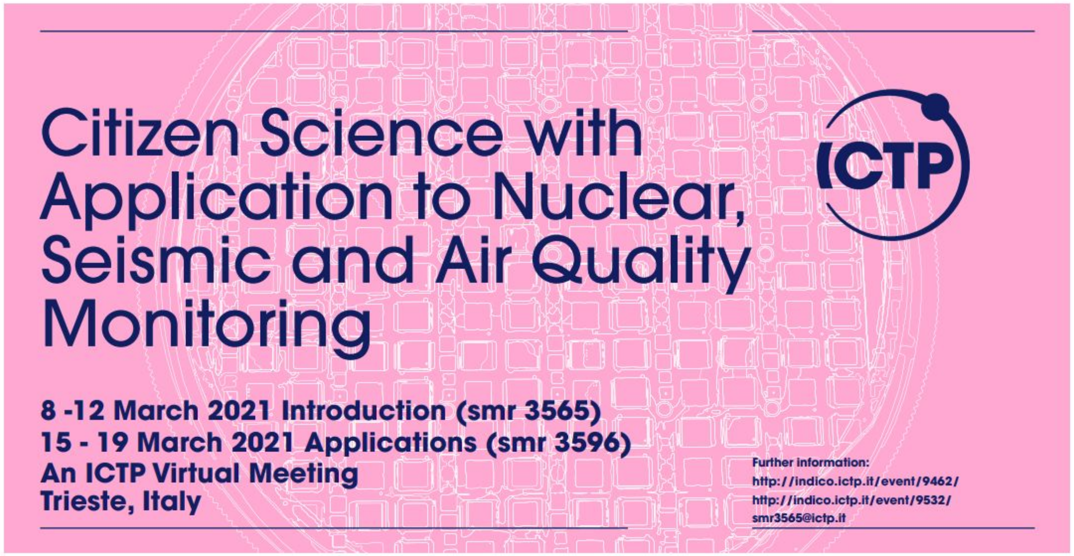
Data quality in Citizen Science

Peter Mooney

Maynooth University
National University of Ireland Maynooth

Lucy Bastin

Aston University
BIRMINGHAM UK



And today we shall talk about ….

The multiple meanings and dimensions of citizen science – how it provokes different responses in scientists (Peter)

Commonly encountered issues around data quality (Peter)

Data Quality and Metadata – why are concepts such as metadata, vocabularies, testing, observation level quality, etc. important for the advancement of Citizen Science. (Lucy)

Introducing PPSR Core (Public Participation in Scientific Research) and the **OGC Citizen Science Interoperability Experiment** (Lucy)

Conclusions and discussions (Peter and Lucy)

Data quality in Citizen Science has different meaning for different people and use cases

Fitness for use?

Fitness for purpose?

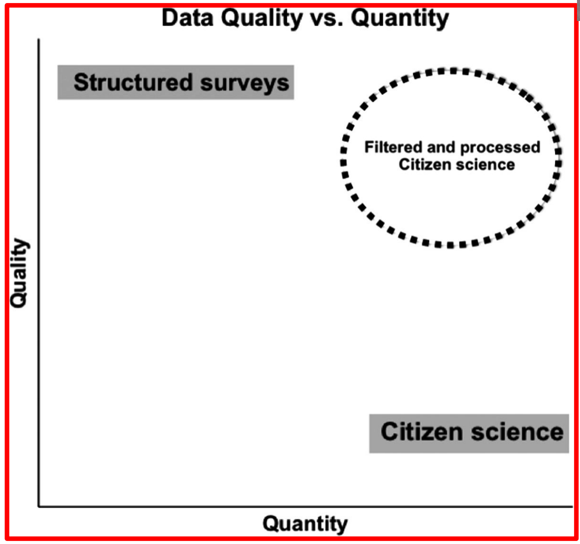
Who baked the cake?

How was the cake baked?

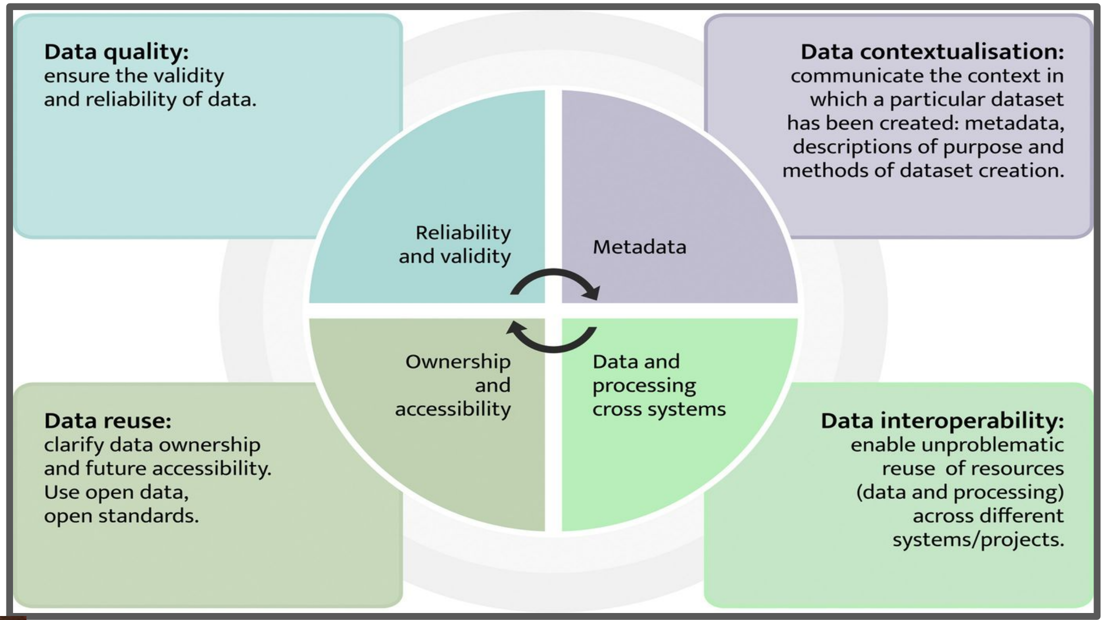
Can I **compare** it to other cakes?

