Data quality considerations in Citizen Science

Peter Mooney

Maynooth University
National University of Ireland Maynooth

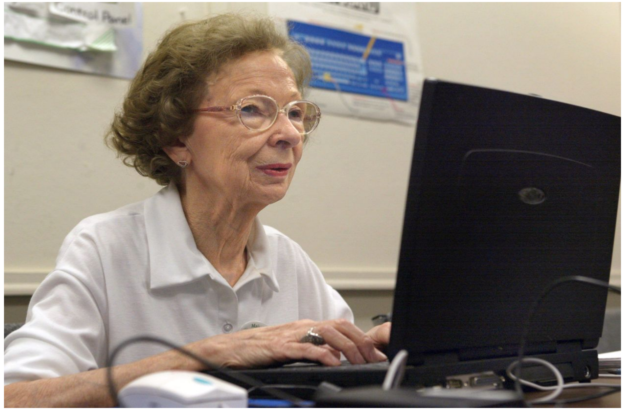
Lucy Bastin

Aston University
BIRMINGHAM UK

ESIP Information Quality Cluster (IQC)

Tuesday, June 22nd 2021

The societal impact of citizen science? There is no ‘template’ citizen scientist. Everyone has a role to play



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

Fitness for use?

Fitness for purpose?

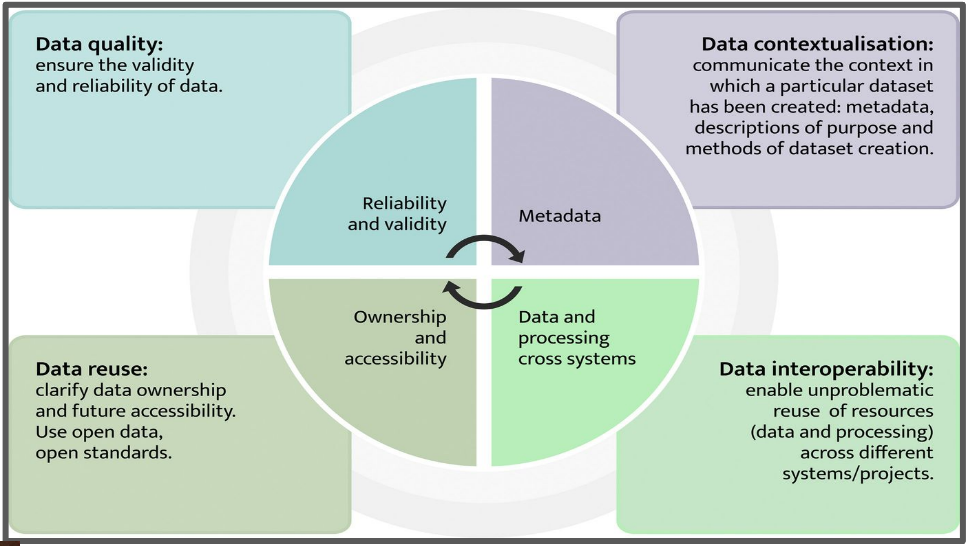
Who baked the cake?

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



Data quality: ensure the validity and reliability of data.

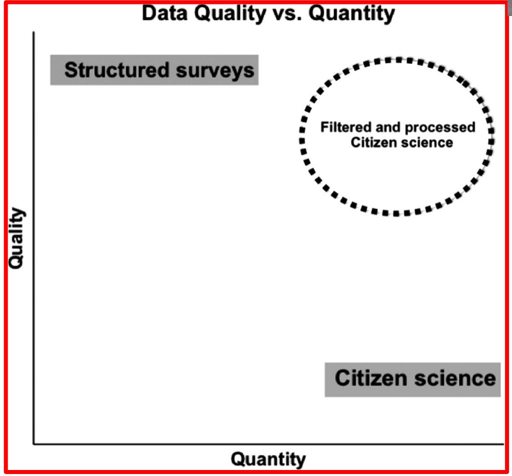
Data contextualisation: communicate the context in which a particular dataset has been created: metadata, descriptions of purpose and methods of dataset creation.

Data reuse: clarify data ownership and future accessibility. Use open data, open standards.

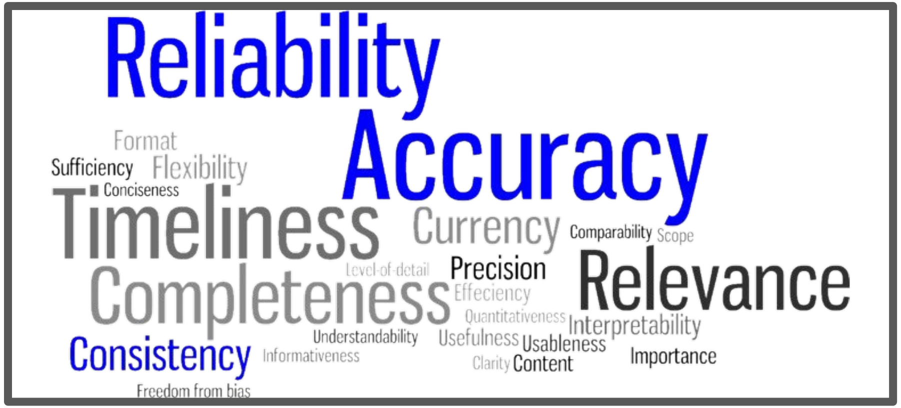
Data interoperability: enable unproblematic reuse of resources (data and processing) across different systems/projects.

Reliability and validity
Metadata
Ownership and accessibility
Data and processing cross systems

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

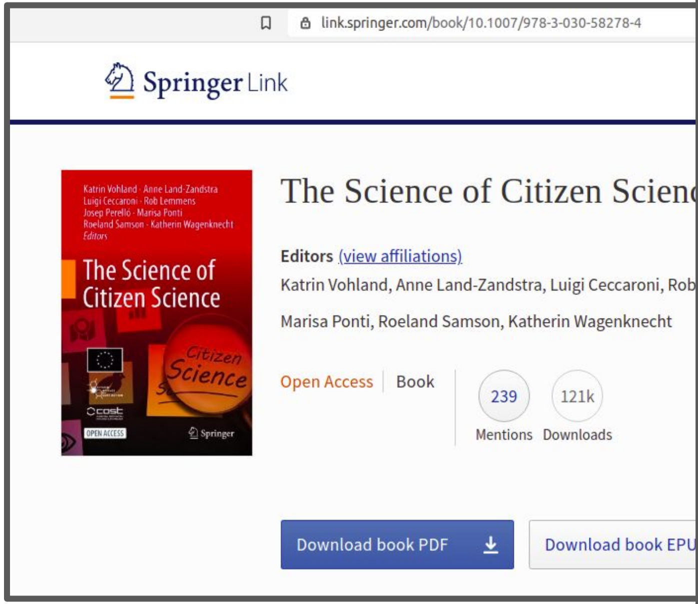


https://doi.org/10.1111/ddi.13068



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

So how did Lucy and I arrive here?



