CHAPTER 11

Data and Metadata Management for Better VGI Reusability

Lucy Bastin[*,†], Sven Schade[*] and Christian Schill[‡]

[*]European Commission, Joint Research Centre, Ispra, Italy,
[†]Aston University, Birmingham UK, lucy.bastin@jrc.ec.europa.eu
[‡]Albert-Ludwig University, Freiburg, Germany

Abstract

The rapid expansion of citizen science projects and crowdsourcing applications is yielding a huge and varied pool of Volunteered Geographic Information (VGI) on a wide variety of themes. This VGI may be of huge value for institutions, individuals and decision-makers, but only if it can be discovered, evaluated for quality and fitness-for-purpose and combined with data from other sources. If VGI data are to be discovered, used and reused to their full potential, they must be actively managed. In this chapter we assess the current state of the art regarding data management practices in VGI, identify some challenges, obstacles and best-practice examples, and review a range of developing and established open source technologies which can underpin robust and sustainable data management for VGI. We conclude that VGI is likely to remain patchy and heterogeneous and that existing standards may not be exploited to their full potential. Nevertheless, automated support for documenting the generation and use of VGI, as well as annotations following the Linked Data paradigm, can help to improve interoperability and reuse. We were able to identify good practices within different existing systems, but more research and development work is needed in order to support their joint application for the

How to cite this book chapter:
Bastin, L, Schade, S and Schill, C. 2017. Data and Metadata Management for Better VGI Reusability. 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. 249–272. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbf.k.

benefit of VGI. New data management methodologies can only succeed if their benefits (for example, simplifying administration or lowering the entry barrier to data publication) exceed the implementation costs.

Keywords

Data Management; Quality Assurance; Quality Control; Interoperability; Open Standards

1 Introduction

The visibility and perceived importance of VGI projects and citizen science is continuously increasing, and this book offers insight into many aspects of user-generated content and VGI collections. In this chapter, we summarise some insights on good practice for the storage and dissemination of this type of data.

Data collection and information retrieval in crowdsourcing or VGI projects may happen on very different spatial and temporal scales and diverse thematic areas, and may involve very varied groups of contributors in terms of expertise and interests. VGI campaigns can include, for example, short-term emergency response projects (e.g. after earthquakes and other natural disasters) that exploit volunteered observations along with repurposed information harvested from social media; Citizens' Observatories such as those funded by the European Commission[1], which have structured and strategic goals to foster '… general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources…'(Socientize, 2013); or well established infrastructures and frameworks such as the Global Biodiversity Information Facility (GBIF), which has collated and registered decades-worth of global species data.

Inherently, such initiatives have quite heterogeneous requirements for data cataloguing, access to data, licensing and long-term availability of data, but they do (or at least they should) share some general 'good practice principles' of data management. These principles include aspects such as how to securely store data; how to grant access and to whom; how to document data so they can be found by humans or machines for specific purposes; and how to develop a common understanding of the meaning of collected information so that data can be understood and used, at the very least within the context of the original project, but potentially also outside that domain.

In 2014, the Joint Research Centre (JRC; the EC's science service) in Ispra, Italy, conducted a 'Citizen Science and Smart Cities Summit' and summarised in a technical report (Craglia and Granell, 2014) that at the time when they wrote '… there [was] little interoperability and reusability of [user-generated]

data, apps, and services developed in each project.' A follow-up survey reinforced these conclusions, especially in relation to data management practices in citizen science projects (Schade and Tsinaraki, 2016). Acknowledging these observations, this chapter summarises good practice recommendations in data/metadata management and curation, as well as details on international standards and cross-community interoperability that can potentially overcome the identified shortcomings. Proper application of these principles could permit seamless integration of data sources from different domains into coherent information that can be reused beyond the scope of the original problem – thus leveraging user-contributed content 'to the next level', i.e. making the data discoverable, easier to reuse and thus even more valuable.

2 Data Management Overview

This section first introduces the required background about the topic. It is then devoted to some of the most central aspects of data management. We focus on those items that cut across all types of data and data sources, and highlight the foundational issues that should be addressed in data management and the related planning processes.

2.1 Background

Data appear in many different forms and originate from an ever-increasing number of sources – and VGI is no exception. VGI has huge potential to enrich the data portfolios of the public sector (e.g. environmental measurement stations, earth observing satellites, land surveys and consultations) and of the private/corporate sectors (e.g. mobile phone data, sensor measurements inside vehicles, market studies, etc.). However, the heterogeneous nature of VGI presents challenges for integrating with these 'traditional' data assets, which are generally structured according to the application domains from which they arise, and formatted according to industry standards, which may or may not be open-source. As seen from the concrete examples in this book, VGI can encompass a wide range of measurement and observation types, including GPS tracks, digitised vector graphics, occurrence information, tagged photographs and sound recordings, and observations of individual species over time.

Each of these datasets is generated/collected for an intended purpose (i.e., to deliver some value for a beneficiary), and is dealt with in a particular way. In other words, it is 'managed' in one way or another – independently of the availability of any form of data management plan. The approaches by which data in general, and VGI in particular, are managed diverge greatly, and are highly dependent on the context of generation and use. For example, data collected locally in a field trip to teach a small group of students about digital cartography

might be kept on an SD card, be copied to several desktop computers at the university and be deleted as soon as the course ends. By contrast, worldwide observations about species occurrences might be fed into a well networked structure in order to contribute to a global collection effort which will curate those data for generations of scientists and environmental organisations.

Although it might be debatable whether every single collected dataset should be preserved for potential future use, sharing of volunteer-generated data is a part of the unspoken contract with the original contributors that underlies citizen science, and can be crucial in maintaining the commitment of volunteers. Bearden (2007) records how, in the absence of feedback on their mapping efforts, volunteer USGS contributors '… would become alienated when they realized that their meticulous work would not be used in the foreseeable future …'. In a broader context, if data are likely to be usable for science, then, following recent moves towards reproducibility, they must be made reusable. These requirements for repeatability, transparency and independent evaluation inevitably suggest a need to curate and preserve data collections. With the growing availability of data storage and data sharing capacities, many of the technical needs are well addressed. However, organisational peculiarities and the differences between communities of practice mean that, in reality, multiple different approaches can be applied. While some thematic areas and communities have well established and internally consistent approaches to data handling and sharing, those experiences and practices are rarely exchanged widely across parties with different interests. To give an example: the geospatial community (or, more strictly speaking, the spatial data infrastructure (SDI) community), has developed in-depth knowledge and best-practice recommendations on managing geographic and other spatial information using web services – especially under the ISO Technical Committee on Geographic Information/Geomatics (ISO/TC211) and the Open Geospatial Consortium (OGC). However, interconnections with the biodiversity and nature conservation community have until recently been limited to a few dedicated projects, including, for example, EU BON[2] and COBWEB[3]. However, as citizen science moves into a new era of data aggregation and harmonisation, this situation is changing fast, making a discussion of data management practices especially topical in the domain of VGI. We will re-visit some of the SDI community standards below, in order to indicate reuse potentials.