The huge remit of Citizen Science Data Quality

Example for context: citizen air quality monitoring in cities



https://doi.org/10.1111/ddi.13068



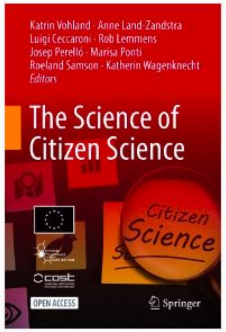
https://doi.org/10.1007/978-3-030-58278-4_8

https://andrewsheppard.net/research/quality-citizen-science/

So how did Lucy and I arrive here?

link.springer.com/book/10.1007/978-3-030-58278-4

Springer Link



The Science of Citizen Scien

Editors (view affiliations)

Katrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni, Rob

Marisa Ponti, Roeland Samson, Katherin Wagenknecht

Open Access | Book

Download book PDF

Download book EPU

Part I Citizen Science as Science

Chapter 8
Data Quality in Citizen Science

Check for updates

Bálint Balázs, Peter Mooney, Eva Nováková, Lucy Bastin, and Jamal Jokar Arsanjani

Abstract This chapter discusses the broad and complex topic of data quality in citizen science – a contested arena because different projects and stakeholders aspire to different levels of data accuracy. In this chapter, we consider how we ensure the validity and reliability of data generated by citizen scientists and citizen science projects. We show that this is an essential methodological question that has emerged within a highly contested field in recent years. Data quality means different things to different stakeholders. This is no surprise as quality is always a broad spectrum, and nearly 200 terms are in use to describe it, regardless of the approach. We seek to deliver a high-level overview of the main themes and issues in data quality in citizen science, mechanisms to ensure and improve quality, and some conclusions on best practice and ways forwards. We encourage citizen science projects to share insights on their data practice failures. Finally, we show how data quality assurance gives credibility, reputation, and sustainability to citizen science projects.

Keywords Peer verification · Expert verification · Quality assessment

https://doi.org/10.1007/978-3-030-58278-4

Several factors combine to make structuring of data quality in citizen science challenging

- New citizen science projects appear daily, the academic literature grows so quickly
- 'The Knock-on Effect' of existing projects are taking different approaches to data quality and data sharing then makes follow-on projects problematic (including reproducibility)
- Different projects consider different dimensions of data quality
- Most citizen science projects have multiple goals ad must all deal with the 'legitimacy' argument waged against them by certain stakeholders

Two objective task independent measures of data quality that prompt the most professional skepticism are accuracy and bias.

"Despite the wealth of information emerging from citizen science projects, **the practice is not universally accepted as a valid method of scientific investigation**" (Bonney et al, 2014) DOI: 10.1126/science.1251554

"**Most types of bias found in citizen-science datasets are also found in professionally produced datasets** and can be mitigated using existing statistical tools" (Kosmala et al, 2016) doi: 10.1002/fee.1436

"The only known bias specific to citizen science is the potentially high variability among volunteers in terms of demographics, ability, effort, and commitment." (Kosmala et al, 2016)

Data Quality in Citizen Science – a multi-dimensional problem?

Lukyanenko et al (2016) https://doi.org/10.1111/cobi.12706

“caution is warranted in emphasizing a particular dimension of data quality in citizen science projects; ***trade-offs in different dimensions of data quality are inevitable***”

> We contend that in trying to hold amateurs to scientific standards, researchers not only ask nonexperts to perform often unrealistic tasks, but also risk missing the opportunity to fully engage with people in the core objective of discovery. The emerging problem of quality in citizen science is, therefore, writing a story in which citizens contribute to the plot.

Some key readings in Citizen Science Data Quality – for self study after the workshop

Wiggins et al. (2011) "**Mechanisms for Data Quality and Validation in Citizen Science**" https://doi.org/10.1109/eScienceW.2011.27

Hochachka et al (2012) "**Data-intensive science applied to broad-scale citizen science**" https://doi.org/10.1016/j.tree.2011.11.006

Sullivan et al. (2014) "**The eBird enterprise: An integrated approach to development and application of citizen science**" https://doi.org/10.1016/j.biocon.2013.11.003

