Doing Better Citizen Science From data quality to project design

Bálint Balázs

CESIS, June 4th, 2018

COST WG1 - Summary of work

- **Dec 2016:** Berlin / general topics of Data Quality. Many voices, many opinions, many directions, etc.
- **April 2017:** Call for Contributions published for first WG1 workshop
- **June 2017:** Call for Contributions ends. WG1 participants chosen.
- **Sept 2017:** Workshop in Budapest. 2 Days. 14 Participants - “getting to work”
- **June 2018:** Workshop in Geneva, Summary
- **2019**: Finalising workshop planned

COST WG1

Systemetic review

DQ approaches through

4 selected *Story - Actor Scenarios:*

Environmental Monitoring

GIS/VGI/Mapping

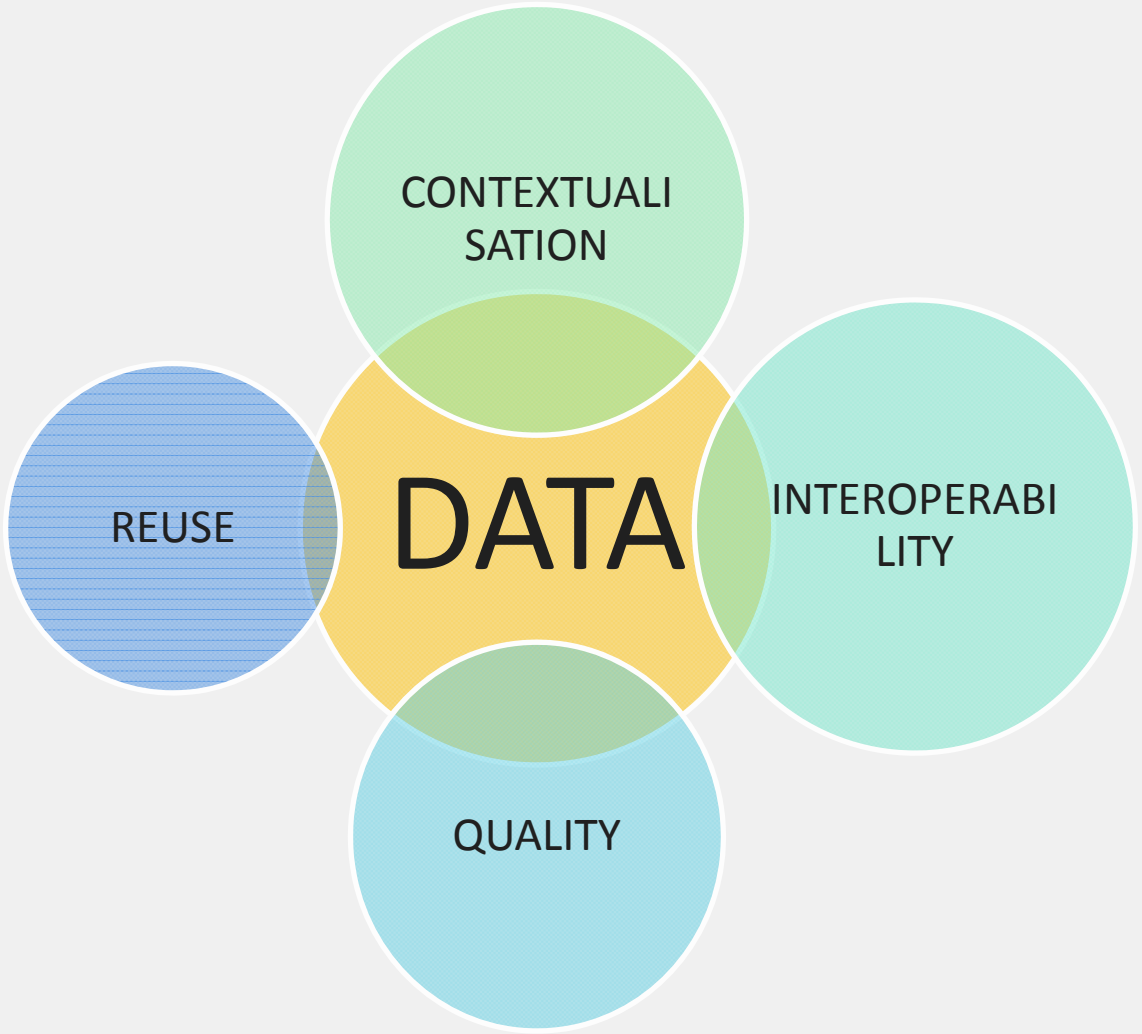
Natural history/BioDiv Observation

Harmful Species Monitoring

goo.gl/EwoA6J

Data is key NEED TO

- present dataset creation: purpose and methods.
- reuse resources across systems/projects.
- ensure the validity and reliability.
- clarify ownership and accessibility.