While each individual collection of VGI is valuable to preserve *per se*, VGI also has reuse potential for purposes that might not have been initially foreseen. These purposes might include longitudinal studies on the use and evolving concept of VGI itself, but could also involve integration with other data sources and interconnection with previously unknown data flows and systems. It is therefore an emerging practice to follow common standards and support interoperability, in order to avoid introducing artificial barriers to such novel and unforeseen usages of VGI. The Group on Earth Observation (GEO) recently published just such a set of data management principles for the Global

Earth Observation System of Systems (GEOSS)[4]. Simultaneously, and along the same lines, the Belmont Forum – a group of the world's major and emerging funders of global environmental change research – released their data principles[5]. The latter principles focus on Findability, Accessibility, Interoperability and Reuse (FAIR) and will be used as a lens through which to assess the state of the art in Section 2.

2.2 Organising Data

One of the very first challenges is the organisation of the data themselves. Before even considering the concrete storage format and structure used, it has to be decided at some point which items are considered data in an 'atomic' form, and how these items might be packaged. As we will see later in the chapter, these early decisions will impact other areas, such as the provision of (persistent) identifiers or the granularity of metadata (data about data). In the context of airborne imagery, the decision could be whether to make accessible as one unit a whole series of images from airborne imagery gathered in a single flight or whether to treat each single scene (image) as a single dataset. Analogously, a species observation could be put into a collection that unites all data relating to a particular day, person, sensor type (e.g. smartphone), administrative region, area of interest (e.g. a natural park), field campaign, etc. The particular choice of grouping will depend on the intended use, which in turn will define the discovery and access needs.

2.3 Persistent Identifiers

Data can only be unambiguously recognised – especially when they are shared with other people – if they can be uniquely and persistently identified. In other words, the data need to be branded in some way that does not change over time. If the data are to be accessible, it must also be possible to resolve that persistent and unique identifier into an appropriate data request.

Without going into too much detail about the meaning of uniqueness and identity, it obviously makes a difference whether a persistent and unique identifier is assigned to every 'atomic' data item or to collections that apply any of the criteria listed above.

The meaning of persistency also has to be challenged: which authorities can guarantee the persistency and uniqueness of identifiers? What if identifiers contain the names of institutions or groups that disappear in real life? Who can guarantee a service that resolves certain identifiers in order to retrieve the actual dataset? Furthermore, it has to be noted that in cases where unique and persistent identifiers are allocated to a data stream, for example one generated by a person or a sensor, the retrieved data will change over time. In practice, the

identifier could resolve to the latest data item that has been collected, or to an accumulated collection. Some specific mechanisms for minting and managing persistent identifiers are detailed and described in Section 3.

2.4 Data Documentation

Are we able to use a dataset that we created ourselves? Can we use it again a few years after we collected it? How are others supposed to find that dataset, understand what it really encapsulates (and assess if it might be valuable for their work), access it and provide their experiences and impressions about it? The answer to all of these questions lies in metadata, or, in other words, the appropriate documentation of data – an answer which is more easily given than implemented.

Documentation is required for a wide range of purposes (e.g. discovery, evaluation and use), and therefore possible forms of documentation vary greatly. Here, again, the packaging of VGI is one determining factor, since one might document a range of possible 'entities', for example: a single observation; observations from one person (including also a description of that person); and VGI collected for a particular area (including also documentation about the area). A dataset stored as a collection of individual observations or measurements might include information about the accuracy of each single value; it has to be determined how this accuracy information is then propagated to a collection of measurements in order to achieve an overall quality measure for the dataset. If a user is filtering this dataset for potential use in an analysis and their fitness-for-purpose criteria include accuracy, then, in theory, this aggregate measure of quality should be recalculated for each candidate set of observations – a considerable challenge for the architecture within which the data are being curated and made accessible for discovery. To give another example, in a VGI dataset where observations can be attributed to an individual, the documentation might include the reputation of this individual in the context of a particular activity or community; but how should such values be propagated when talking about a group of people? At the time of writing, accessible and robust tools for this type of aggregation are lacking.

Another important feature of documentation is the semantics used to describe what is actually being measured. Terms and units that are implicit in one domain are often taken for granted, and not necessarily well recorded for communication with potential users in other fields. For example, the choice of code list, (i.e. determined terminologies of a particular community) to constrain keywords about a data collection might hinder others in finding the data collection because they use other words to say the same thing, or might confuse people expecting something completely different because they use the same word to say something else. Only where semantic mappings between code lists are available can these cross-domain discoveries be made possible and reliable.

Such 'cross-walking' initiatives are very valuable, because, by contrast to free text, which is complicated and laborious to parse and mine, code lists and

restricted vocabularies are extremely valuable ways to speed up the filtering and fitness-for-purpose assessment of datasets. Natural language processing is powerful and becoming more so, as can be seen from the increasing support for automated systems such as chatbots. However, these systems model primarily social contexts, and are not yet coupled to the kind of semantic matching and inference that are needed to distinguish the correct context in which a word is being used to describe an indicator, unit of measure or phenomenon across different scientific fields. For example, if a user is searching globally for datasets that include numerical estimates of uncertainty or variability, they could search for free text descriptions that include terms such as '*variance*', '*standard deviation*', '*ecart-type*' or '*intervalo de confianza*'. However, the presence of such words does not guarantee that variability is indeed mathematically described within the dataset, since, for example, the word 'variance' can also be used in a qualitative sense. By contrast, a URI[6] identifies, via the vocabulary server of the UK's National Environmental Research Council, a definition of 'variance' that is explicitly mathematical and that can be related to other defined statistical concepts, across spoken languages and scientific domains. A similar clarification of terms such as 'sea level' can be seen at the SeaDataNet vocabulary server[7].