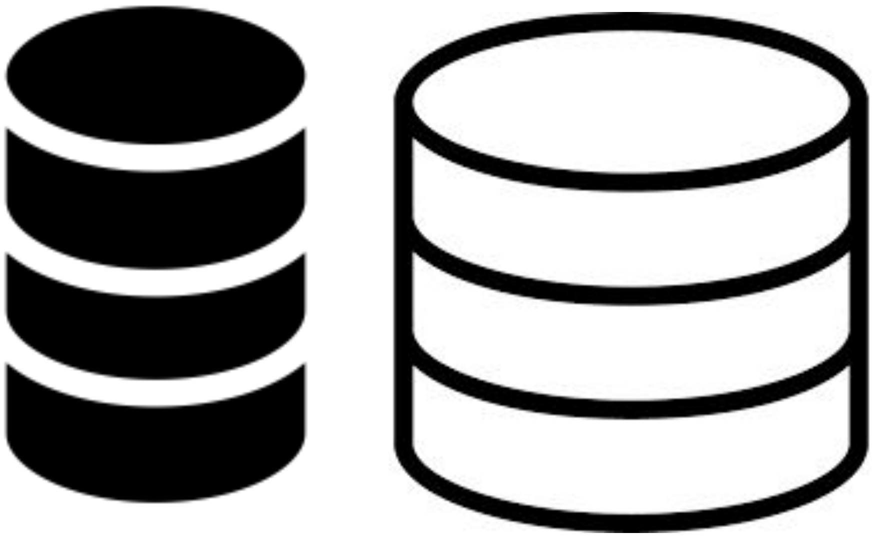
Burgess et al. (2017) "**The science of citizen science: Exploring barriers to use as a primary research tool**" https://doi.org/10.1016/j.biocon.2016.05.014

Fraisl et al. (2020) "**Mapping citizen science contributions to the UN sustainable development goals**" https://doi.org/10.1007/s11625-020-00833-7

If a dataset was not explicitly identified as Citizen Science how would you know?

Given two datasets, how could you tell which is the professional dataset and which is the citizen science dataset?

Suppose the dataset(s) are PM10 concentration measurements (hourly) in a city of population 50,000

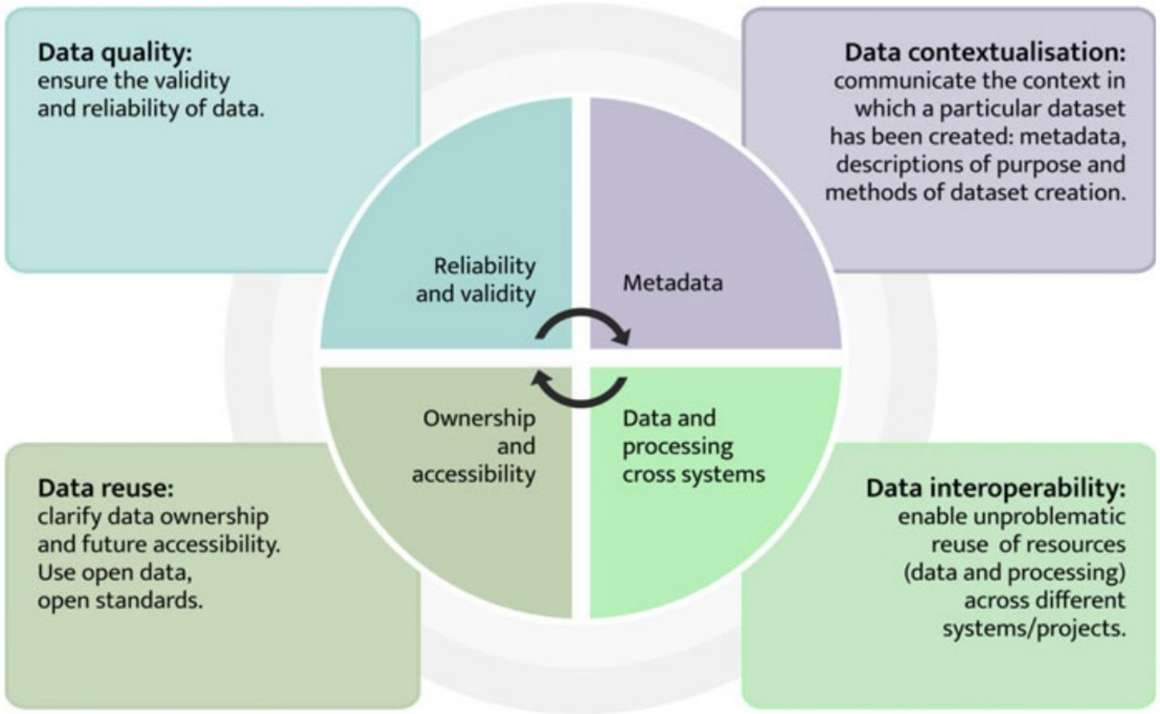


Blind tasting removes many opportunities for bias and levels the playing field for all wines to receive the same analysis without any preexisting expectations.

Data as a risk factor in Citizen Science

Data from citizen science is unparalleled as it represents evidence that is otherwise difficult for professional science to generate or obtain.

For every stakeholder in citizen science, there appears to be a different definition of what constitutes data quality from an epistemological point of view, the question is how accurately does the data represent the real-world constructs to which they refer.



Kosmala et al (2016) **Questions to consider when evaluating citizen science projects for data quality**

- Does the project use **iterative design**?
- How **easy** or **hard** are the tasks?
- How systematic are the **task procedures** and data entry?
- What **equipment** are volunteers using?
- Does the project record relevant **metadata**?
- Are **good data management practices** used?
- Are the **data appropriate for the project's management objectives** or research questions?
- Does the project assess data quality by **appropriate comparison with professionals**?
- Is **collection effort standardized** or accounted for in data analysis?