Katrin Vohland · Anne Land-Zandstra
Luigi Ceccaroni · Rob Lemmens
Josep Perelló · Marisa Ponti
Roeland Samson · Katherin Wagenknecht
Editors

The Science of Citizen Science

The Science of Citizen Scienc

Editors (view affiliations)

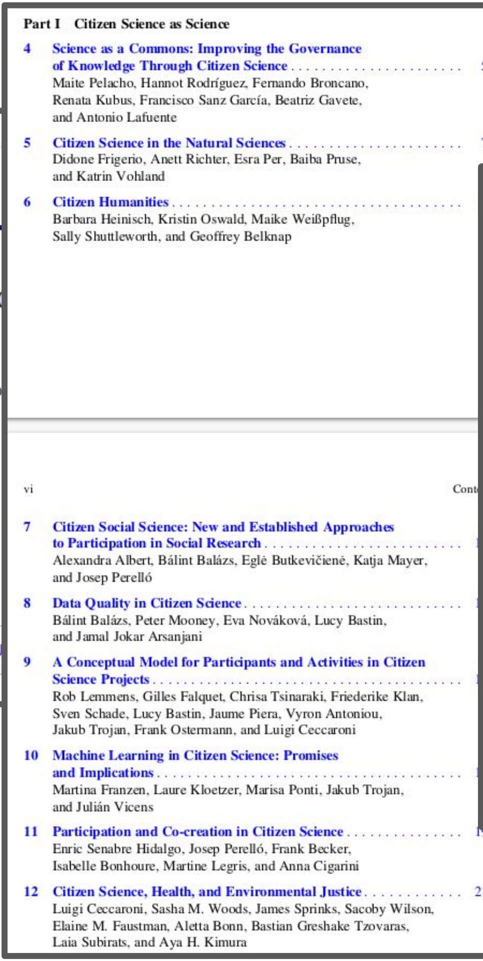
Katrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni, Rob

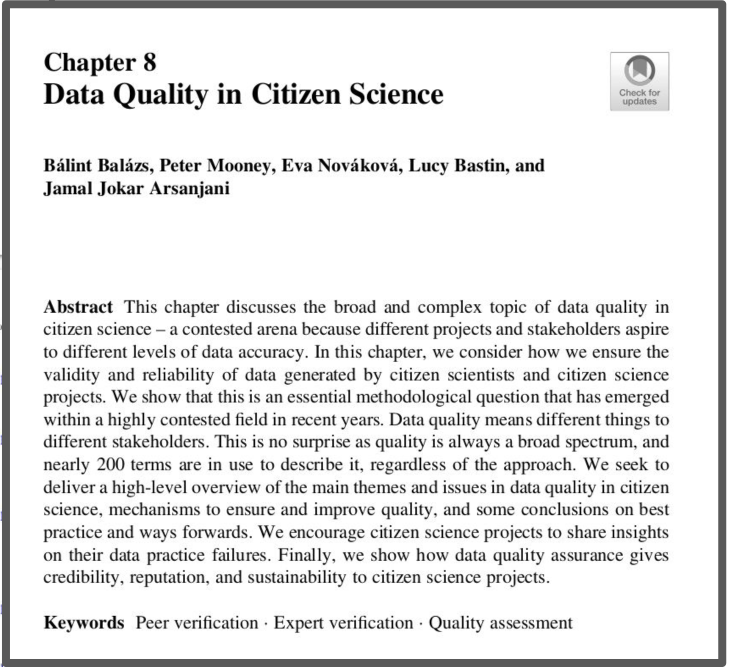
Marisa Ponti, Roeland Samson, Katherin Wagenknecht

Open Access | Book

239 Mentions 121k Downloads

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





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

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

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

> “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***” Lukyanenko et al (2016) https://doi.org/10.1111/cobi.12706

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

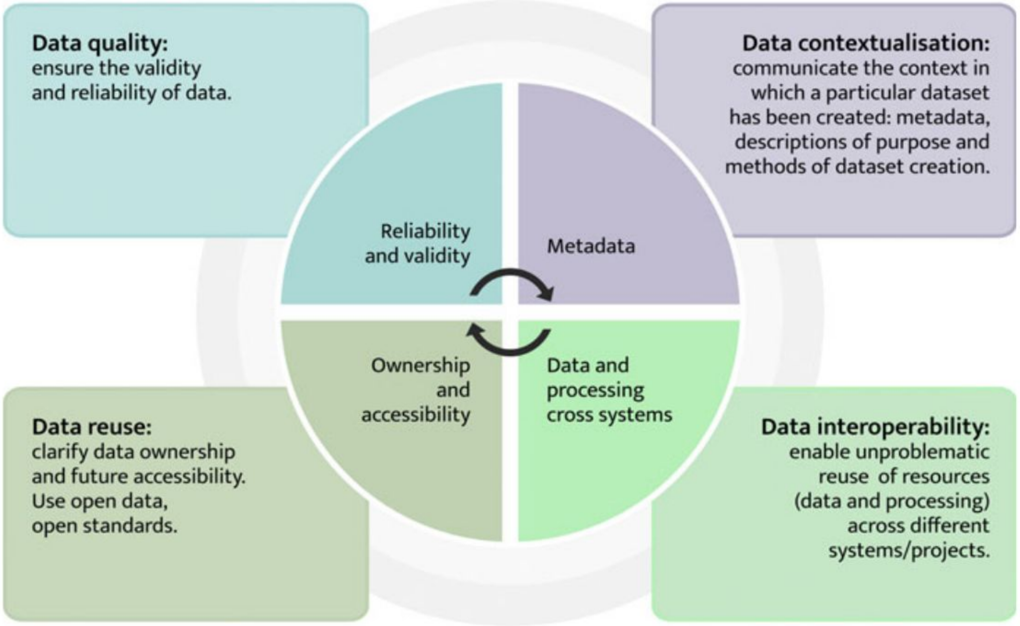
“**Most types of bias found in citizen-science datasets are also found in professionally produced datasets** and can be mitigated using existing statistical tools” (Kosmala et al, 2016) doi: 10.1002/fee.1436

“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 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.



Data quality: ensure the validity and reliability of data.

Data contextualisation: communicate the context in which a particular dataset has been created: metadata, descriptions of purpose and methods of dataset creation.

Data reuse: clarify data ownership and future accessibility. Use open data, open standards.

Data interoperability: enable unproblematic reuse of resources (data and processing) across different systems/projects.