For this reason, many classic metadata elements allow free text only for titles and descriptions but require selection from code lists for everything else. We will consider some examples of this practice below, in the section relating to standards. However, there are times when there is no substitute for human-readable material such as manuals and descriptions of research methods, and so methods for adding or linking these to VGI datasets as annotations must be considered. Such documentation can encourage the dissemination of a dataset and might raise the reputation of those who created it – see, for example, the first publication within the newly established geospatial dataset description section of the *International Journal of Spatial Data Infrastructures Research* [8], or the recently launched *Data in Brief* journal[9]. Such documents can convey organisational priorities that are hard to capture otherwise: they can help others to understand the deeper intentions behind why a dataset has been collected, and the reasons for organisational decisions, thereby contributing to the understanding of the overall purpose and potential reusability of a dataset.

Last but not least, it should be considered whether feedback can be collected on the dataset (at whatever level of granularity the packaging allows). Such feedback might include ratings, written statements and references to cases of reuse, but also more direct indications of potential error, identified needs for updating, etc.

2.5 Sharing - With Whom?

The management and curation of datasets not only is an exercise for those gathering and hosting data, but also benefits the users, whether those are the originally-intended beneficiaries or new user groups that find value

in reusing a dataset for their own purposes. Access and use conditions may vary – e.g. depending on privacy and legal issues (see also Chapter 6, Mooney and Minghini, 2017 on privacy, legal issues and ethics), commercial interests, or an organisation's commitment to Open Science. However, VGI can only be exploited to its full potential if these conditions are clearly articulated and, ideally, accompanied by the relevant licences. The decision to integrate or split VGI into collections will have an impact here, since permissions on different elements of a VGI dataset could be different, meaning that different consumers would access different collections of records.

Having persistent identifiers and a minimum set of documentation (including contributors, title and release date) in place also enables proper data citation – an element that should not be underestimated. On the one hand, citable VGI allows clear reuse, since reference can now be made not only to other scientific articles, but also unambiguously to data used within a particular activity. On the other hand, data citation also provides a means of acknowledging the source – thereby contributing to the recognition of the data contributors and owners and providing an incentive for the provision of metadata and curation of VGI. It is likely that new metrics for scientific reputation (altmetrics) will very soon take these achievements into account; the cross-referencing of datasets and the numbers of citations will become essential measures of impact.

3 The Role of Open Standards for VGI Data Management

In the above discussion we have identified a number of crucial practices for ensuring the usability and usefulness of VGI data. A number of tools and protocols exist which can support these practices, and key among these are the various open standards which allow data to be described, structured, exchanged, discovered and documented in ways which best promote interoperability and reuse. In this context, we use the word 'standards' not to denote quality standards, which are addressed in Chapter 7, but agreed schemas, formats and protocols from bodies such as the World Wide Web Consortium (W3C)[10] and OGC[11], which, by virtue of being open for free use, are accessible to a wide range of users across scientific and other domains.

In the following section, the FAIR principles will be used to structure discussion of the tools and approaches that are available. This minimum set of foundational principles originally derives from a 2014 workshop that brought together a wide range of 'academic and private stakeholders all of whom had an interest in overcoming data discovery and reuse obstacles'. The principles have been subsequently developed and refined with the goal of ensuring that 'research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people' – allowing stakeholders to 'more easily discover, access, appropriately integrate and re-use, and adequately cite, the vast quantities of information being generated by contemporary data-intensive

science' (Wilkinson et al., 2016). FAIR is intended to be domain-independent and to be applicable to data archival, management, exploration, discovery and reuse across a range of research fields and scholarly disciplines.

Examples have been chosen from the current practice of the Global Biodiversity Information Facility to illustrate certain sections of FAIR. The reason for this choice is that GBIF is an extremely good example of cross-domain strategic thinking where standards from different fields have been employed, adapted, influenced and developed in order to generate a highly usable, scientifically robust repository of data from hugely varying sources that supports hundreds of high-quality peer-reviewed scientific analyses each year[12].

The FAIR principles are as follows:

> **F1. (meta)data are assigned a globally unique and persistent identifier**
> **F3. metadata clearly and explicitly include the identifier of the data it describes**

As described above, data can only be sensibly shared and reused if the data resource can be identified and reliably retrieved. Persistent identifiers are unique strings of numbers and/or characters that are assigned to a digital resource (e.g. datasets, documents, images) in order to allow long-term, reliable access to that specific item. Persistent identifiers should ideally be managed separately from the physical location of the resource, ensuring the continued accessibility and discoverability of the resource 'no matter how many times the object moves to different servers or property rights owners' (USGS, 2017). *Actionable* persistent identifiers permit access to the resource via a link, which should remain resolvable for the long term. An example that is widely used in the scientific domain is the Digital Object Identifier (DOI; ISO standard 26324:2012)[13], which allows published documents and datasets to be tracked and cited, and which is assigned to journal publications (or prepublications) by CrossRef[14], Figshare[15], Zenodo[16] and other platforms. Recent moves towards data DOIs have been hugely supported by initiatives such as DataCite[17], NOAA's EZID[18], or DryadLab[19], which enable a data producer to mint a DOI and, in some cases, register associated metadata.

An example current practice for VGI is the ability of the GBIF website to produce and maintain a DataCite DOI for a specific user request, guaranteeing that this request can be reliably repeated at a future date. Different query filters (date, type of record, species' scientific name, country, etc.) are collated and stamped with a DOI, which is supplied to the user to ensure future retrieval of records according to the same filters.

A DOI can be allocated at a level of granularity specified by the user, but the maintenance of relationships (e.g. hierarchical 'nestings' of DOIs) is the responsibility of the resource owner, and can be challenging. The ability to discover related datasets in this way is extremely powerful, and can support the Linked Data approach described more fully in the next section. Attention to versioning

is also important: a DOI may represent the final version of a resource, approved for release; an extension or annotation of a resource; or a model/algorithm version used in a reproducible workflow (in this context, a github or subversion version ID can be adapted to fulfil at least some of the role of a DOI). However, there are cases where a DOI will always return 'the latest version' of a resource, and, here, scientific reproducibility is not guaranteed. GBIF DOIs are a good example: the data underlying a query are regularly improved and updated, and historical records may be retrospectively added, meaning that the exact same set of records is not guaranteed to be returned when a DOI is used at a later date.