Cross-section of the most commonly encountered issues around data quality in citizen science

1. Data collection **protocols are not followed by participants**.

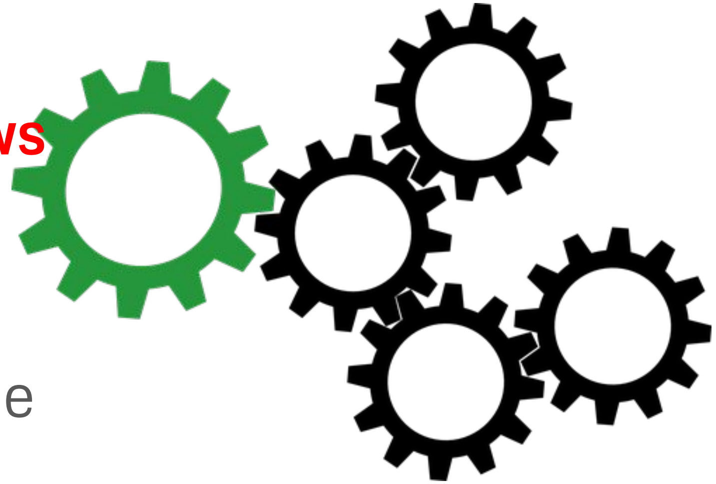
2. Data collection **protocols do not match the goals of the project or the probable participants**.

3. Data collection **protocols are incorrectly implemented**.

4. Data collection **protocols are not comprehensive** and are used by stakeholders with **different data quality expectation levels**.

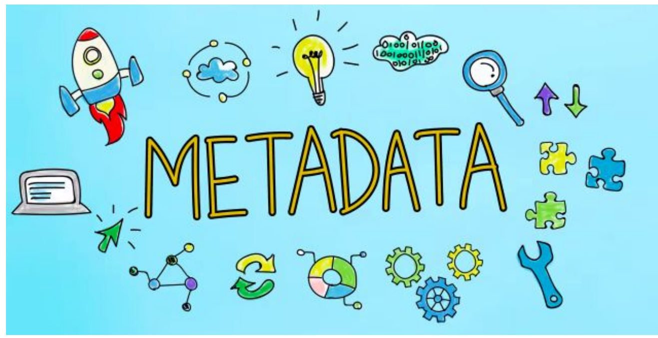
5. **Data used are not fit for purpose**.

Protocols, data quality processes, observations, etc. are linked by metadata



Metadata is what makes protocols happen, it allows us to 'describe' the processes, record experiences, make systems & data interoperable etc.

As two Computer Scientists we appreciate the value of metadata (*but we also know that most practitioners find it very boring*)



For the remainder of the presentation we turn our focus to metadata.

Discovering data… and metadata

There is huge potential for citizen science data to be combined together, and with other data, to understand earth systems and human impacts in a more powerful way.

This approach might cross traditional disciplinary boundaries…

… for example,

a museums project interpreting historic painting and documents might be combined with modern datasets on weather, air quality and health to uncover trends and patterns.

Discovering data… and metadata

But to do this, we need to understand the nature and quality of all the data sources.

e.g.

What's being measured / recorded / observed, how and where?

and…

What measures are being taken to ensure a certain level of quality?

The importance of metadata

Some useful elements for quality evaluation:

- **completeness, consistency and representativity**: do observers sample at random or according to some plan?

- **accuracy and precision**: are the volunteers trained, and is their data double-checked?

If metadata communicates this provenance, we can decide whether it’s scientifically **appropriate** to re-use datasets.

Ideally, the metadata needs some level of machine-readability

- and **interoperability.**

In the wider scientific field, there are several standards that try to achieve interoperability for data and metadata.

You can record the quality of a geospatial dataset (along with many other dataset characteristics) with a standard like ISO 19115 or FGDC.

Typically, an XML document, with structured information embedded in it.

http://www.fao.org/geonetwork/srv/en/iso19139.xml?id=37134


```xml
▼<gmd:dataQualityInfo>
  ▼<gmd:DQ_DataQuality>
    ▼<gmd:scope>
      ▼<gmd:DQ_Scope>
        ▼<gmd:level>
            <gmd:MD_ScopeCode
            codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_ScopeCode"
            codeListValue="dataset"/>
          </gmd:level>
        </gmd:DQ_Scope>
      </gmd:scope>
    ▼<gmd:lineage>
      ▼<gmd:LI_Lineage>
        ▼<gmd:statement>
            <gco:CharacterString>Due to the map generation method, the quality of the map can never be
            uniform. The overall quality of the map depends heavily on the individual quality of the
            data for the different countries.</gco:CharacterString>
          </gmd:statement>
        </gmd:LI_Lineage>
      </gmd:lineage>
    </gmd:DQ_DataQuality>
  </gmd:dataQualityInfo>
```

Metadata for citizen science

Historically not standardised.

Can be laborious to produce, especially for small projects with little resource.

Often very descriptive, but can contain a wealth of useful information.

The challenge is to discover, harmonise and interpret that information.

dataQualityAssuranceMethod	-Data owner curated -Subject matter expert record verification -Crowd-sourced record verification -Record annotation -System supported data attribute configuration -No DQ methods used -Not applicable

A set of possible labels for citizen science to describe how data QA was carried out.

Work in progress – more on this example later
https://core.citizenscience.org/

Does dataset-level quality make sense?

Many citizen science repositories are not static 'datasets'

They can be 'sliced and diced' and queried in a range of ways.