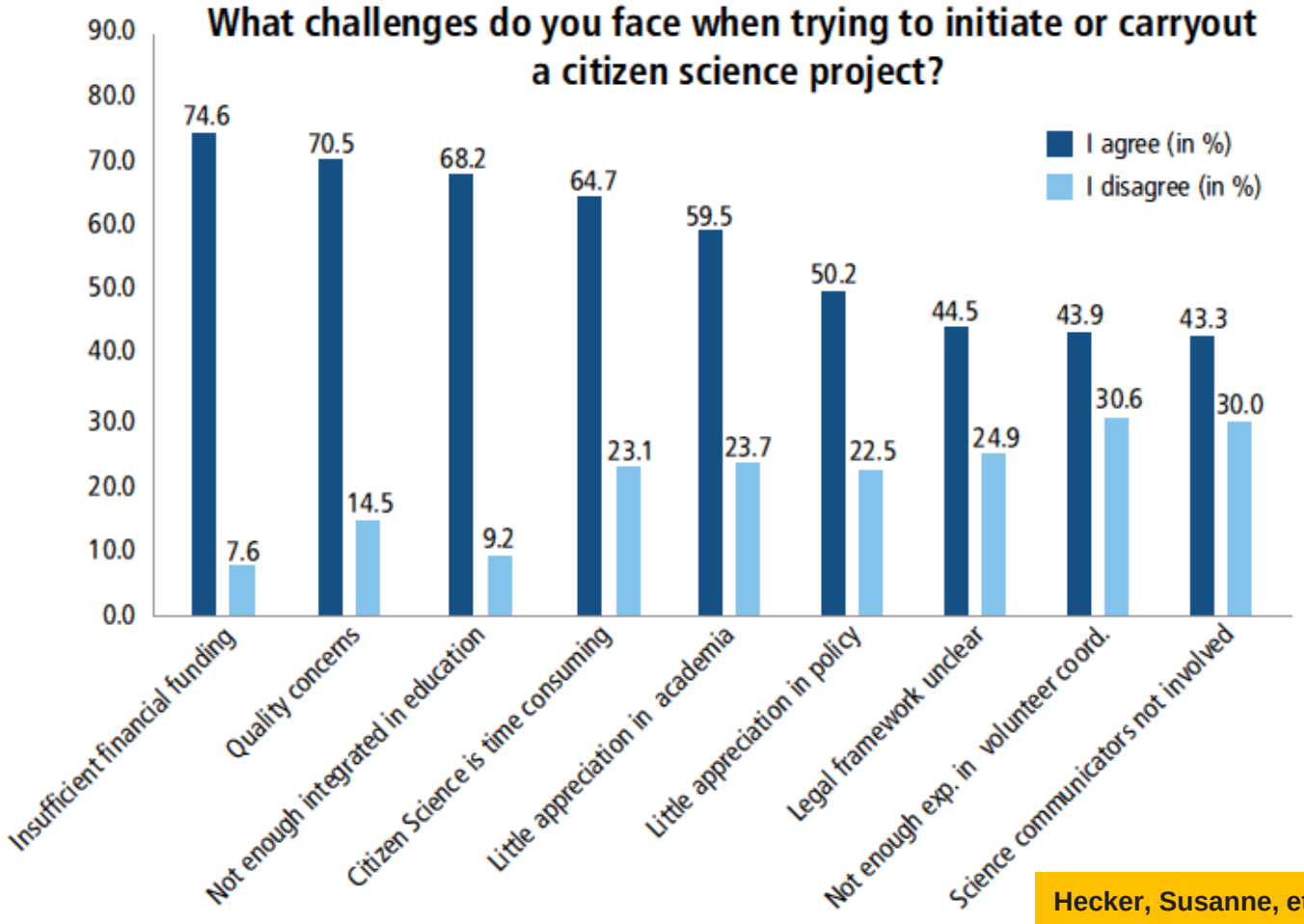