It is possible to embed dataset identifiers within metadata using existing geospatial metadata standards, such as ISO 19115[20], which offers a CI_Citation element that allows an identifier such as a DOI to be supplied in a structured manner and to be associated with a namespace that can help to ensure the uniqueness of the identifier. However, the real-world practice is less consistent, as evidenced when exploring records in the GEOSS Common Infrastructure (GCI): here, metadata and data identifiers are found in a wide variety of locations within catalogued metadata documents, and are sometimes completely absent. This problem is more cultural than technical: because ISO 19115:2003 is not completely clear about the difference between data and metadata identifiers, and lacks a clear recommendation on the use of Unique and Universal Identifiers (UUIDs), profilers have generated a variety of different identifiers (if they have generated them at all in the first place) and have located these identifiers in at least four different locations within metadata documents (Maso, 2013). The US FGDC metadata standard also allows the encoding of a variety of references to data and metadata[21], but also requires some investment of time and effort for proper use. In the next section we discuss the implications of these standards' complexity for VGI initiatives that may be ephemeral and poorly resourced.

I3. (meta)data include qualified references to other (meta)data
R1.2. (meta)data are associated with detailed provenance

In the above section, we described potential ways in which the identifier of a dataset can be embedded in a traditional geospatial metadata document. However, an important consideration in the context of VGI is the rather complex and laborious nature of generating such 'traditional' metadata documents, which require a significant investment of time and effort. Geospatial metadata standards such as ISO 19115/19157 and FGDC offer a rich and expressive range of descriptive elements, but the reality is that many VGI initiatives are unlikely to generate such detailed documentation. In the face of this reality, other, more lightweight alternatives are likely to be taken up.

In those cases where metadata that are compliant with the ISO standard are generated, there is a huge opportunity for documenting provenance in a

machine-readable way that can, if necessary, encode a full production workflow. The Lineage element of an ISO document, stored as part of the data quality statement, permits the description of any number of processing steps, complete with references to input and output data, descriptions of algorithms of software processing and citations of published reports/articles[22]. Figure 1 shows a single ProcessStep taken from such a lineage statement, rendered in a more human-readable format. It consists of a description of the processing that was carried out, and the three data sources (all of which may be optionally identified with persistent identifiers) that were used in the processing.

The standard and schema implementations of ISO 19115/19157 allow for a series of such ProcessSteps to be combined to generate a highly detailed, and, to some extent, machine-readable description of a dataset's provenance. However, in practice, the rich array of available elements are rarely used as intended, and it is far more common, if a lineage statement is provided at all, to see a single ProcessStep with a long and descriptive text account of the means by which the data were produced. This is in part because of the basic nature of many editing tools for ISO metadata and the lack of best-practice examples, but it is also evidence of the investment required to generate detailed metadata compliant to standards, and of the fact that this investment is not always budgeted into research projects – especially not citizen science projects. The FGDC approach to documenting data provenance is simpler, relying primarily on citations to scientific papers rather than on a fully modular description of the processing, but it is still common to find FGDC-compliant metadata with no real information on data provenance.

An alternative, or potentially a complement, to traditional geospatial metadata is a Linked Data approach (Heath and Bizer, 2011). Here, triples (in the form of subject-predicate-object) are used to describe relationships between entities. This mechanism, further discussed in Section 4.3, extends the potential for resource discovery to off-the-shelf web browsers, rather than just specialised portals and catalogues. Such an encoding, which is, in effect, returning to the roots of Geography Markup Language (GML) – GML version 1.0 came with an encoding in the Resource Description Framework (RDF) – can be adapted to include provenance information on a dataset. This strategy is of particular interest because it could be used to improve or enrich data documentation after data are published, or when they are reused for a different purpose than the original intended use case. For example, user reviews, reports of usage, discovered issues relating to particular observations, spatial regions or observers could be attached, post-hoc, to a published dataset and used in filtering and assessing fitness-for-purpose. Initial research along these lines can be seen in the outputs of the CHARMe project[23], which adapted the proposed OGC Geospatial User Feedback standard (Maso and Bastin, 2015) to permit lightweight annotations to be added to climate data in order to document quality issues, anomalies and user opinions on the value of the data. Another promising approach is the use

ProcessStep
description
Discriminant Analysis (DA) involves a linear combination of the original variables to produce a new set of variables that maximise the statistical difference between the predefined groups. DA acts as a standard classifier (applied to each date) because it enables an unknown pixel to be assigned to one of the predefined classes using discriminant functions obtained from a set of training areas. Training areas were obtained from fieldwork carried out in the Ebro Delta on 29 October 2006 and 17 January 2007. The surface of the training areas collected during fieldwork was 79.7 ha and 40% of them were reserved for an independent test of the results (random sampling).
source
description
Training areas collection (47.8 ha): Several sites representative of each class were visited and georeferenced with the aid of a Global Positioning System (Garmin etreX VistaC, Garmin International, Olathe, KS, USA), the cadastre cartography and the most recently available Landsat image.
source
description
Test areas collection (31.9 ha): Several sites representative of each class were visited and georeferenced with the aid of a Global Positioning System (Garmin etreX VistaC, Garmin International, Olathe, KS, USA), the cadastre cartography and the most recently available Landsat image.
source
description
Each of the 5 previous Landsat-5 images after geometric and radiometric correction and SIGPAC masking

Fig. 1: The content of a ProcessStep in an ISO 19115 metadata document. Namespaces and XML-specific formatting have been removed for clarity.

of the W3C PROV specifications in combination with RDF triples to create queryable databases representing the steps by which a dataset has been generated. A particular advantage of this approach is its amenability to extension when products are derived by some process which needs to be documented. In particular, the documentation of uncertainty introduced by data processing has been explored by Car et al. (2015), who combined UncertML (Williams et al., 2009) – a model and schema for documenting probabilistic uncertainty – with

the PROV-O provenance ontology in such a way that quality issues in multi-part datasets can be encoded, and automated uncertainty propagation is made much more feasible.

F4. (meta)data are registered or indexed in a searchable resource
A2. metadata are accessible, even when the data are no longer available

The geospatial community has widely adopted the use of catalogues, which can be harvested, aggregated and searched in order to yield metadata that in turn reference the location of data resources. In many cases, the data referenced in these metadata documents are no longer available at the specified locations – though this is usually an accidental result of poor curation, rather than a demonstration of conscious compliance with principle A2. The prevalent standard underlying geospatial catalogues is the OGC's Catalogue Service standard[24], of which there are many free and open-source implementations, including the Java-based GeoNetwork and the Python implementation pycsw. Acknowledging that the OGC and SDI community to a large extent complements mainstream Internet developments through specific additions and extensions, the provision of metadata in the form of indexing files for common Internet search engines should also be considered.