GBIF
Global Biodiversity Information Facility

Download details

IDENTIFIER	DOI doi:10.15468/dl.wjrus4	
CITE AS	GBIF.org (12th July 2015) GBIF Occurrence Download http://doi.org/10.15468/dl.wjrus4	
QUERY	TAXON	*Ruwenzorornis johnstoni (Sharpe, 1901)*
	COUNTRY	*Rwanda*
	GEOREFERENCED	*true*
FORMAT	DwCA	
STATUS	Preparing	

4 datasets contributed data to this download

DATASET	rmca-albertine-rift-birds
RECORDS	35 records from this dataset included at time of download
IDENTIFIER	doi:10.15468/i2phti
CITATION	BeBIF Provider: rmca-albertine-rift-birds

DataCite

DATASET	EOD - eBird Observation Dataset
RECORDS	6 records from this dataset included at time of download
IDENTIFIER	doi:10.15468/aomfnb
CITATION	2013. EOD - eBird Observation Dataset.

DATASET	Royal Museum of Central Africa - Albertian Rift Birds (ENBI wp13)
RECORDS	35 records from this dataset included at time of download
IDENTIFIER	doi:10.15468/evhiqt
CITATION	BeBIF Provider: Royal Museum of Central Africa - Albertian Rift Birds (ENBI wp13)

DATASET	iNaturalist research-grade observations
RECORDS	1 records from this dataset included at time of download
IDENTIFIER	doi:10.15468/ab3s5x
CITATION	iNaturalist.org: iNaturalist research-grade observations

Variability among volunteer weather stations…
7 typical examples, co-located with a gold-standard weather station.

Bell, S, Cornford, D & Bastin, L, 2015. Weather, 70 (3), pp. 75-84

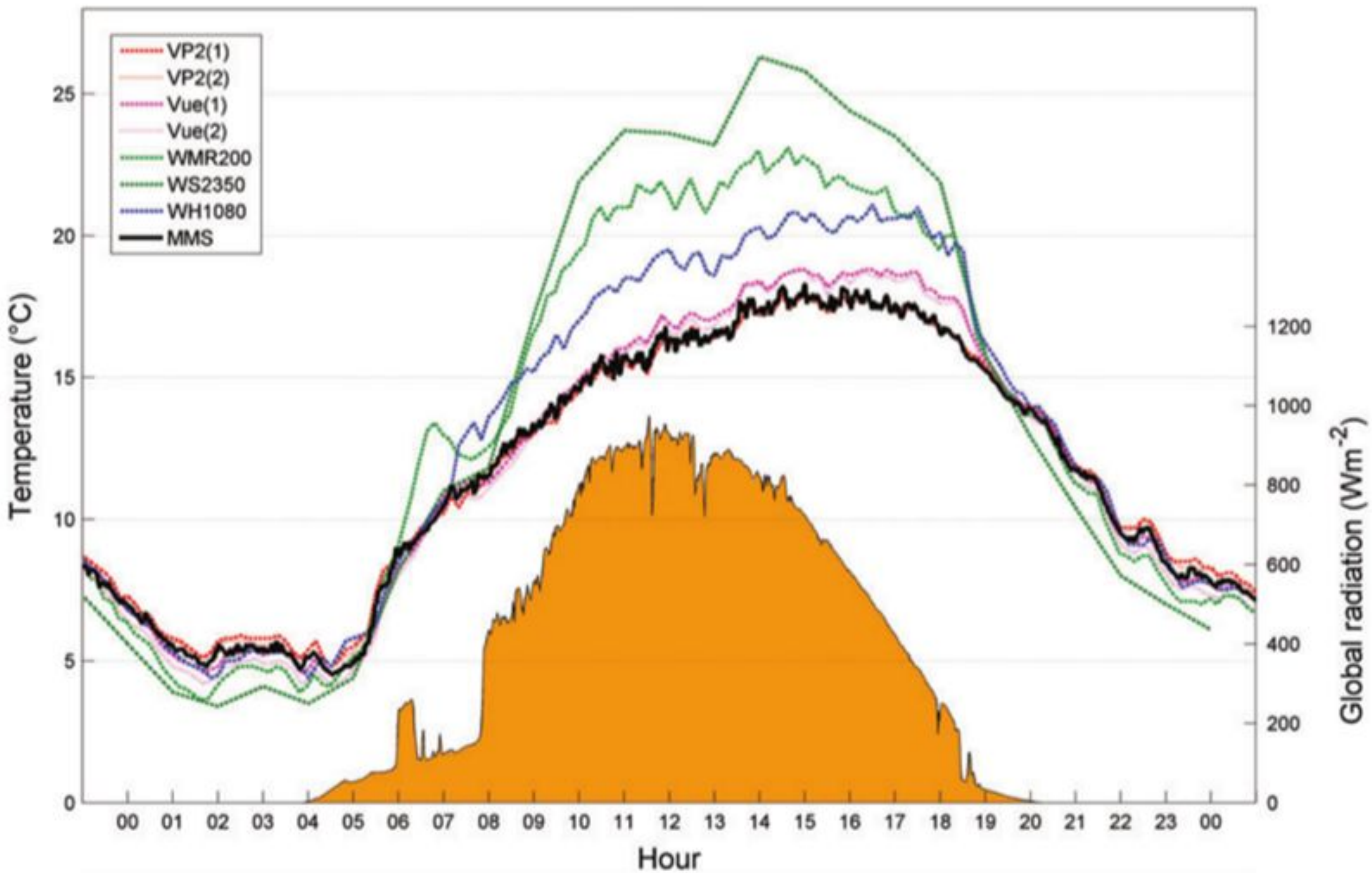
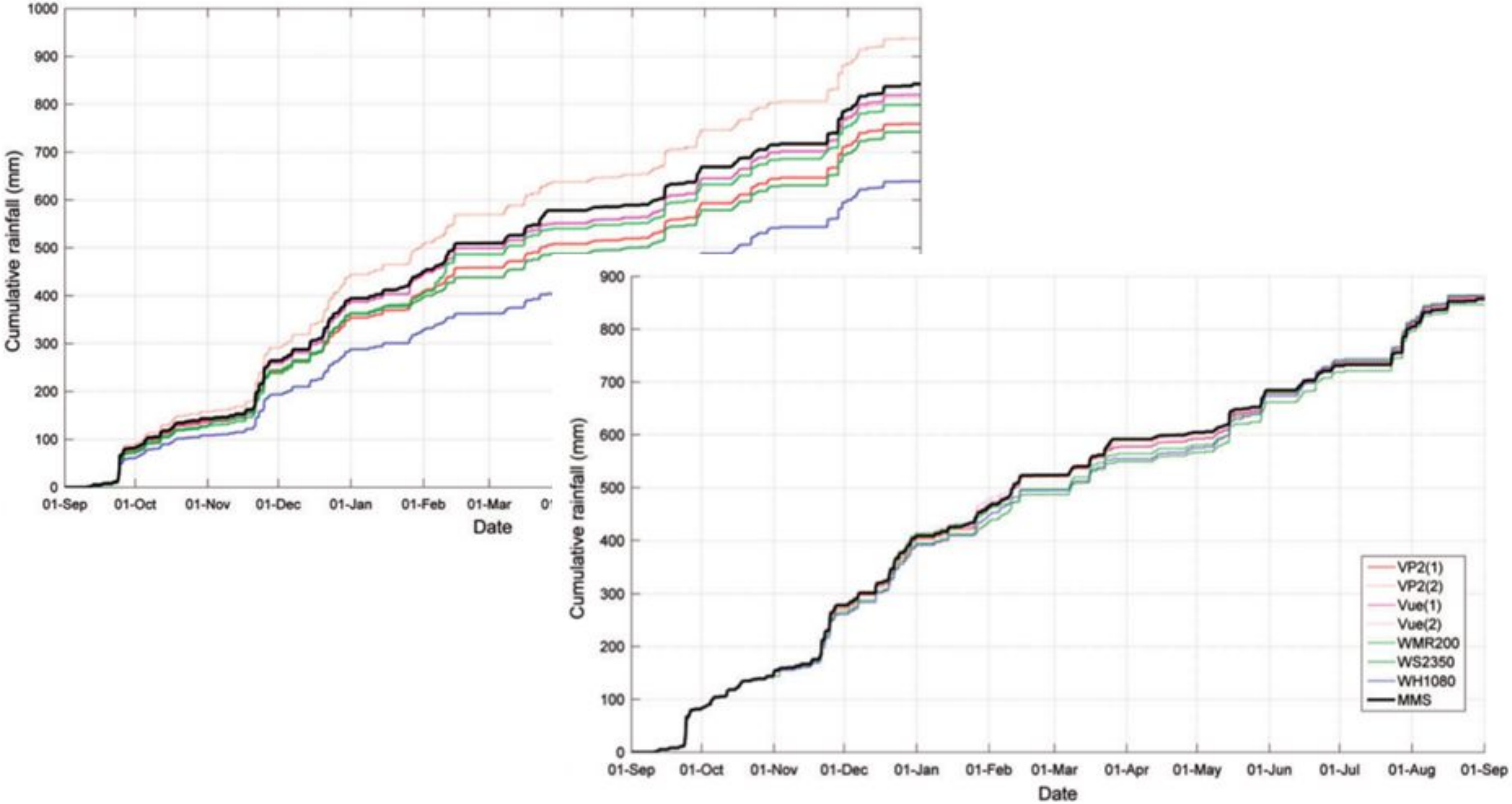