Reliability and validity
Metadata
Ownership and accessibility
Data and processing cross systems

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

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

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

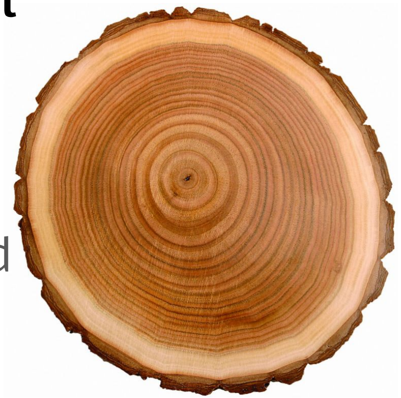
1. Data collection **protocols are not followed by participants**.

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

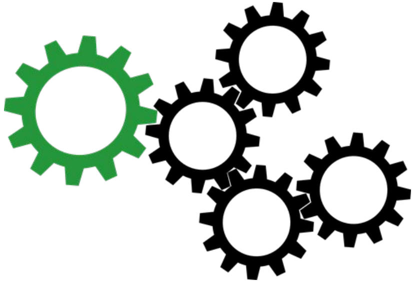
3. Data collection **protocols are incorrectly implemented**.

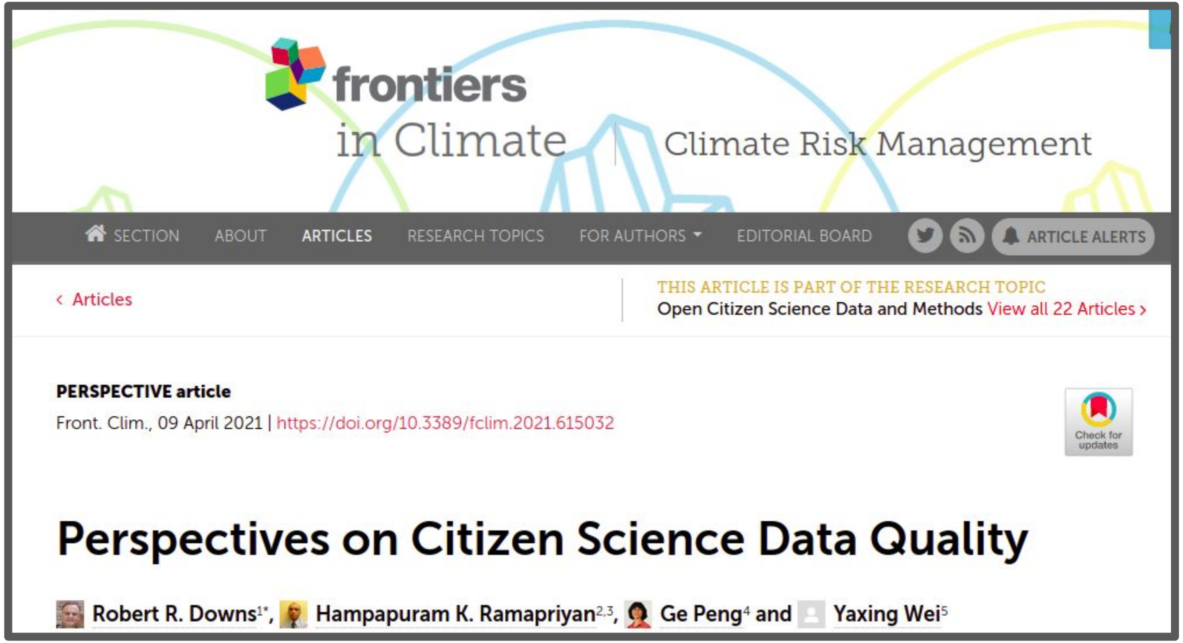
4. Data collection **protocols are not comprehensive** and are used by stakeholders with **different data quality expectation levels**.

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



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





THIS ARTICLE IS PART OF THE RESEARCH TOPIC
Open Citizen Science Data and Methods View all 22 Articles ›

PERSPECTIVE article

Front. Clim., 09 April 2021 | https://doi.org/10.3389/fclim.2021.615032

Perspectives on Citizen Science Data Quality

Robert R. Downs[1*], Hampapuram K. Ramapriyan[2,3], Ge Peng[4] and Yaxing Wei[5]

https://doi.org/10.3389/fclim.2021.615032

“.....**documenting CSD (Citizen Science Data) quality can improve trust in CS** within the scientific community and reflects ethical approaches to conducting CS. …. **Investigators should describe data quality in the metadata and data documentation**, as well as in data papers and publications. Documentation should differentiate between various quality issues to **avoid confusing potential users.**”

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…

- 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. But we need to know:

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

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

Fitness-for-purpose in citizen science

Producers or managers of ‘authoritative’ datasets have a relatively standardised set of QA tools and procedures to document

(Even so, the documentation can be highly variable!)

Potential users can evaluate the quality of that data against their needs.

(These users are becoming more numerous and variable)

With citizen science, the **communication challenge is multiplied**:

- The ways of producing data proliferate and become more variable
- So do the strategies for assuring data quality
- So do the ways in which a producer values / describes quality

Quality evaluation in citizen science

Some useful elements for assessing fitness-for-purpose:

- **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.