A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary

As described above, a variety of free and open standards exist for the search and retrieval of metadata from catalogues through an identifier. In terms of data service protocols, a powerful and widely adopted set of standards has been agreed to and maintained by the OGC: namely, the Web Map Service (for images), Web Feature Service (for data about geospatial objects) and Web Coverage Service (for data about geospatial fields). These standards are widely used, and implemented in a variety of languages and off-the-shelf toolkits such as GeoServer, MapServer, THREDDS and GeoNode, which are free to install and require relatively little configuration effort on the part of a user. When accessing data or imagery via OGC services, a simple HTTP request is parameterised with various user-specified options such as the area of interest and the projection in which the data should be returned. However, it is not specifically the identifier of the data that is used to identify the resource of interest; more commonly, one or more URLs are embedded in the metadata document, incorporating the layer name and namespace and enabling the retrieval of the resource from the service in question, which may not incorporate that

unique identifier at all. For example, a typical WFS request contains a parameter with a namespace and layername defining the data to be retrieved (e.g. '*typeName=lrm:wdpa_latest*'), but there is no requirement to use a persistent identifier for the layer name.

Authorisation and authentication are possible with some implementations of these standards, for example GeoServer[25].

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
R1.3. (meta)data meet domain-relevant community standards
F2. data are described with rich metadata
R1. meta(data) are richly described with a plurality of accurate and relevant attributes

In order to represent the knowledge of data producers, some clear and well structured approaches have been developed. These identify core sets of vital information which **must** be provided, and supplement these cores with optional descriptive elements that can enrich the metadata and assist in assessment of fitness-for-purpose. For example, both ISO and FGDC standards have a subset of compulsory elements without which the metadata are invalid, and a wide array of optional descriptors that can be extremely detailed – for example, reports on quality, representativity, licensing and data provenance. Thus these standards support the generation of rich and informative metadata. In order to make these metadata more easily machine-readable and avoid large amounts of text mining, many elements can be populated with strings selected from code lists, which map to defined meanings in vocabularies and may be further maps to terms in other vocabularies. A good example of this is the 'occurrence issue' vocabulary used by GBIF to describe potential problems with a record, ranging from swapped coordinates to incorrectly inferred country origin for a record. Using values constrained by this list, extremely detailed information about quality assurance can be recorded in a very systematic way, which enables easy filtering and querying of records based on the nature of their errors, and avoids confusion where different assessors might describe an issue using different technical terms[26].

Similar vocabularies have been devised for ISO standards[27] and for taxonomic terms that allow the FDGC standard to be extended to cover biological data[28]. This last point is another strength of these agreed standards: they can be profiled to produce domain-relevant standards, while core elements remain consistent and interoperable with metadata produced using the base standard. In the context of GBIF, the Darwin Core standard, which is fundamental for structuring and harmonising species occurrence data, has been recently extended with new elements that permit the representation of sample data reporting species abundance information[29].

4 Representative Examples of Cross-Community Interoperability Approaches

Following the considerations so far, GBIF has already been considered as a good example to learn from. In addition to some of the highlights of the underlying approach, we see additional value in including two more examples in order to cover a wider spectrum of existing (or emerging) good practices in VGI data management.

4.1 The GBIF Data Publishing Framework

GBIF[30] was founded in 2001 upon a recommendation of the Biodiversity Informatics Subgroup of the Megascience Forum and a subsequent endorsement by the OECD science ministers, to 'enable users to navigate and put to use vast quantities of biodiversity information, advancing scientific research … serving the economic and quality-of-life interests of society, and providing a basis from which our knowledge of the natural world can grow rapidly and in a manner that avoids duplication of effort and expenditure.'[31]

Since then, GBIF has established a renowned cross-community data and metadata infrastructure to function as a single point of access to hundreds of institutions and services offering biodiversity data, based upon a data publishing framework as advised by the GBIF Data Publishing Framework Task Group with the central recommendation that 'all data relevant to the understanding of biodiversity and to biodiversity conservation should be made freely, openly and effectively available' (Moritz et al., 2011). GBIF facilitates responsible use and sharing of data by emphasising the need for proper publishing and citation, and by citing contributing nodes as data curators. It claims to offer data about more than 1.6 million species, collected in 300 years of exploration, from volunteers, researchers and monitoring programmes (see the organisation's 'what is GBIF' website section[32] and the GBIF Data Policy[33]).

As a mature and open infrastructure, the GBIF architecture supports several standards, the most important ones being Darwin Core, Ecological Metadata Language (EML[34]), Access to Biological Collections Data (ABCD[35]) for metadata and also access protocols like TDWG Access Protocol for Information Retrieval (TAPIR[36]) and Distributed Generic Information Retrieval (DiGIR[37]), in order to register and connect hundreds of different data holders and service providers within the GBIF portal. Most of the 'biodiversity standards' are being developed in the context of the Taxonomic Databases Working Group (TDWG)[38].

The principal workflow within the GBIF (2011) infrastructure is described as follows:

1. Digitization: The initial capturing of information in electronic form, through imaging, databasing, maintaining spreadsheets etc.

2. Publishing: The act of making data sources available in a well known format (standard) and with appropriate metadata for access on the internet.
3. Integration: The process of aggregating published datasets, applying consistent quality control routines and normalizing formats.
4. Discovery and access: By building network wide indexes, discovery services are offered for users through portals and for machines by extensive web service APIs (GBIF, 2011).[39]

In order to collect standardised information from contributing nodes, GBIF offers its community several tools, the most prominent one being the Integrated Publishing Toolkit (IPT):

> The IPT's two primary functions are to
> 1) encode existing species occurrence datasets and checklists, such as records from natural history collections or observations, in the Darwin Core standard to enhance interoperability of data, and
> 2) publish and archive data and metadata for broad use in a Darwin Core Archive, a set of files following a standard format (Robertson et al., 2014).

A further functionality is the possibility to convert metadata into 'data papers' that may be published as peer-reviewed scholarly articles in a journal. This is a direct incentive for publishing, as data can then be cited, raising the profile of the researcher or institution[40]. It also encourages the user to directly choose a public domain licence for the data (which is in line with GBIF's data policy and also leads to easier reuse of the data; see FAIR principles in previous section).

