where someone decides what the data is about. It's the instant of context. That needs repeating:
>
> You decide what data is about the moment you define its schema.
>
> With the new, data-is-abundant model, we collect first and ask questions later. The schema comes after the collection. Indeed, big data success stories like Splunk (http://www.splunk.com), Palantir (https://www.palantir.com), and others are prized because of their ability to make sense of content well after it's been collected — sometimes called a schema-less query. This means we collect information long before we decide what it's for. And this is a dangerous thing.

There is of course a mixture of the two approaches. Typically, data may be collected for one purpose, but its impact is expanded through creative applications not envisioned in the original schema. One only has to look at satellite remote sensing, GPS or the applications of software (www.palentir.com) to see creative evolution of applications.

4. The quality argument

There is often a perception that citizen science and poor data quality go hand-in-hand. This is equally often used as an argument to dismiss citizen science in favour of traditional scientific approaches and/or data collection by official agencies. The typology provided in Section 3, is designed to clarify that citizen science is not a single concept, but there is a range of approaches, objectives, and methods available.

Citizen science and crowdsourcing projects do not have greater data quality issues than other research projects. They are based, by and large, on scientifically–designed methodologies, so the question, like in any other project, is whether the methodology is sound and appropriate to respond to the projects' objectives. As an example, Haklay (2010) comparing the quality of OSM to Ordnance Survey official data concludes that: 'you can expect OSM data to be within positional accuracy of over 70%, with the occasional drop down to 20%' (700) and 'This preliminary study has shown that VGI can reach very good spatial data quality' (p. 701). This of course does not mean that there are no quality issues with VGI or citizen science! Only that they are not inherently greater than those of other research projects.

With respect to the objectives, Section 3 also articulated different classes: projects aiming at scientific discovery or policy analysis require greater data quality than projects the aim of which is education or community building. In the latter cases, the outcome does not depend (as much) on the quality of measurements but on the process of designing the project and the active participation of citizens. As an example, the Sensebox project (www.sensebox.de) aims to teach science in general and GIS in particular in schools. Students learn how to build a sensor platform; ask scientific questions about the relationship between height, temperature and pressure; learn how to do some basic programming; and then collect and analyse the data to answer their questions. This project is a great example on how to communicate science to students and how the kit can be successfully used to ask challenging questions to students in order to

stimulate them in the scientific/spatial thinking. In this case the accuracy of the data collected is not that important. The process is the outcome.

As mentioned earlier, citizen science is also not a new phenomenon. It has existed for well over a century as most 'scientific' observations, particularly in the natural sciences, were made in the past by amateurs and volunteers. For Example, the Audubon bird count in the United States has been carried out annually since the year 1900 (http://www.audubon.org) while in the United Kingdom the British Trust for Ornithology has been running counts since the 1930s.

What has changed of course is that the availability of mobile technologies and the Internet has increased manifold the number of projects and participants. Citizens are making very important, but not always fully recognized, contributions to science in several fields. For example, Cooper, Shirk, and Zuckerber (2014) analysed 250 peer-reviewed papers used by Knudsen et al. (2011) to challenge some of the claims in the relationships between climate change and the timing of spring migration among migratory birds. The analysis reveals that almost 50% of these peer-reviewed papers were based on citizen science, i.e. on data collected by the public. This is a clear statement on the undisputable contribution to science by citizen-generated data.

What is equally interesting in the analysis by Cooper is that the term 'citizen science' was never used in any of the 250 papers reviewed! The term 'volunteer' was used in 45% of the papers, and another 45% used a range of terms such as 'birder', 'people', 'ringer' or 'public' (the remaining 10% required contacting the authors for details of the methodology). This means that if somebody had tried to substantiate the contribution of citizen science to this area of research using the term 'citizen science', he/she would have drawn the false conclusion that there was no contribution!