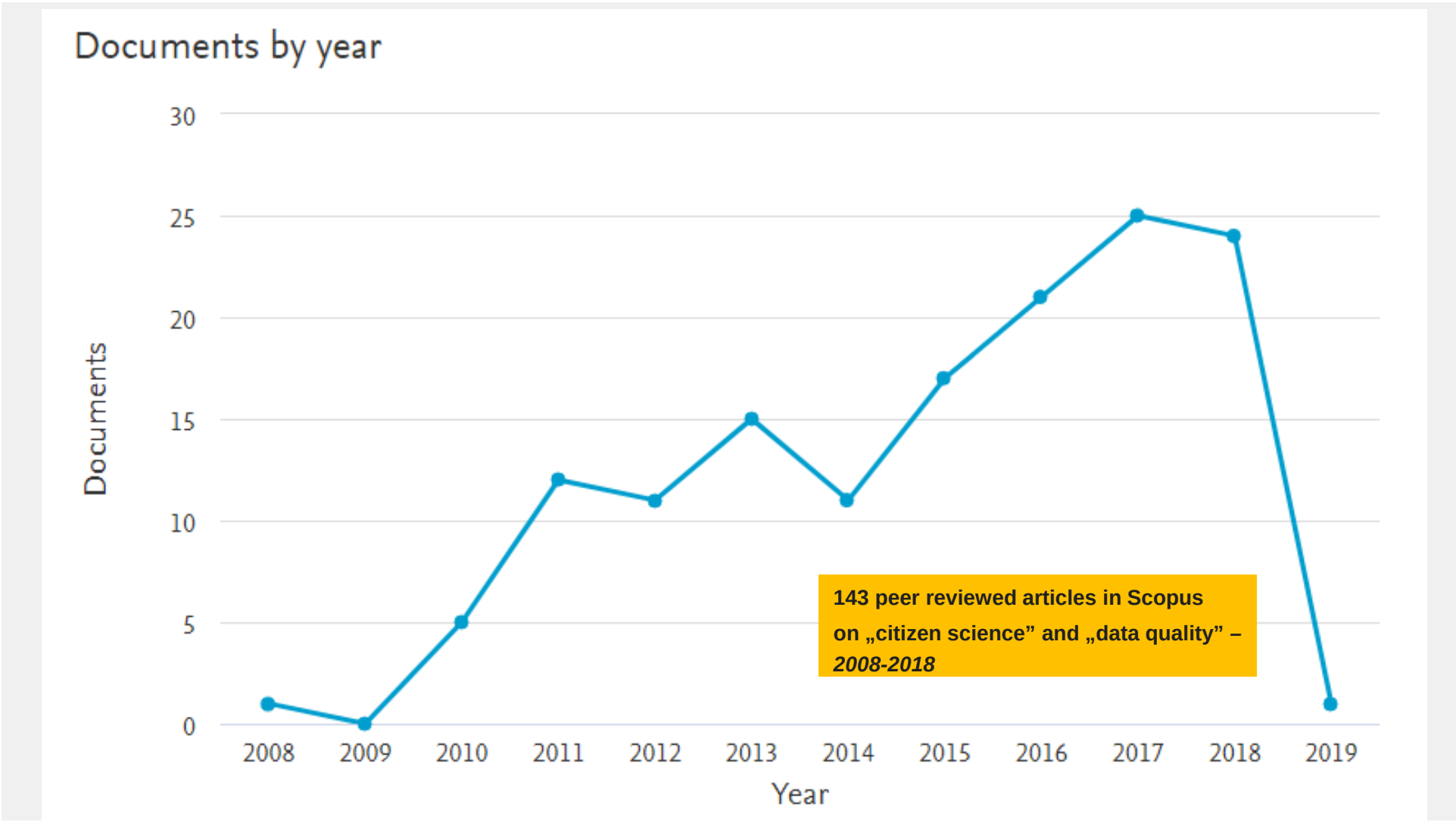