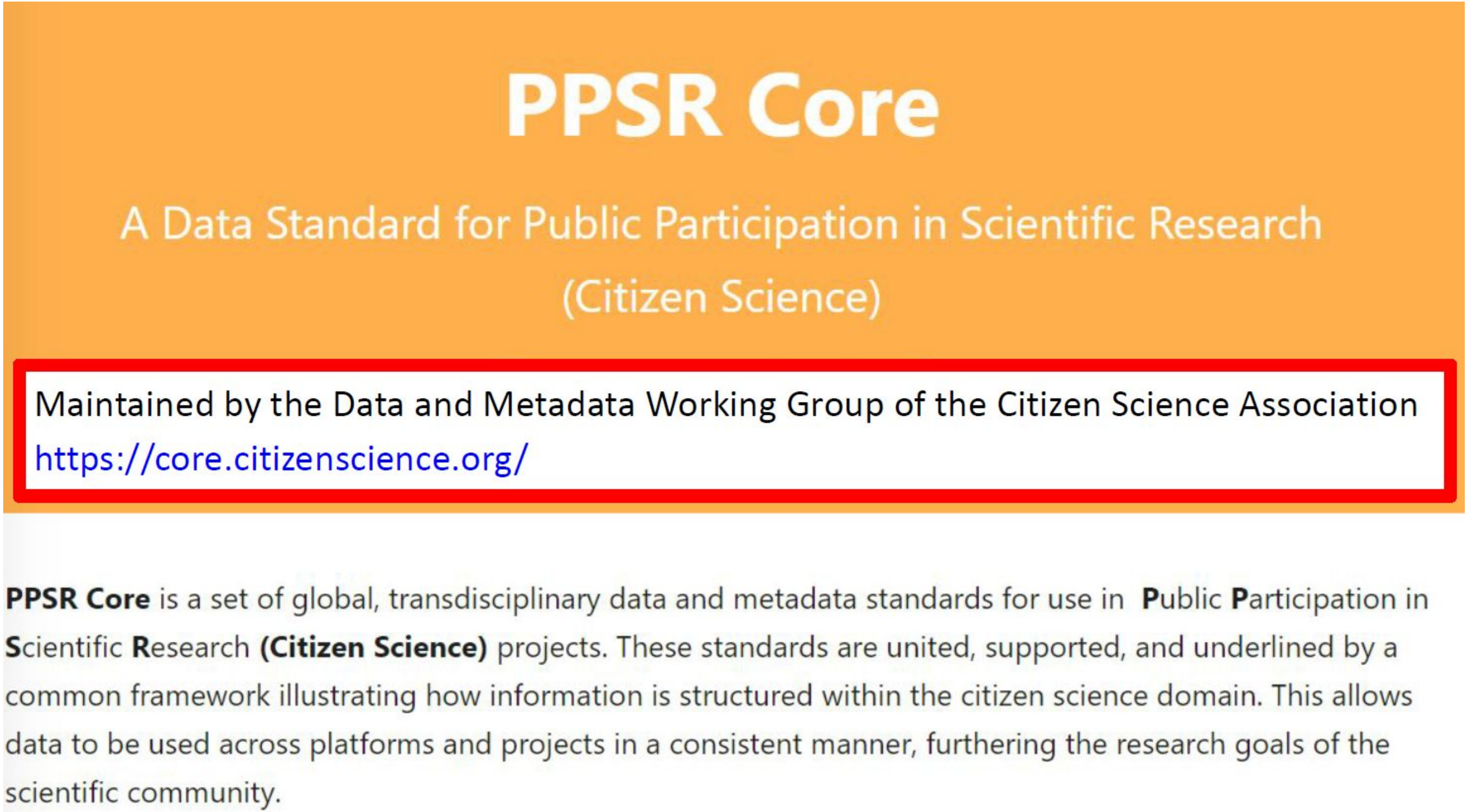
Metadata for citizen science

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.



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.

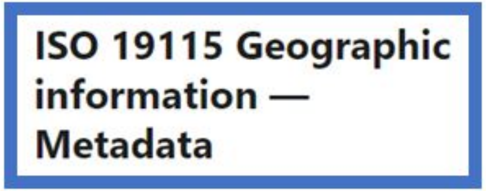
Aims to unify EXISTING standards and ontologies and re-use or map to definitions which already exist.

Dublin Core Metadata Element Set

Darwin Core Terms

Core Ontology of Scientific Investigation

PROVenance Interchange Ontology



ISO 19115 Geographic information — Metadata

OGC Observations and Measurements

OGC Earth Observations GeoJSON

Friend of a Friend Ontology

Data Catalog Vocabulary

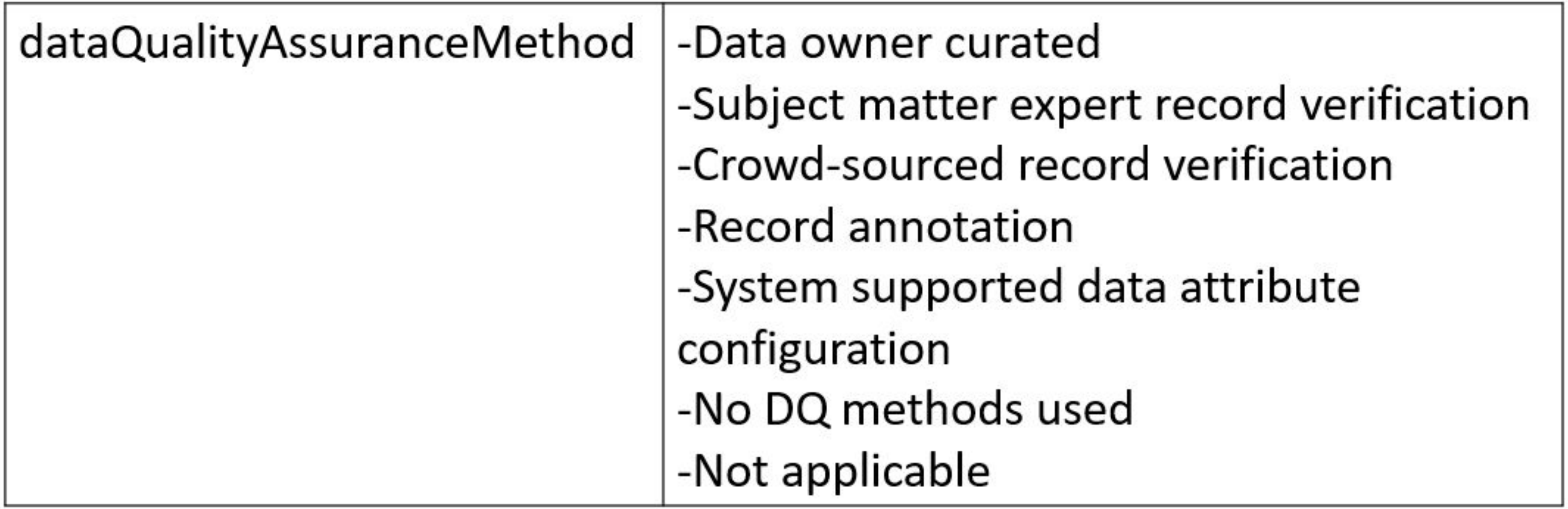
PPSR Core quality component is pretty minimal

- The expected usage is through extended profiles, which as far as possible use existing standards and information models
- As ever, this gives opportunity for duplication / redundancy
- Active engagement with initiatives like the 19157 Data Quality Measures Register* will be crucial

* Described by Ivana Ivanova in last month's meeting

dataQualityAssuranceMethod	Description of the types of data quality assurance methods that were applied in capturing, curating and managing the dataset.	Vocabulary
dataQualityAssuranceDescription	Detailed description of the methods used to quality assure the dataset both during capture and post processing. This is important for data users to understand the processes applied to the data to verify or enhance its quality for use.	Text
spatialAccuracy, temporalAccuracy, nonTaxonomicAccuracy	A generalised category that best reflects the least accurate record in the dataset.	Vocabulary (e.g., Low, Medium, High)
speciesIdentificationAccuracy	A generalised category that best reflects the least accurate record in the dataset for species identification. Choose 'Not applicable' if species fields are not included in the dataset.	Vocabulary
methodSpecification	**Details of the methodology or sampling protocol used to collect the dataset.**	**cosi:hasRelatedMaterial**

cosi = Core Ontology of Scientific Investigation



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 proposed labels for citizen science to describe how data QA was carried out.