The Integrated Publishing Toolkit is one prominent example of how GBIF tries to lower the barriers for new data publishers and to promote this community's standards.

4.2 The OGC Interoperability Program, Cross Community Interoperability

VGI data often lack a common understanding associated to the meaning of the data or are user-contributed without any specific purpose, via social media platforms such as Twitter and Flickr. Nonetheless, often these data contain geographic reference and are *tagged* with other useful and queryable information, and the social media platforms offer application programming interfaces (APIs) to harvest from their services. In photo-community platforms, for example, the position of the published image may be (sometimes unintentionally) recorded in the GPS tags of EXIF metadata. This is likely to increase with the widespread use of smartphones equipped with capable GPS sensors. These sensors may eventually provide even more sophisticated information – for example, orientation

and tilt angle of the camera. Such ancillary information is useful in a wide variety of use cases: for example as additional 'ground truth data' in the validation of global land cover products, or as one source among others in realtime crisis management. Several authors (Goodchild, 2007; Jürrens et al., 2009; Schade et al., 2011) have suggested viewing citizens [or humans] as *sensors* and using the OGC Sensor Web Enablement (SWE) as a reference framework to describe these *sensors* and their readings (or *observations*). In short, this framework aims at making sensor readings of all kinds discoverable and accessible via the net as near real-time streams in a standardised way, thus allowing for e.g. additional information streams beyond authoritative data from satellite images (in the case of crisis response for example). The SWE consists of a set of relevant standards, for example:

- *O&M* – Observations and Measurements: This standard describes the general data model and specifies XML encodings on how to represent data.
- *SOS* – Sensor Observation Service: The standard description of the service offering sensor descriptions and their observations.
- *SensorML* – Sensor Model Language: The standard models and XML Schema for describing the processes within sensor and observation processing systems.

(See the OGC website's Sensor Web Enablement description[41] for details.)

The data model of O&M is generic in the sense that its core element, an observation event, can be mapped against all kinds of physical properties:

'An observation is an act associated with a discrete time instant or period through which a number, term, or other symbol is assigned to a phenomenon. It involves application of a specified procedure, such as a sensor, instrument, algorithm, or process chain. The procedure may be applied in situ, remotely, or ex situ with respect to sampling location. The result of an observation is an estimate of the value of a property of some feature' (Cox, 2013).

In a series of so-called testbeds, the OGC Interoperability Program (IP) addresses fundamental questions regarding testing, prototyping and early adoption of OGC standards. These testbeds consist of several threads in specific application domains, such as aviation. In one of these threads – on Cross-Community-Interoperability (CCI) – the OGC has taken up the idea of mapping VGI information against the O&M data model (see *testbed 10 CCI VGI Engineering report* (OGC, 2014)). By transforming social media content into the O&M data model, the data can further be served by OGC service components in a standardised way, as observations made by the human observer, by using the Sensor Observation Service (SOS). The testbed report also states some real-world problems – since the prototype was tested against several clients, some of which could not deal with the SOS interface (at the time of writing SOS is not yet as widespread as the Web Feature Service (WFS) interface), the data were also encoded as features for usage within a WFS. In this scenario,

the social media content was harvested by using the REST interface of the service (Flickr in their example) and uploaded as observations to the SOS after being transformed into the O&M model. This development was taken up as 'SWE for Citizen Science' as part of the discussions that led to the proposal of a new OGC Domain Working Group on Citizen Science (that was adopted at the OGC Technical Committee Meeting in September 2016).

4.3 The Provision of OpenStreetMap (OSM) as Linked Data

An interesting case builds on one of the most prominent VGI initiatives so far: OpenStreetMap (OSM). In the provision of OSM as Linked Data (Stadler et al., 2012), the traditional OSM dataset gets translated into a model that implements the Linked Data paradigm using RDF. Technically, the OSM data are periodically extracted from the official web page (openstreetmap.org), transformed into an RDF representation and loaded into a publicly available triple store that is essentially an RDF database. This processing is enabled by the open licensing model of OSM.

Apart from changing the data model (i.e. data formats and structures that are used to encode the points, lines, polygons, etc. that are used within OSM), the transition to a Linked Data approach also provides a step change in respect to (semantic) interoperability. While OSM defined its own structures and map elements (features) that are at most known to its own community, RDF is a recognised standard of the W3C and thereby well known to web developers around the globe, i.e. far beyond the original OSM contributors and the geospatial community. As such, datasets that are translated to so-called RDF triples (subject-predicate-object) can be easily connected to other triples by adding standard or self-defined relationships. In this way, datasets from multiple providers become interconnected and can be cross-navigated within the Linked Data Cloud[42].

In addition to introducing a standard way of modelling and related encodings, RDF also provides the possibility to reuse existing vocabularies so that the expressions used to represent subjects, predicates and objects are understood by many different communities (and not only by those that are familiar with a particular VGI dataset, such as, in this case, OSM). Considering geospatial data, for example, one might use the Location Core Vocabulary[43] for describing any place in terms of its name, address or geometry. In a similar manner vocabularies exist to describe persons and their social network[44] or even relationships between terms in two different vocabularies[45]. The most important point here is that the use of RDF is a well established step to breaking down the silos between closed communities, such as the SDI or the VGI community (see also Schade and Smits, 2012). Compared to many current OGC standards, which mostly evolve in parallel worlds, RDF provides common grounds for all sorts of different communities. This is because RDF builds on the (semantic)

web as the common denominator and enables the specification of community-specific vocabularies, together with shared terms and well defined mappings. The mechanisms of vocabulary reuse and matching avoid the need for additional architectural approaches to join information from separately operating communities, such as wrappers, brokers or proxies.

While the above holds for all data models, it particularly also holds for models of data quality. Returning to the concrete example of OSM, the overall quality assurance and data management mechanisms remain core business within the traditional platform that underlies OSM (available from openstreetmap.org). The architecturally loosely coupled Linked Data representation adds, for example, the possibility to apply W3C vocabularies related to data quality – most notably the W3C Data on the Web Best Practices: Dataset Quality Vocabulary (W3C, 2016a) and Data Usage Vocabulary (W3C, 2016b). Whereas DQV provides the means to describe 'the quality of a dataset …, whether by the dataset publisher or by a broader community of users' (W3C, 2016a), DUV specifies 'a number of foundational concepts used to collect dataset consumer feedback, experiences, and cite references associated with a dataset' (W3C, 2016b). Together, both vocabularies could also be used for VGI, in order to support providers to express quality parameters of their offerings, but also to enable users to add their experiences and feedback to these parameters.