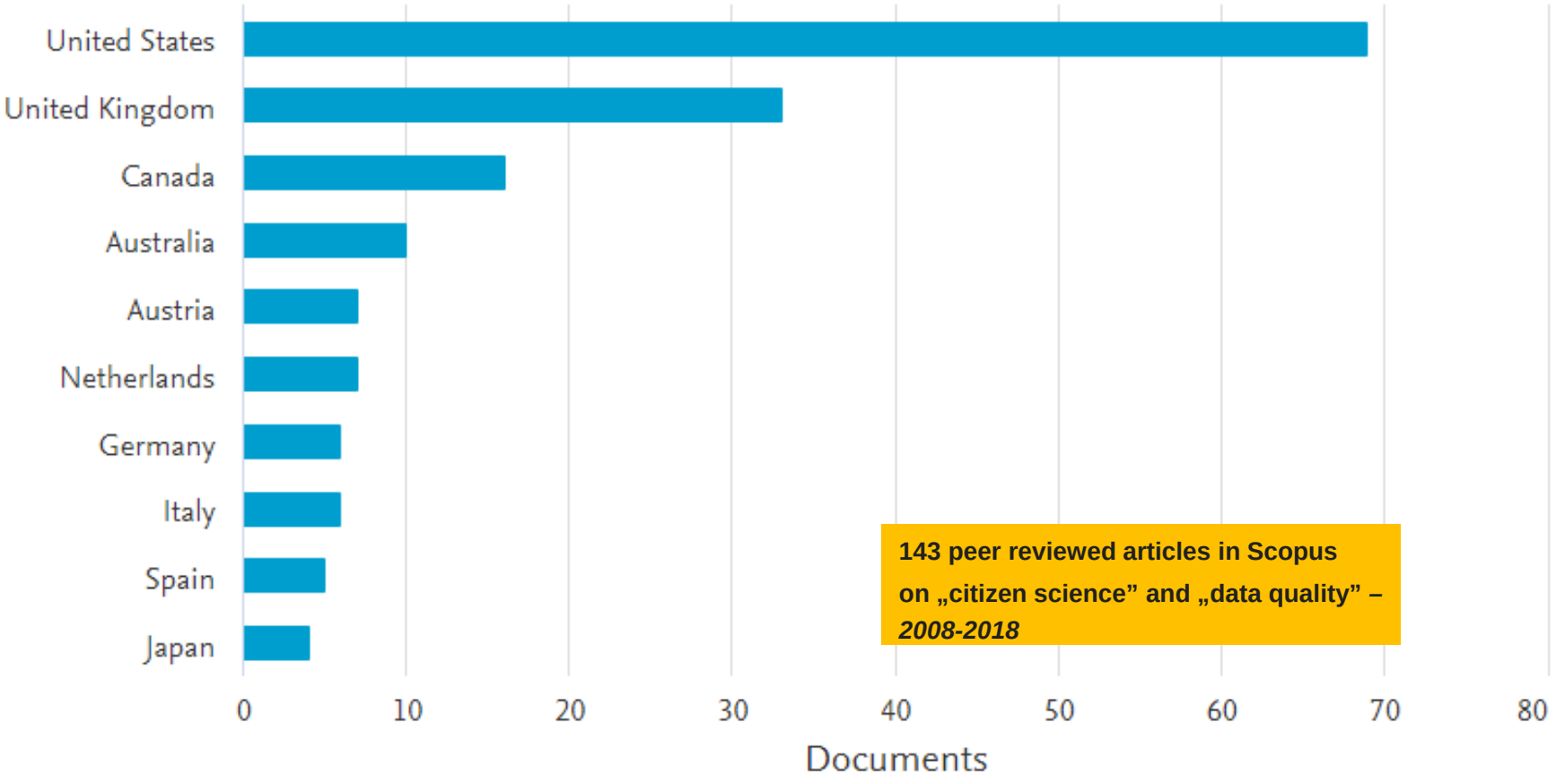
Fig. 13.7 Challenges for citizen science projects