Crowdsourcing methods can also provide excellent results beyond contributions to science such as OSM, Wikimapia, Wikipedia, etc., demonstrate. As indicated earlier, an early comparison of the quality of OSM versus the official cartography by Ordnance Survey (Hacklay, 210) showed that OSM was comparable in quality to commercially available products, with the main difference that errors were not randomly distributed but where dependent on the abilities of the contributor. That study, however, was made in 2008 when OSM was still relatively young and covered only 29% of England. Since then, more stringent QA/QC methods, training of volunteers, wisdom of the crowd peer review, triangulation with different sources and so on are all methods that have been deployed to ensure quality or at least fitness-for-purpose of projects like OSM (see, e.g. Haklay [2010] and Goodchild and Li [2012] for a review of methods).

The area where issues of data quality become particularly challenging is in the data mining/analysis of data generated by the public prior to the analysis. Examples are the many projects analysing social media (Twitter, Facebook, FourSquare, Flickr, etc.) to address a wide range of issues from natural disaster and crisis mapping (Shanley et al. 2013) to what Oboler, Welsh, and Cruz (2012) label as computational social science, i.e. the study of social dynamics through advanced computation.

For most researchers, the issues of quality using social media arise from the use of a small percentage of the data universe that is available free of charge from public APIs. As an example, Twitter APIs provide access to a small percentage of the 500 million tweets generated each day. Even a small percentage makes a lot of tweets, but the researcher never knows how representative this percentage is of the total number of tweets, how representative of the underlying population, what other biases have been introduced in reply to the search or in the stream, etc.

Equally, if one was concerned with the geographical dimension of the data, only about 5% of Tweets have latitude–longitude coordinates, and even then it is not always clear if these are the coordinates of the user's home, as provided in the profile or the place from which the tweet is sent, and/or if this is the place to which the tweet refers. As a result, researchers may need to try and parse the text of the tweets to find name places to geocode, and then triangulate with other datasets from a range of sources, including other social media and official data sources to try and assess the fitness-for-purpose of the data harvested. In Craglia, Osterman, and Spinsanti 2012, we show that the automatic workflow we developed to find forest fires in the South of France delivered some 70–80% of correct results compared to official data sources based on MODIS satellite data. Therefore, even relatively DIY methods, on an unknown sample of the data, can deliver some useful results, but there is nevertheless significant variability in the results of different projects (see Morstatter et al. [2013] for a useful comparison of sampled versus full Twitter data).

Could one build mission-critical systems around such outcome? The answer to this question is: it depends. Largely, it depends on resources available, alternatives, and context. In case of emergency, and where time is of the essence, it may often be the case that data from social media is more timely and useful than alternatives, including data from official sources, which maybe late in coming or out of date. In case of an emergency, if you have to decide whether to evacuate your house or not, false positives (Type I errors) may be better than false negatives (Type II)! In many countries where official data, such as maps, are severely out of date, crowd-sourced initiatives like OSM are valid alternatives as a basis for decisions to deliver aid or services. On the other hand, where good official data exists and is accessible, crowd-generated data could complement it and provide additional perspectives, without being needed as replacement.

Of course, if resources are not an issue and you own or can acquire very large pools of social media data, and other data sources with which to triangulate, the results are more likely to be better, at least potentially. This is why the social media giants like Google, Microsoft, and Facebook increasingly force users to sign on with a single ID to be able to triangulate better across multiple platforms and improve user profiling. Large corporations have invested heavily in Big Data and data analytics methods to improve their business intelligence, and governments are also increasingly doing so for a range of reasons including security, policy assessment, or to influence electoral results.

These are clearly mission critical operational systems, although they are not exempt from quality challenges as the increase in data volumes and computer power does not necessarily result in improved methodologies or guarantee results free from spurious correlations (see for example Marcus and Davis 2014).

Summarizing the discussion above, data quality, or more appropriately fitness-for-purpose, is an important issue, but, by and large, not much greater for projects based on citizen-generated content than for any other research project. Either there is a good fit between research questions, methods, data, and conclusions, or there isn't, regardless of methods or data source.