Work in progress

https://core.citizenscience.org/

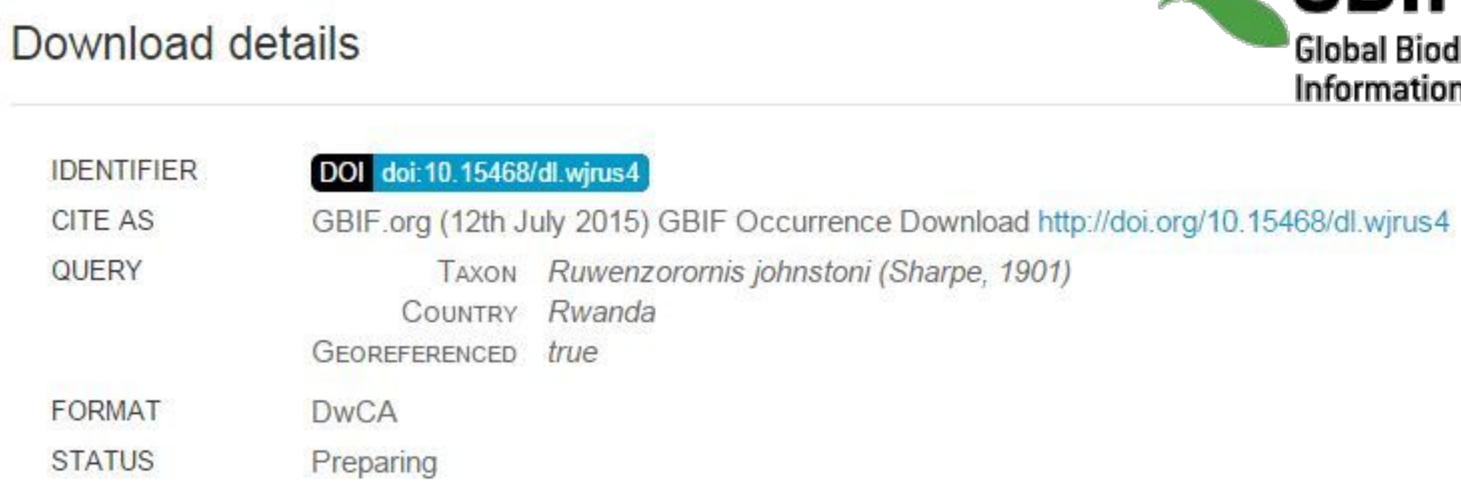
PPSR Core

Are dataset-level quality metrics sufficient?

Many citizen science repositories are not static ‘datasets’

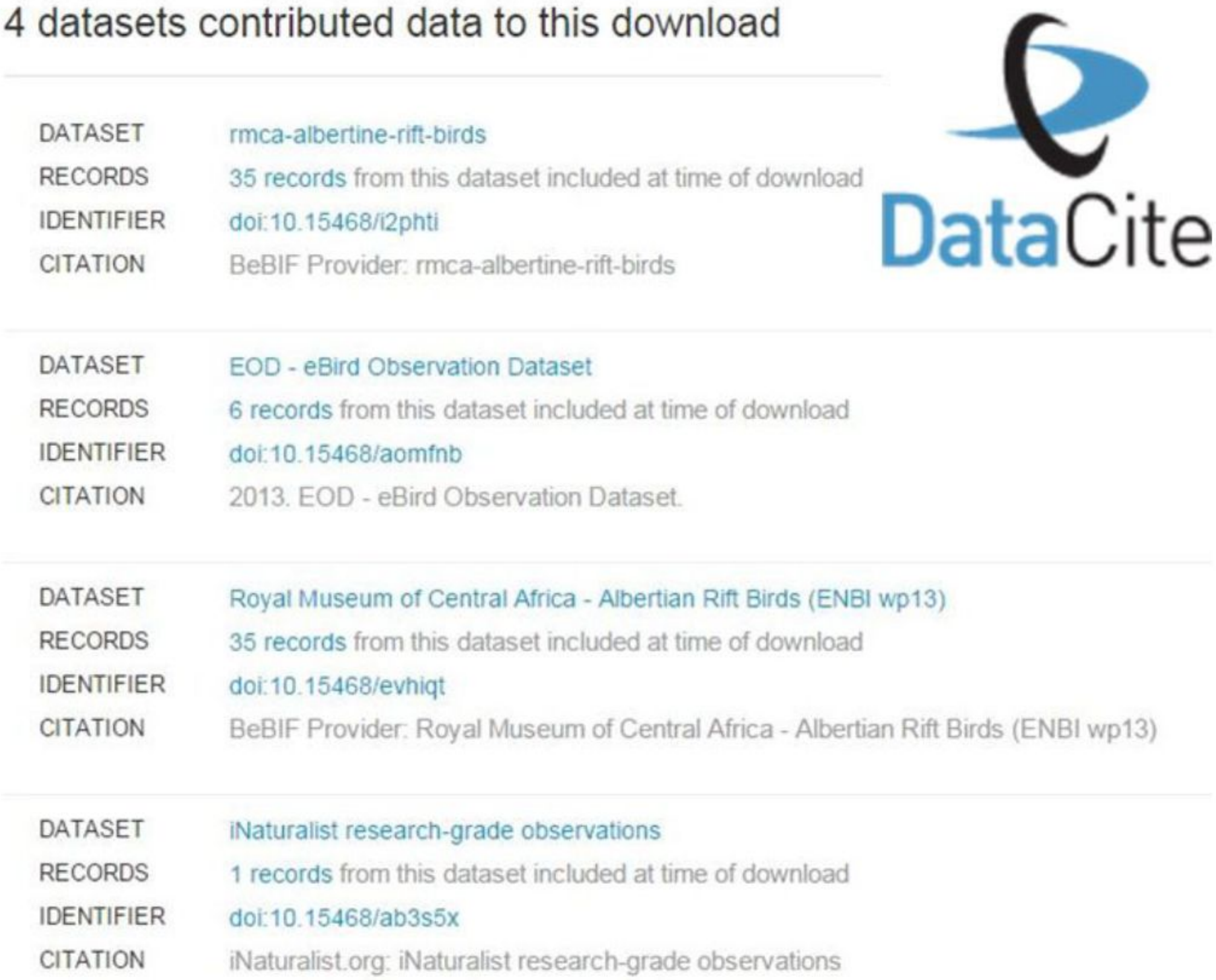
They can be ‘sliced and diced’ and queried in a range of ways.