Figure 4. Time series plot of air temperature recorded by the seven CWS and the professional platinum resistance thermometer housed within a Stevenson screen for 26 May 2013. A time series of MMS global radiation is shown in orange.

Bell, S, Cornford, D & Bastin, L, 2015. Weather, 70 (3), pp. 75-84



Bell, S, Cornford, D & Bastin, L (2015)
How good are citizen weather stations? Addressing a biased opinion.
Weather, 70 (3), pp. 75-84.

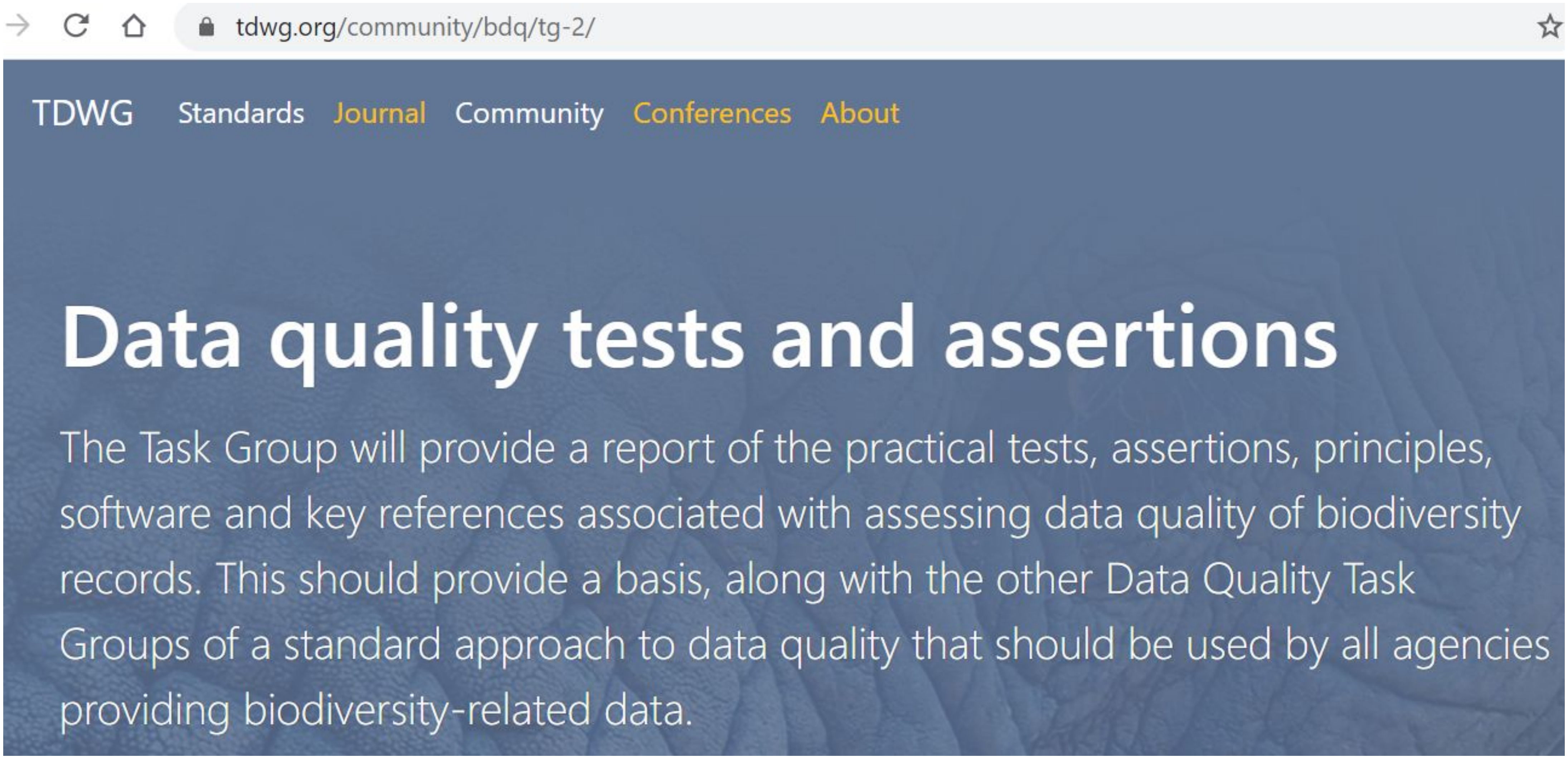
Observation-level quality

- more useful in a context where an individual outlier will have a large effect on a decision or modelling output

e.g in contexts where the decisions are high-stakes,

Allows filtering, where, to be fit for **your** purpose, all data points MUST conform to a certain standard.

An example from the Biodiversity Information Standards working group (TDWG)



For EACH observation, record whether tests are passed

{"name":"zeroCoordinates","code":4,"isFatal":true,"description":"Supplied coordinates are zero", "category":"warning","fatal":true},

{"name":"countryCoordinateMismatch","code":16,"isFatal":false,"description":"Coordinates dont match supplied country", "category":"error", "fatal":false},

{"name":"invertedCoordinates","code":3,"isFatal":false,"description":"Coordinates are transposed","category":"warning","fatal":false},

https://biocache.ala.org.au/ws/assertions/codes

Many of these errors are not specific to biodiversity data

- for example, typical errors like getting the x and y coordinates the wrong way round.

The definition is openly available - anyone can find out the meaning of a particular test failure, and decide whether that observation is acceptable for their own purpose.

- Like a shared **vocabulary**

"name":"invertedCoordinates",

"code":3,

"description":"Coordinates are transposed",

"fatal":false

Vocabularies, dictionaries, thesauri...

- There are many such contexts where it might be useful to share or even re-use a definition.

- Many scientific communities have collated terms for their domain so they can be unambiguously referenced.

- This often involves hosting the definition on the Web and referencing it via a **URI**

European Environment Agency

Login | >> Eionet portal

EIONET
Data Dictionary

You are here: **Eionet** » **Data Dictionary** » Vocabulary

- Help and documentation
- Datasets
- Tables
- Data elements
- Schemas
- **Vocabularies**
- Services
- Namespaces

Concept: *Particulate matter < 10 µm (aerosol)* in the *pollutant* vocabulary

Back to vocabulary

Concept URI	**http://dd.eionet.europa.eu/vocabulary/aq/pollutant/5**
Preferred label	Particulate matter < 10 µm (aerosol)
Definition	PM10 - recommended unit: µg/m3
Notation	PM10
Status	Valid
Status Modified	20.09.2013
Accepted Date	20.09.2013

https://dd.eionet.europa.eu/

Natural Environment Research Council

National Oceanography Centre
British Oceanographic Data Centre BODC

The NERC Vocabulary Server (NVS)