On the other hand, there are major benefits from these new data sources: involving the public in science and policy analysis can raise awareness about key issues being studied, help change behaviours (for example in support of more sustainable consumption), narrow the gaps between science, policy and society, support communities, and support good governance. The use of social media also may provide valuable sources of real-time data, including qualitative data, about policy-relevant issues and help address

them at an early stage before social and political positions become too entrenched. Last but not least, the use of crowd-generated content can be cost-effective. For example, Theobald et al. (2014) analysed 388 biodiversity citizen science projects (in English) from around the world and found that 1.3 to 2.3 million people volunteered each year in these projects with an economic in-kind contribution worth up to $2.5 billion per annum, equivalent to 40% of the annual budget of the National Science Foundation (in 2013). On a national scale, Mackechnie et al. (2011) reported that terrestrial biodiversity in the UK involved some 30 different organizations, often relying on volunteers. Their contributions had an estimated value, in 2007–2008, of £20million against government funding of £7 million. Moreover, POST (2014) indicates that:

> Seven out of the 26 biodiversity indicators rely on volunteer-collected data and it has been estimated that volunteers are capable of monitoring 63% of the 186 indicators that the UK is obliged to monitor through twelve international biodiversity agreements. (Pg. 3)

Beyond biodiversity, not much research has been carried out to date quantifying the value (both economic and social) of crowd-generated data. This is therefore an important gap needing further research.

5. Open issues

5.1. Sustainability

A major challenge of citizen science and crowdsourcing projects is to sustain the commitment of the participants through the project. Understanding the motivation of the participants is important, and equally important is to manage expectations, and try and align the objectives of the project with the expectations and motivations of the participants. Volunteers may be motivated by the need for social interaction, a deep interest in the topic, gaining new employment opportunities and skills, contributing to environmental or community-focused policy, and so on.

Some topics and communities are easier to engage with than others. For example, there is a very wide and well-established community of bird watchers. This community is not only motivated but also supported by multiple infrastructures to share their findings and get visibility, and is also highly competitive, so approaches based on gaming may also be successful in gaining and retaining the participation and commitments. Other communities are not as well established, or the topic is not as popular. For example, monitoring eels may not be as attractive and projects working in this area, such as the Hudson River Eel Project,[9] may face greater difficulties in recruiting and retaining volunteers (Bowser and Shanley 2013).

Gaming is a useful way to engage individuals and groups in social activities, including those with serious aims (e.g. http://www.seriousgamesinstitute.co.uk/), although as a strategy it may not always be appropriate. In some instances, forms of community acknowledgment, reward or open awards may be more useful, and it is important to tailor the method of engagement to the specificity of the project and the participants (Newman et al. 2012).

5.2. Reproducibility

This is a key issue of many citizen-based projects. To be able to reproduce the results of a project, it must be possible to have access to its documentation, the data collected, and the methods used. Many projects strive to do this, but a very large number do not, faced with

several challenges, legal, organizational and policy related. In Craglia and Granell (2014), we identified some of these challenges:

5.2.1. To publish or not to publish?

Although there are many policy drivers pushing for the open publication of data, it is clear that not all data collected by individuals can be published as is. In some cases, it is confidential from a commercial or environmental point of view (e.g. rare species location), in others there may be identifiers of the individual collecting or generating the data that need to be suppressed (Bowser et al. 2014). One possibility is to aggregate the data so that it can be reused. This may help overcome privacy concerns but may not always give the reusability and transparency needed. In all cases, the rights of the contributors to the project need to be respected with clear rules of engagement and informed consent about subsequent re-use.

There are many cases of good practice in information sharing and growth in open datasets, for example in the biodiversity domain (see the Global Biodiversity Information Facility[10]). In other areas there are still challenges: For example, many projects take place in collaboration with local authorities that recognize the potential of citizen participation and data collecting contributions but are not clear yet how to respond to the inputs provided by the public and how to integrate them into the well-established information flows, which are often regulated by legal requirements. How can data collected by the public on air quality, water quality or noise, often with equipment of low quality, be reconciled with better quality but more sparse observations from official sources? How to manage the debates between the measured magnitude of a phenomenon and the public perception of the same phenomenon informed by observations maybe of lower quality but amplified by the very large numbers of observers? These are not easy questions to address in this early stage of the Citizen Science phenomena, and it is not surprising that many public authorities have difficulties in finding consistent answers.