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

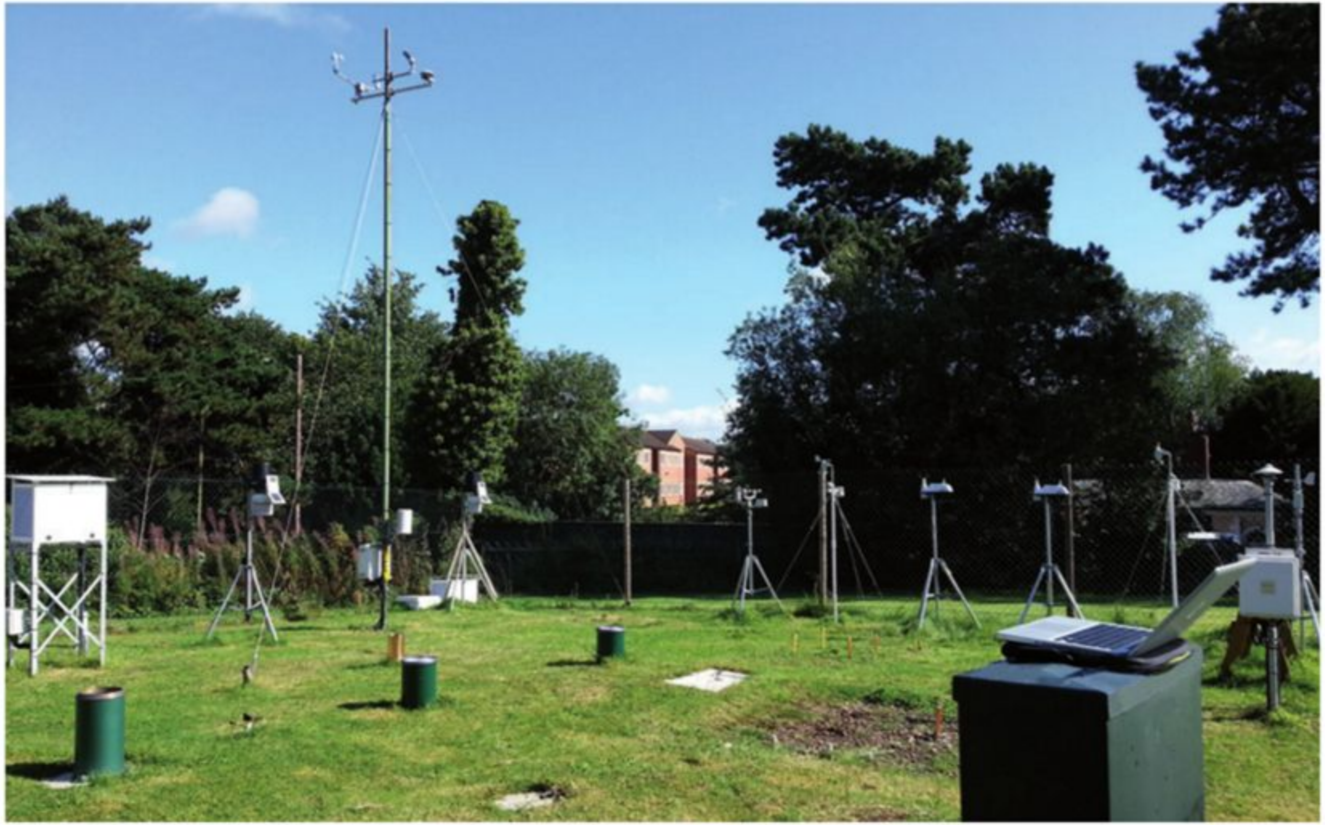
DataCite

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
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

Observation-level metadata

- more useful in a context where an individual outlier will have a large effect on a decision or modelling output
-
- Or where you EXPECT data points to have varying reliability

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



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

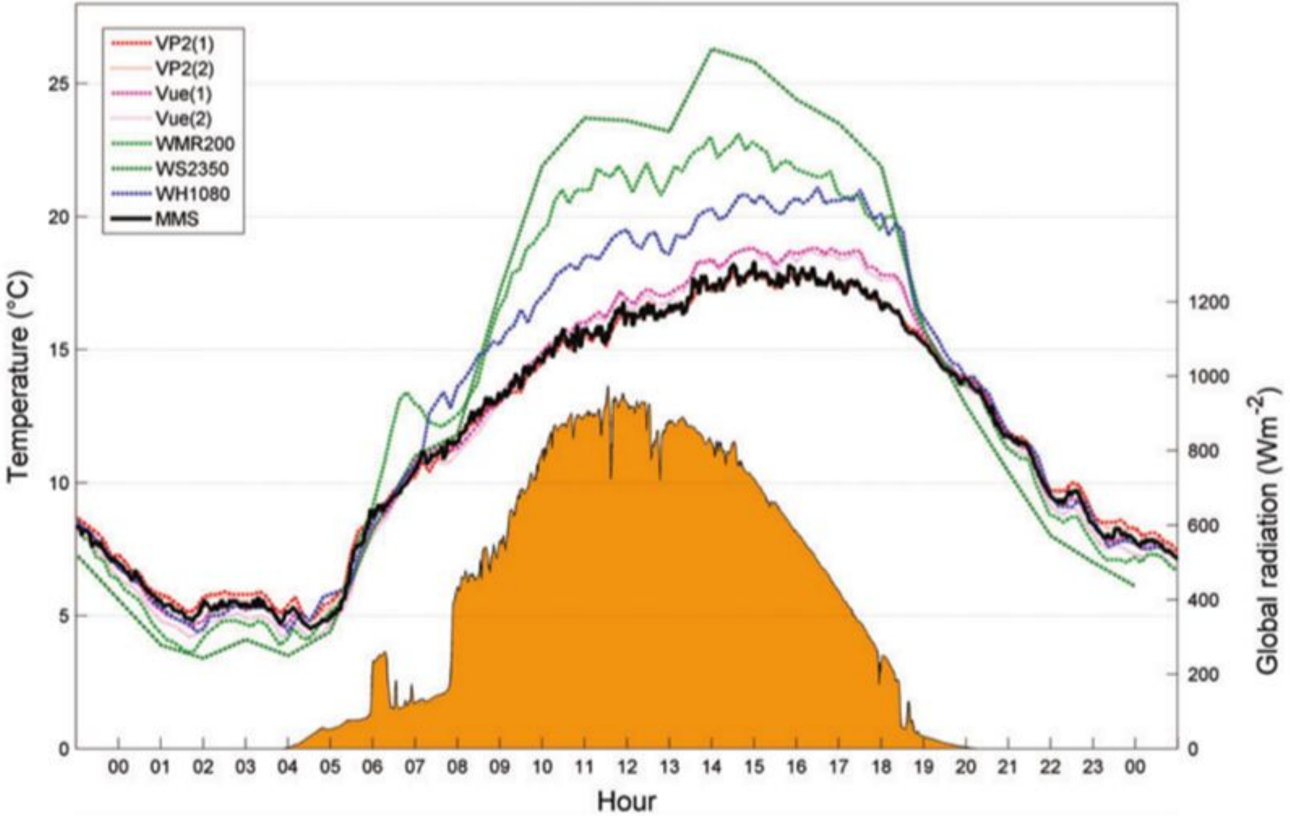
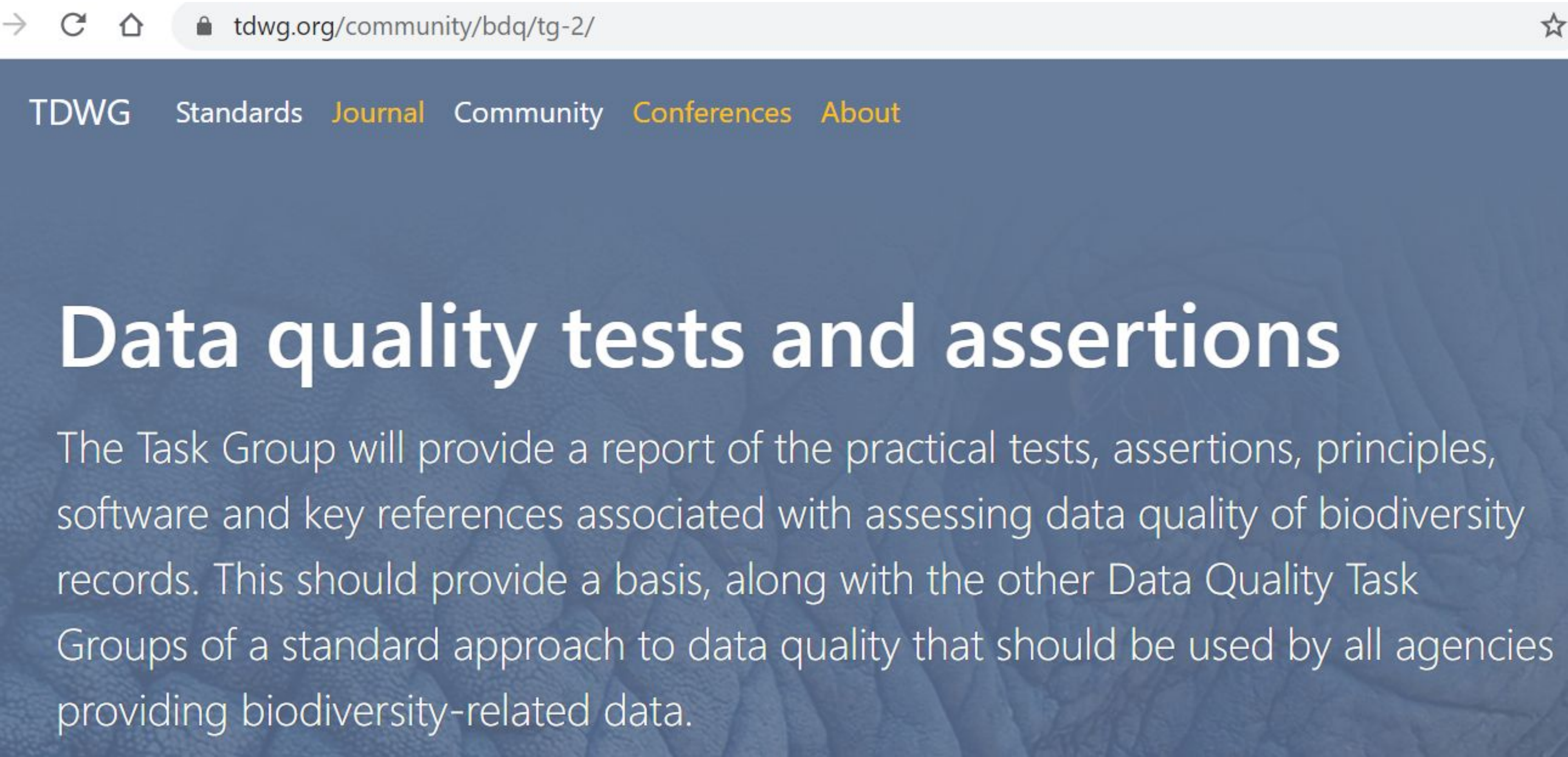