Service Status

Concept

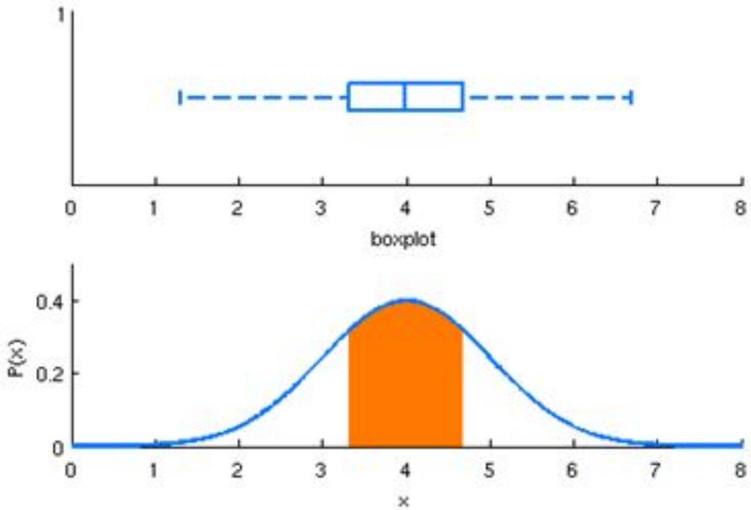
Not usable

URI	http://vocab.nerc.ac.uk/collection/L31/current/4/
Within Vocab	Geo-Seas data object quality flags
Preferred Label	Not usable
Definition	The data object (such as a seismic section) quality is so poor that it cannot be exploited
Note	accepted
Deprecated	false
Alternative Label	bad

Some vocabulary terms refer specifically to **quality conformance** and the methods used to measure it. For example, this URI takes you to a page with a clear definition of what the quality code means, and who it is used by.

UncertML describing and exchanging uncertainty...

| About UncertML | API documentation | User guide

URI:	http://www.uncertml.org/statistics/interquartile-range
UncertML name:	InterquartileRange
Alternative names:	IQR
Definition:	 The interquartile range is the range between the 1st and 3rd quartiles. It contains the middle 50% of the sample realisations (or of the sample

This vocabulary unambiguously defines statistical terms, so that users can be sure they are talking about the same clearly-defined measure or metric.

More at http://www.qualityml.org/

Citizens as reviewers?

Emerging tools allow a user to **annotate or tag** a dataset or an observation.

- can describe how and where they used the data.
- can flag up problems that they discovered.

Zabala et al (2021) *Geospatial User Feedback: How to Raise Users' Voices and Collectively Build Knowledge at the Same Time*. https://doi.org/10.3390/ijgi10030141

These examples have been rather biased towards spatial and biodiversity concepts.

…in your fields of expertise, you may be using other standards for documenting the data you create and publish.

If interoperability and metadata interest you… I would like to highlight two international initiatives working to bring together these standards, **specifically for citizen science.**

PPSR Core

A Data Standard for Public Participation in Scientific Research (Citizen Science)

Maintained by the Data and Metadata Working Group of the Citizen Science Association
https://core.citizenscience.org/

PPSR Core is a set of global, transdisciplinary data and metadata standards for use in **P**ublic **P**articipation in **S**cientific **R**esearch **(Citizen Science)** projects. These standards are united, supported, and underlined by a common framework illustrating how information is structured within the citizen science domain. This allows data to be used across platforms and projects in a consistent manner, furthering the research goals of the scientific community.

PPSR-Core - not about creating a whole new standard for the sake of it.

Aims to unify EXISTING standards and ontologies and re-use or map to definitions which already exist.

- **Dublin Core Metadata Element Set**
- **ISO 19115 Geographic information — Metadata**
- **Friend of a Friend Ontology**
- **Darwin Core Terms**
- **Data Catalog Vocabulary**
- **OGC Observations and Measurements**
- **Core Ontology of Scientific Investigation**
- **PROVenance Interchange Ontology**
- **OGC Earth Observations GeoJSON**

dataQualityAssuranceMethod	-Data owner curated -Subject matter expert record verification -Crowd-sourced record verification -Record annotation -System supported data attribute configuration -No DQ methods used -Not applicable

Elements of the PPSR-CORE remit:

-decide what information is essential

-construct vocabularies that reflect actual practice across citizen science

https://core.citizenscience.org/

The OGC* Citizen Science Interoperability Experiment

*Open Geospatial Consortium

https://external.ogc.org/twiki_public/CitSciIE/WebHome

Ongoing initiative to demonstrate how current ICT-based tools can be applied to allow easier citizen participation and better data reuse.
2019 Engineering report at http://docs.opengeospatial.org/per/19-083.html

Some outputs specifically address quality:
e.g. https://doi.org/10.1117/12.2570814

Assess citizen science based land cover maps with remote sensing products: the Ground Truth 2.0 data quality tool

Summary

There is **often huge suspicion about citizen science data quality**

It can be an excellent complement to research datasets; sometimes of equivalent or better quality.

Often, **it contains rich information, additional to what scientists want**:

- e.g., *where are people observing, and which people*: tells you something about digital inclusion and how different social groups experience their local surroundings.

We have to be transparent about the quality aspects of all data, so that a user can decide if it is fit for their purpose.

Huge momentum right now - potential for a really open Citizen Science data ecosystem that crosses disciplinary boundaries.

Some references and further links

Website of the PPSR-CORE initiative https://core.citizenscience.org/

Engineering Report of the OGC Citizen Science Interoperability experiment http://docs.opengeospatial.org/per/19-083.html#DataQuality

Yu et al. (2015) Towards Linked Data Conventions for Delivery of Environmental Data Using netCDF. https://hal.inria.fr/hal-01328530/document

A collection of resources related to dataset quality and FAIR principles. https://wiki.esipfed.org/FAIR_Dataset_Quality_Information

Thanks for watching and listening

✉ peter.mooney@mu.ie
✉ l.bastin@aston.ac.uk