5.2.2. Where and how to publish?

The exponential rise in number of citizen science projects (the Scistarter[11] site lists, for example, over 600 ongoing projects) raises the question of if and how the data that are collected can be accessed and re-used. In some projects, there may be tensions between researchers wanting to hold on to the data until they have published their academic papers and volunteers wanting to have the data published quickly. In many other instances, the data are published on the project website which then disappears shortly after the end of the project. This is a common issue in many if not most research projects in general, not just citizen science projects. As an example, Pepe et al. (2012) analysed the URL links embedded in Astronomy publications over 15 years and found that 44% of links were broken 10 years after publications. Only 15–20% of links pointing to curated data archives were broken, while links to project or personal websites decayed at a much faster rate. Considering that astronomy is a very well organized community with a significant number of institutional data archives, the situation is clearly much worse for research projects in other disciplines not so well equipped with underlying infrastructures for data repository, curation and long-term access. This is an issue that needs to be addressed in general terms also in the new research programmes (like the Horizon 2020 in Europe), which support the policy of opening as much research data, models and scientific output

as possible. This policy must be matched by a strengthened network of open data archives worldwide.

Reproducibility means not only being able to access (and understand) the data but also the models and algorithm used for the analysis. Hence, it is equally important to publish these processes or workflows, with appropriate documentation written in ways that can be understood from different disciplines and backgrounds. Platforms and software developed by different projects should also strive to be interoperable so that they can be easily reused and combined with others. This may take quite a while to achieve, given the difficulty we are witnessing even in developing basic metadata for datasets, let alone models and processes.

6. Data democracy and way forwards

The issues of reproducibility of scientific evidence, and policy advice discussed in the previous section in the context of the many, but relatively small, citizen science and crowdsourcing projects, take a different life when discussing projects based on social media and citizen-generated content. Here we enter the realm of Big Data that is being pushed heavily by major corporations and governments. Without entering into issues of privacy, competition and data protection, which are beyond the scope of this paper, it is worth noting that the increasing opportunities of Big Data analytics risk taking us towards less transparency and accountability rather than more, i.e. less data democracy because Big Data analytics requires access to very large datasets, often commercially protected, specialized skills, software and high-intensity computing available to relatively few. Opening up the bases of decisions by government agencies and private corporations is a matter of democratic right that should not be underestimated (Croll 2012). It would be perverse if all the movement towards Open Data ended up as providing greater wealth of information for the few to the detriment of the many. From this perspective, it may well be that citizen science is in fact not a matter for concern but a necessary development to foster a vibrant, open and informed society. To help do so, there are a number of issues that need to be addressed:

(1) How can we integrate citizen science and ‘data education’ in the school curricula and help educate children from an early age about the conduct of science, the handling of uncertainty, as well as fostering multi-disciplinary system thinking?
(2) How can we engage communities (scientific, issues-based and area-based) to take ownership of the data and analytical tools at their disposal so that projects are sustained over time and contribute to their ability to mobilize and act?
(3) How can we develop frameworks for sharing citizen-generated content across projects and develop the culture, and supporting infrastructure, for data management (publishing, archiving and curating) so that evidence based on citizen-generated content can be reproduced and analyzed over time?
(4) Are the existing regulatory and practice frameworks adequate to minimize risks with respect of privacy, security, Intellectual Property and liability when handling citizen-generated content? Is harmonization of such frameworks necessary?
(5) How can we design an international research programme on the value of citizen-generated content for policy, science and society? What controlled experiments and real-life portfolio of case studies need to be devised to have comparable data across contexts and countries?

We hope that this paper contributes to a collective reflection on these issues and promotes citizen science practice across scientific disciplines.

Acknowledgements

A previous version of this paper was presented and discussed at the workshop 'Assessing the Socio-economic Impacts and Value of "Open" Geospatial Information', 28 and 29 October 2014, George Washington University, Washington D.C. The authors are grateful to the discussants and reviewers for their useful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. http://ec.europa.eu/enterprise/policies/space/copernicus/index_en.htm.
2. http://ec.europa.eu/enterprise/policies/satnav/galileo/why/index_en.htm.
3. https://www.planet.com/flock1/.
4. https://www.urthecast.com/features.
5. https://www.abiresearch.com/press/the-internet-of-things-will-drive-wireless-connect.
6. See, for example, http://www.pewinternet.org/2014/05/14/internet-of-things/.
7. http://www.washington.edu/news/2011/09/12/new-study-quantifies-use-of-social-media-in-arab-spring/
8. http://www.zerosixright.com/how-to-use-tomnod-to-find-mh370/.
9. http://www.dec.ny.gov/lands/49580.html.
10. http://www.gbif.org/.
11. http://scistarter.com/about.html.