Hecker, Susanne, et al., eds. *Citizen Science: Innovation Open Scien*. UCL Press, 2018.

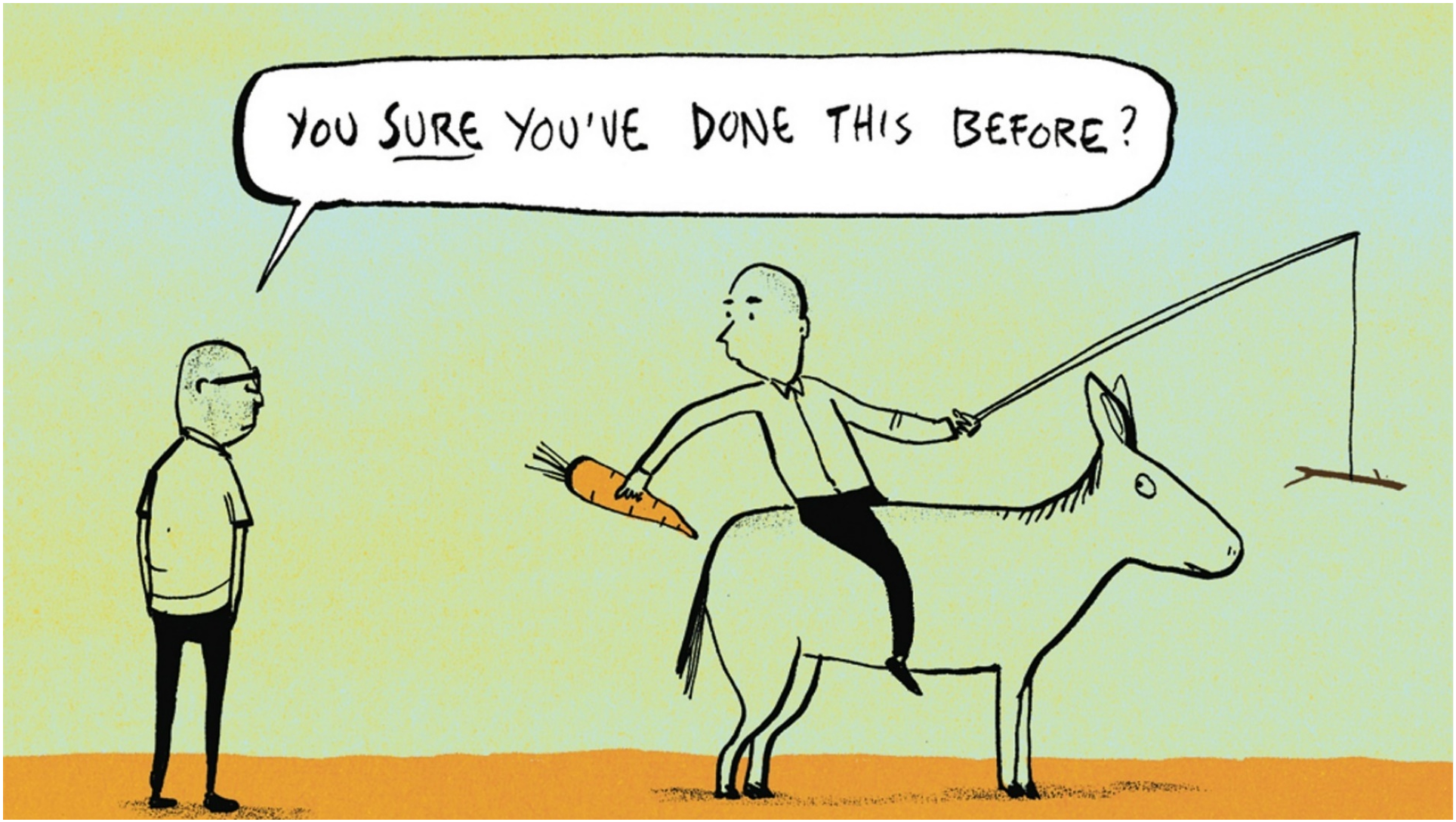

Documents by year
Documents
30
25
20
15
10
5
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year
143 peer reviewed articles in Scopus
on „citizen science" and „data quality" –
2008-2018

Documents by country or territory

Compare the document counts for up to 15 countries/territories



143 peer reviewed articles in Scopus on „citizen science" and „data quality" – *2008-2018*


YOU SURE YOU'VE DONE THIS BEFORE?

DQ perspectives

I am a motivated individual Citizen Scientist - Data Quality is important because… and this is what I do

I am a user of Citizen Science data in R&I - Data Quality is important because… what I do is…

I am a long-term Citizen Science Project of an NGO - Data Quality is important because… what I do is…

I am a policy maker using Citizen Science data - Data Quality is important and this is what I do for Data Quality

Data quality definitions

- **Practical**: fit for purpose, or intended uses (in operations, decision making and planning)
- **Philosophical**: How correctly represents the real-world construct to which it refers
- **Data consistency** is becoming a problem with increasing size of CS datasets
- **Hard to agree** on the quality of same data used for the same purpose
- **Typical terms** to define DQ: completeness, availability, standards based, validity, consistency, timeliness and accuracy
  - Nearly 200 such terms and there is little agreement in their nature (concepts, goals or criteria?), their definitions or measures (Wang et al., 1993)

What is data quality in CS

- Disciplinary conventions, standards for quality (not a closed book)
- CS data quality definitions: taken for grantedness