Figure 4. Time series plot of air temperature recorded by the seven CWS and the professional platinum resistance thermometer housed within a Stevenson screen for 26 May 2013. A time series of MMS global radiation is shown in orange.

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

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



Data quality tests and assertions

The Task Group will provide a report of the practical tests, assertions, principles, software and key references associated with assessing data quality of biodiversity records. This should provide a basis, along with the other Data Quality Task Groups of a standard approach to data quality that should be used by all agencies providing biodiversity-related data.

For EACH observation, record whether tests are passed

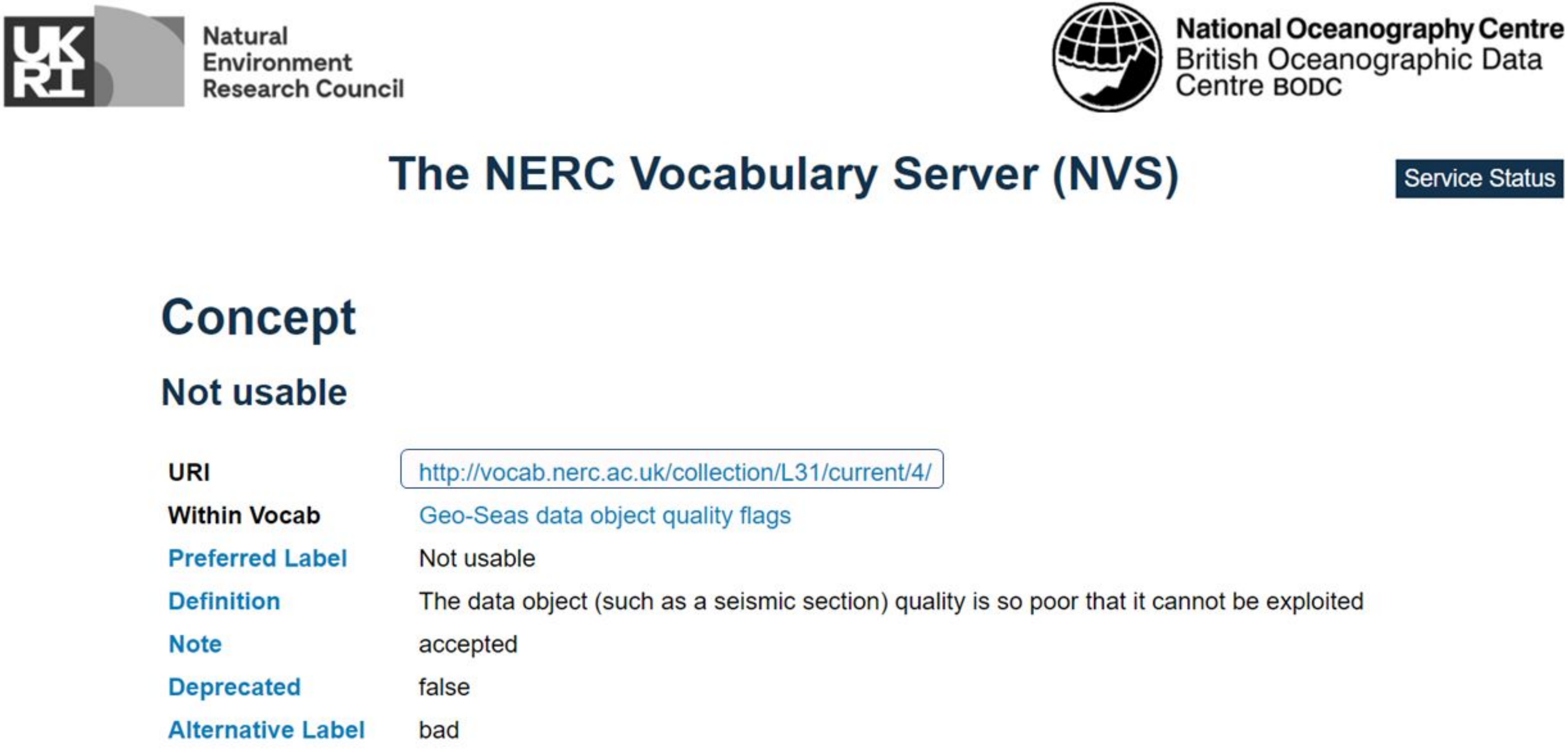
{"name":"zeroCoordinates","code":4,"isFatal":true,"description":"Supplied coordinates are zero", "category":"warning","fatal":true},