ORCID

Max Craglia http://orcid.org/0000-0001-6244-6428

References

Antoniou, V., J. Morley, and M. Haklay. 2010. "Web 2.0 Geotagged Photos: Assessing the Spatial Dimension of the Phenomenon." *Geomatica* 64 (1): 99–110.

Belward, A., and J. Skøien. 2014. "Who Launched What, When, and Why: Trends in Global Land-Cover Observation Capacity from Civilian Earth Observation Satellites." *ISPRS Journal of Photogrammetry and Remote Sensing*. doi:10.1016/j.isprsjprs.2014.03.009.

Bonney, R. 2014. "Citizen Science: A Vision for the Future." Keynote at the Citizen Cyberscience Summit 2014, London, February 21.

Bonney, R., H. Ballard, R. Jordan, E. McCallie, T. Phillips, J. Shirk, and C. C. Wilderman. 2009. "Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education, CAISE." http://informalscience.org/images/research/PublicParticipationinScientificResearch.pdf.

Bowser, A., and L. Shanley. 2013. "New Visions in Citizen Science." Woodrow Wilson Centre. http://www.wilsoncenter.org/publication/new-visions-citizen-science.

Bowser, A., A. Wiggins, L. Shanley, J. Preece, and S. Henderson. 2014. Sharing Data While Protecting Privacy in Citizens Science. http://www.wilsoncenter.org/publication/sharing-data-while-protecting-privacy-citizen-science.

Coleman, D., Y. Georgiadou, and J. Labonte. 2009. "Volunteered Geographic Information: The Nature and Motivation of Producers." *International Journal of Spatial Data Infrastructures Research* 4: 332–358. http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/140/223.

Cooper, C. B., J. Shirk, and B. Zuckerber. 2014. "The Invisible Prevalence of Citizen Science in Global Climate Change Research." *PLOS ONE* 9 (9): e106508. doi:10.1371/journal.pone.0106508.

Covington, S. 2014. "The USGS Landsat Big data Experience." Paper presented at the Copernicus and Big Data Workshop, Brussels, March 13. http://www.copernicus.eu/pages-principales/library/presentations/copernicus-big-data-workshop/international-experiences/?no_cache=1&cHash=4d31533f1115000bb8dffce2f6131318.

Craglia, M., and C. Granell. 2014. "Citizen Science and Smart Cities." Luxembourg: Publications Office of the European Union. http://digitalearthlab.eu/Citizen_Science_and_Smart_Cities_Full_Report.pdf.

Craglia, M., F. Osterman, and L. Spinsanti. 2012. "Digital Earth from Vision to Practice: Making Sense of Citizen-Generated Content." *International Journal of Digital Earth* 5: 398–416. doi:10.1080/17538947.2012.712273.

Croll, A. 2012. "Big Data is Our Generation's Civil Rights Issue, and We Don't Know It." http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html.

Dittrich, A., and C. Lucas. 2014. "Is This Twitter Event a Disaster?" In *Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science*, edited by J. Huerta, S. Schade, and C. Granell, Castellón, June 3–6. http://www.agile-online.org/Conference_Paper/cds/agile_2014/agile2014_97.pdf.

Fox, G. 2011. Africa's Mobile Economic Revolution. http://www.theguardian.com/technology/2011/jul/24/mobile-phones-africa-microfinance-farming.

Goodchild, M. F., and L. Li. 2012. "Assuring the Quality of Volunteered Geographic Information." *Spatial Statistics* 1: 110–120. doi:10.1016/j.spasta.2012.03.002.

Haklay, M. 2010. "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning B: Planning and Design* 37: 682–703. doi:10.1068/b35097

Haklay, M. 2011. "Citizen Science as Participatory Science." http://povesham.wordpress.com/2011/11/27/citizen-science-as-participatory-science/.