Yet, at the time of writing, both of these best practices are only available in draft versions and so far (to our knowledge) we still lack tangible access to using this concrete approach in a VGI context. We consider it as an extremely exciting area that is worth exploring (and comparing to dedicated OGC-centric approaches) in respect to VGI data management. The example of OSM as Linked Data may be the most straightforward use case for testing these possibilities.

5 Conclusion

In this chapter, we have looked into some generic – and not only VGI projects-specific – principles and good practices of data management, with the central paradigm being the FAIR principle: data should be findable, accessible, interoperable and reusable. To be reusable, it is vital that (meta)data are released with a clear and accessible data usage licence (see Chapter 6, Mooney and Minghini, 2017). Furthermore, we have summarised standards that support these principles, both from the Open Geospatial Consortium and from ISO TC/211, as well as from W3C, and we have investigated three examples where these principles and standards are utilised to maximise cross-discipline interoperability.

A key conclusion from this review into the current state of the art is that metadata for VGI are, and are likely to remain, patchy and extremely heterogeneous. 'Traditional' standards aimed at complete documentation of a one-

off production workflow, such as ISO 19115/19157, are rich in descriptive elements that, if used properly, can enable the provenance and quality of geospatial data to be documented in very useful and machine-readable ways that support uncertainty propagation and fitness-for-use assessment. However, an investigation of open geospatial catalogues quickly shows that these standards are not being exploited to their full potential, even by large institutional data producers – partly because of the resource-intensive nature of metadata generation, and partly because of an ongoing shortage of tools and examples to simplify the process. For VGI, where even a single 'dataset' can contain observations produced by a wide variety of observers, instruments and methods, such monolithic standards may only be of use for periodic review and documentation of aggregated and quality-controlled data. In addition, the nature of VGI is such that observations may be accessed and used in a variety of different combinations and groupings. With such a fluid granularity, tools and APIs that allow annotation and documentation of individual records or groups of records are likely to be more useful, as are any tools and processing methods that permit the collection and storage of metadata automatically at the point of observation. Ongoing developments in RDF and Linked Data appear very promising for supporting data annotation, but are still too immature to be easily usable within most VGI initiatives. However, this is a key angle of research that should be developed, not least because the annotation/commentary approach to metadata permits information and quality reports to be attached to data after their production, so that VGI can be mobilised and made more usable and reusable.

We have not looked into software solutions of how to access, store and back up data, for example which database management solution to use, such as PostgreSQL (with its language extension PostGIS), MySQL or the lightweight SpatiaLite, to name a few. We have also only touched the surface of the topic of software suites like GeoServer, deegree or GeoNetwork, all of which offer substantial building blocks for Spatial Data Infrastructures. We encourage the use of Open Source software like these, as well as open and freely accessible standards.

In this text we have not addressed Environmental Sensor Networks (ESNs) that may comprise a backbone in data assessment from distributed heterogeneous sensors. We expect that the Sensor Web Enablement, as an OGC reference framework, will play an important role in citizen sensing. For further reading, the FP7 funded Citizen Observatory 'COBWEB' has defined a 'Generic Infrastructure Platform to facilitate the collection of Citizen Science data for Environmental Monitoring'(Higgins et al., 2016).

In terms of actual formulation of Data Management Plans, substantial resources are available; see for example DataOne's 'Data Management Guide for Public Participation in Scientific Research'[46] or COBWEB's 'Generic Data Management Plan Check' in their 'deliverable 7.1 on Data Management Guidelines.'[47]

Data management methodologies can only succeed if their benefits overcome their implementation costs; i.e. existing solutions and best practices will have to be tailored to the needs and capabilities of individual projects, and feasibility needs to be assessed on a case by case basis. However, it is imperative to recognise that a precise knowledge of the provenance and meaning of data is a most precious asset that should be highly valued.

Notes

[1] https://ec.europa.eu/programmes/horizon2020/en/news/citizens%E2%80%99-observatories-empowering-european-society-open-conference
[2] http://www.eubon.eu/
[3] https://cobwebproject.eu/
[4] https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf
[5] http://www.bfe-inf.org/info/data-principles
[6] E.g. http://vocab.nerc.ac.uk/collection/P15/current/CFCM0010/
[7] E.g. http://seadatanet.maris2.nl/v_bodc_vocab_v2/vocab_relations.asp?lib=P02
[8] http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/389
[9] https://www.journals.elsevier.com/data-in-brief
[10] https://www.w3.org/
[11] http://www.opengeospatial.org/
[12] http://www.gbif.org/mendeley
[13] http://www.iso.org/iso/catalogue_detail?csnumber=43506
[14] https://www.crossref.org/
[15] https://figshare.com/
[16] https://zenodo.org/
[17] https://www.datacite.org/
[18] http://ezid.cdlib.org/
[19] http://datadryad.org/
[20] http://www.iso.org/iso/catalogue_detail.htm?csnumber=53798
[21] http://www.ngdc.noaa.gov/wiki/index.php/Data_Set_Identifiers_and_other_Unique_IDs
[22] https://geo-ide.noaa.gov/wiki/index.php?title=File:LI_Lineage-2.png
[23] http://charme.org.uk/
[24] http://www.opengeospatial.org/standards/cat
[25] http://docs.geoserver.org/stable/en/user/security/service.html
[26] http://gbif.github.io/gbif-api/apidocs/org/gbif/api/vocabulary/OccurrenceIssue.html
[27] https://geo-ide.noaa.gov/wiki/index.php?title=ISO_19115_and_19115-2_CodeList_Dictionaries

[28] http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/biodatap.pdf
[29] http://www.gbif.org/sites/default/files/gbif_IPT-sample-data-primer_en.pdf
[30] http://www.gbif.org
[31] http://www.gbif.org/what-is-gbif#background
[32] http://www.gbif.org/what-is-gbif
[33] http://www.gbif.org/resource/80527
[34] https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
[35] http://www.tdwg.org/activities/abcd/
[36] http://www.tdwg.org/activities/tapir/
[37] http://digir.sourceforge.net/
[38] http://www.tdwg.org/standards/
[39] http://www.gbif.org/infrastructure/summary
[40] http://www.gbif.org/publishing-data/data-papers
[41] http://www.opengeospatial.org/ogc/markets-technologies/swe
[42] http://lod-cloud.net/
[43] https://www.w3.org/ns/locn
[44] http://www.foaf-project.org/
[45] https://www.w3.org/2004/02/skos/
[46] https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf
[47] https://cobwebproject.eu/sites/default/files/COBWEB%20D7.1%20Data%20Management%20Guidelines%20v1_0.pdf