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

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

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**


Natural Environment Research Council
National Oceanography Centre
British Oceanographic Data Centre BODC
Service Status

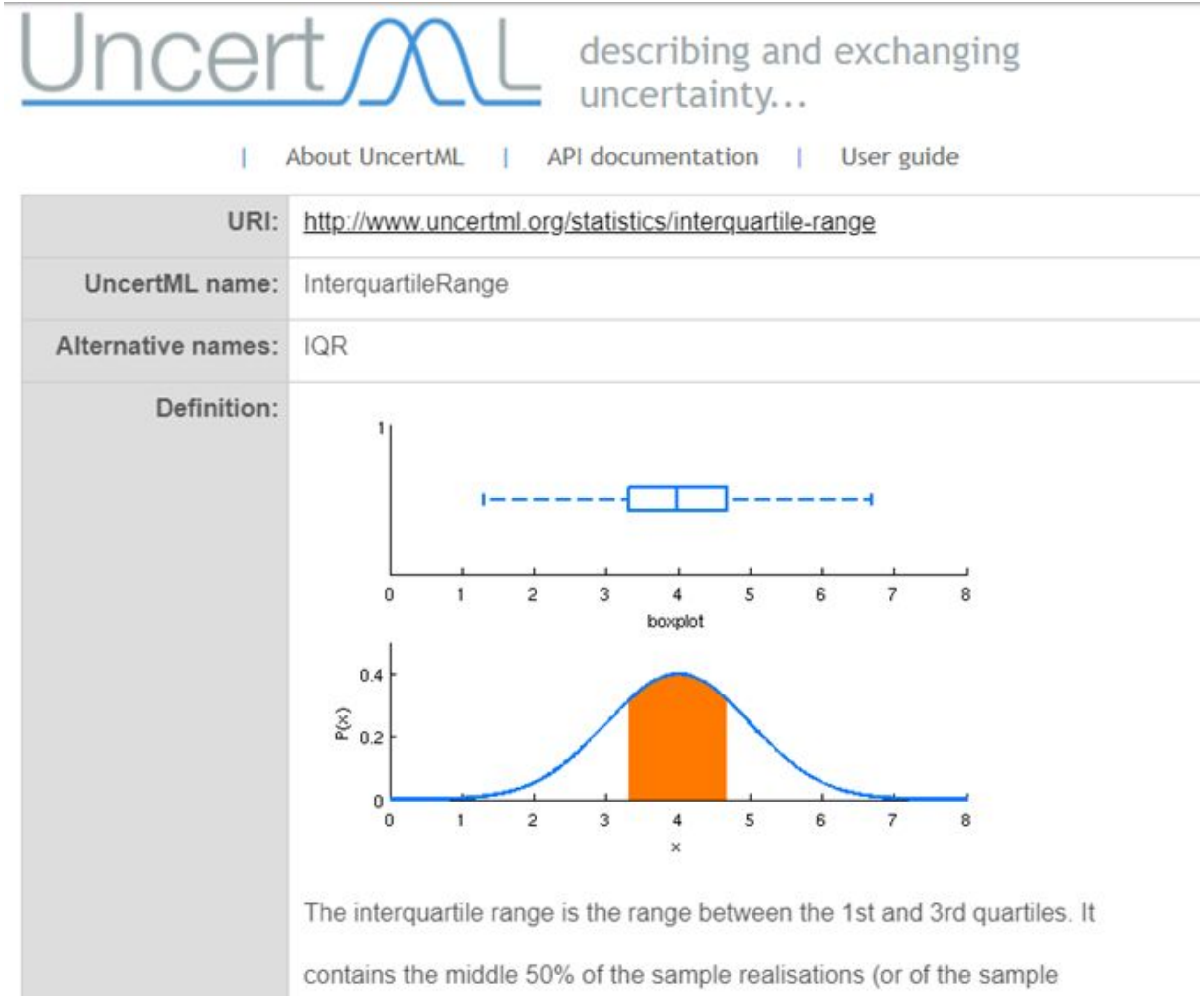
The NERC Vocabulary Server (NVS)

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.

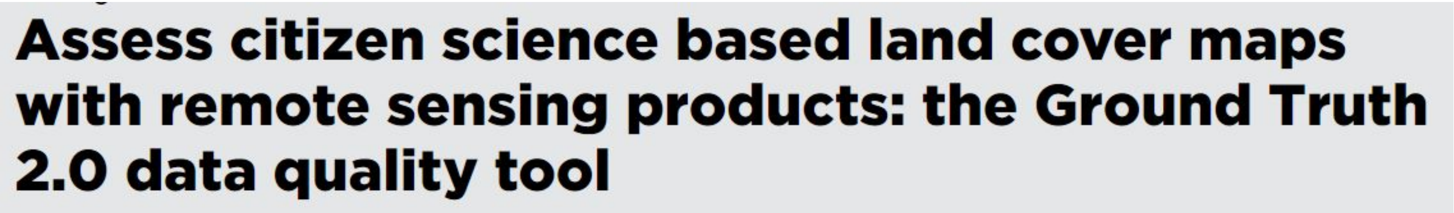
More at http://www.qualityml.org/

The OGC Citizen Science Interoperability Experiment

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: Huge momentum right now - potential for a truly open Citizen Science multidisciplinary data ecosystem. We need to overcome CS skepticism

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

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.

Crucial role of metadata: If metadata communicates provenance and quality, we can decide whether it's scientifically appropriate to re-use Citizen Science datasets. Example: PPSR Core efforts. Unify existing standards rather than re-inventing the wheel

Some useful references on Citizen Science Data Quality

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

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

Questions?

Comments?

✉ peter.mooney@mu.ie

✉ l.bastin@aston.ac.uk