Knudsen, E., A. Lindén, C. Both, N. Jonzén, F. Pulido, N. Saino, W. J. Sutherland, et al. 2011. "Challenging Claims in the Study of Migratory Birds and Climate Change." *Biological Reviews* 86: 928–946. doi:10.1111/j.1469-185X.2011.00179.x.

Jiang, M., and W. L. McGill. 2010. "Human-Centered Sensing for Crisis Response and Management Analysis Campaigns." In *Proceedings of the 7th International ISCRAM Conference*, edited by S. French. http://www.iscram.org/ISCRAM2010/Papers/171-Jiang_etal.pdf.

Mackechnie, C., L. Maskell, L. Norton, and D. Roy. 2011. "The Role of 'Big Society' in Monitoring the State of the Natural Environment." *Journal of Environmental Monitoring* 13: 2687–2691. doi:10.1039/c1em10615e.

Marchisio, G. 2014. "The Future of Geospatial Big Data." Presentation at the Geospatial World Forum, Geneva, May 5–9. http://www.geospatialworldforum.org/2014/presentation/Big%20Data/WGF%202014%20-%20Giovanni%20M%20-%20DigitalGlobe.pdf.

Marcus, G., and E. Davis. 2014. Nine Large Problems with Big Data. http://www.brw.com.au/p/tech-gadgets/nine_large_problems_with_big_data_BOkbvT5G7f6Y2Jc2qiMgGM.

Masser, I. 1999. "All Shape and Sizes: The First Generation of National Spatial Data Infrastructures." *International Journal of Geographic Information Science* 13 (1): 67–84. doi:10.1080/136588199241463.

Masser, I., H. Campbell, and M. Craglia. 1993. *GIS Diffusion*. London: Taylor and Francis.

Meeker M. 2014. Internet Trends 2014 - Code Conference. http://www.kpcb.com/internet-trends.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter Firehose." Proceedings of the Seventh International Conference on Weblogs and Social Media of the Association or the Advancement of Artificial Intelligence, Cambridge MA, July 8–11. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6071/6379.

Newman, G., A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston. 2012. "The Future of Citizen Science: Emerging Technologies and Shifting Paradigms." *Frontiers in Ecology and Environment* 10: 298–304. http://www.esajournals.org/doi/abs/10.1890/110294.

Oboler, A., K. Welsh, and L. Cruz. 2012. "The Danger of Big Data: Social Media as Computational Social Science." *First Monday* 17, July 2. http://journals.uic.edu/ojs/index.php/fm/article/view/3993/3269.

Parliamentary Office for Science and Technology (POST). 2014. "Environmental Citizen Science". POST 476. http://www.parliament.uk/briefing-papers/POST-PN-476/environmental-citizen-science.

Pepe, A., M. Crosas, A. Muench, A. Goodman, and C. Erdmann. 2012. How do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. https://authorea.com/users/3/articles/288/_show_article.

Purves, R. S., A. J. Edwardes, and J. Wood. 2011. "Describing Place through User Generated Content." *First Monday* 16 (9), September 5. http://firstmonday.org/ojs/index.php/fm/article/view/3710/3035.

Shanley, L., R. Burns, Z. Bastian, and E. Robson. 2013. "Tweeting up a Storm." Photogrammetric Engineering & Remote Sensing, October. http://www.scribd.com/doc/174769846/Tweeting-Up-a-Storm-The-Promise-and-Perils-of-Crisis-Mapping.

Shirk, J. L., H. L. Ballard, C. C. Wilderman, T. Phillips, A. Wiggins, R. Jordan, E. McCallie, et al. 2012. "Public Participation in Scientific Research: A Framework for Deliberate Design." *Ecology and Society* 17 (2): 29. http://www.ecologyandsociety.org/vol17/iss2/art29/.

Theobald, E. J., A. K. Ettinger, H. K. Burgess, L. B. DeBey, N. R. Schmidt, H. E. Froehlich, C. Wagner, et al. 2014. "Global Change and Local Solutions: Tapping the Unrealized Potential of Citizen Science for Biodiversity Research." *Biological Conservation* 181: 236–244. doi:10.1016/j.biocon.2014.10.021.