Reference list

Bearden, M.J., 2007. The National Map Corps. Presented at the Specialist Meeting on Volunteered Geographic Information, University of California at Santa Barbara. Available at http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Bearden_paper.pdf [Last accessed 1 May 2017]

Car, N., Cox, S., Fitch, P., 2015. *Associating uncertainty with datasets using Linked Data and allowing propagation via provenance chains*. EGU General Assembly Conference Abstracts, p. 4392. Available at http://adsabs.harvard.edu/abs/2015EGUGA..17.4392C [Last accessed 1 May 2017].

Global Biodiversity Information Facility (GBIF), 2011. GBIF Position Paper on Data Hosting Infrastructure for Primary Biodiversity Data. Version 1.0. (Authored by Goddard, A., Wilson, N., Cryer, P., & Yamashita, G.) Available at http://www.gbif.org/resource/80733 [Last accessed 1 May 2017]

Cox, S., 2013. OGC and ISO 19156:2011(E) OGC Abstract Specification Geographic information — Observations and measurements. Available at http://portal.opengeospatial.org/files/?artifact_id=41579 [Last accessed 1 May 2017].

Craglia, M., Granell, C., 2014. *Citizen Science and Smart Cities*. Publications Office of the European Union, Luxembourg. DOI: https://doi.org/10.2788/80461

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 211–221. DOI: https://doi.org/10.1007/s10708-007-9111-y

Heath, T., Bizer, C., 2011. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology* 1, 1–136. DOI: https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Higgins, C.I., Williams, J., Leibovici, D.G., Simonis, I., Davis, M.J., Muldoon, C., van Genuchten, P., O'Hare, G., Wiemann, S., 2016. Citizen OBservatory WEB (COBWEB): A Generic Infrastructure Platform to Facilitate the Collection of Citizen Science data for Environmental Monitoring. *International Journal of Spatial Data Infrastructures Research* 11, 20–48. DOI: https://doi.org/10.2902/1725-0463.2016.11.art3

Jürrens, E.H., Bröring, A., Jirka, S., 2009. A Human Sensor Web for Water Availability Monitoring, in *Proceedings of OneSpace 2009 – 2nd International Workshop on Blending Physical and Digital Spaces on the Internet*, Berlin, Germany, 1 Sep 2009. Available at: http://onespace.ace.ed.ac.uk/2009/docs/onespace2009_submission_2.pdf [Last accessed 16 May 2017].

Maso, J., 2013. Unique Identifiers within Systems of Systems, Slide 8: Id's in metadata records. Metadata and data id's. GEOSS Future Products Workshop, NOAA, Silver Springs, USA, 27 March 2013. Available at: https://portal.opengeospatial.org/files/?artifact_id=53340 [Last accessed 1 May 2017].

Maso, J., Bastin, L., 2015. OGC Geospatial User Feedback Standard. Conceptual Model and XML Encoding Extension. Available at: http://www.opengeospatial.org/standards/requests/144 [Last accessed 1 May 2017].

Mooney, P and Minghini, M. 2017. A Review of OpenStreetMap Data. 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. 37–59. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbf.c.

Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., Cockerill, M., Chavan, V., 2011. Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics* 12, 1–10. DOI: https://doi.org/10.1186/1471-2105-12-S15-S1

OGC, 2014. OGC Testbed-10 CCI VGI Engineering Report. Edited by A. Broring, S. Jirka, M. Rieke, B. Pross. Available at: https://portal.opengeospatial.org/files/?artifact_id=58925 [Last accessed 1 May 2017].

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L., Desmet, P., 2014. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLOS ONE* 9, e102623. DOI: https://doi.org/10.1371/journal.pone.0102623

Schade, S., Díaz, L., Ostermann, F., Spinsanti, L., Luraschi, G., Cox, S., Nuñez, M., Longueville, B.D., 2011. Citizen-based sensing of crisis events: sensor

web enablement for volunteered geographic information. *Applied Geomatics* 5, 3–18. DOI: https://doi.org/10.1007/s12518-011-0056-y

Schade, S., Smits, P., 2012. Why linked data should not lead to next generation SDI, in: 2012 IEEE International Geoscience and Remote Sensing Symposium. Presented at the 2012 IEEE International Geoscience and Remote Sensing Symposium, pp. 2894–2897. DOI: https://doi.org/10.1109/IGARSS.2012.6350721

Schade, S., Tsinaraki, C., 2016. *Survey report: data management in Citizen Science projects*. JRC Technical Report; EUR 27920 EN, Luxembourg (Luxembourg), Publications Office of the European Union. DOI: https://doi.org/10.2788/539115.

Socientize, 2013. White Paper on Citizen Science in Europe. Available at http://www.socientize.eu/?q=eu/content/download-socientize-white-paper [Last accessed 1 May 2017].

Stadler, C., Lehmann, J., Höffner, K., Auer, S., 2012. LinkedGeoData: A core for a web of spatial open data. *Semantic Web* 3, 333–354. DOI: https://doi.org/10.3233/SW-2011-0052

USGS, 2017. Persistent Identifiers. USGS Data Management. 17 February 2017. Available at https://www2.usgs.gov/datamanagement/preserve/persistentIDs.php [Last accessed 4 April 2017].

W3C, 2016a. Data on the Web Best Practices: Data Quality Vocabulary. Edited by A. Albertoni, A. Isaac. Available at: https://www.w3.org/TR/vocab-dqv/ [Last accessed 1 May 2017].

W3C, 2016b. Data on the Web Best Practices: Data Usage Vocabulary. Edited by B Farais-Losco, E. Stephan, S. Purohit. Available at: https://www.w3.org/TR/vocab-duv/ [Last accessed 1 May 2017].

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L.B. da S., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Williams, M., Cornford, D., Bastin, L., Pebesma, E. 2009. Uncertainty Markup Language (UncertML). OGC Discussion Paper, Document Number: 08–122r1. Available at: http://www.opengeospatial.org/docs/discussionpapers [Last accessed 16 May 2017].