  - **Validity**: accuracy, confidence, completeness, error-free
  - **Reliability**: trusted and aligned with policy requirements/stakeholders

Why data quality?

- **Scepticism and distrust** by scientists and policymakers (Kosmala et al. 2016; Bonney et al. 2014, Nascimento et al. 2018; Bonn et al. 2018): citizen science is backward, marginal, unprofessional… - „public engagement" or "informal education" or "Science with and for Society"
- **Weakness in methodology** boils down to two main questions of DQ:
  - Does the project have clear processes defined to validate and guarantee high data quality?
  - Does the data adhere to common standards?

Typical anomalies

How DQ goes wrong?

- Data collection protocols are not respected by participants
  - People don’t know how to collect data
  - People “lie”
- Data collection protocols are incorrectly implemented
  - Devices are not accurate
  - Technological problems
- Data collection protocols are not verified by authorities/ stakeholders
  - Spatial inequality

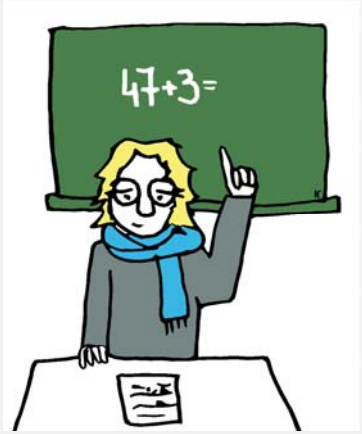
Data validation methods

1. **Peer verification**: project participants help identify and validate the observations provided by new users
    - Wikipedia or Open Street Map
2. **Expert verification**: who are your stakeholders? data curation communities
    - Once the needs of data usability are defined, solutions for data quality can be formulated (Veiga et al. 2017)
3. **Automatic QA**: filters, data mining algorithms, qualifying systems, vote for the best
    - COBWEB: human mobile-enabled sensors (Meek et al., 2016)
    - iSpot reputation score for participants (8 groups of species). The contributor's reputation acts as a quality measure of trust and can be used to evaluate their identifications over alternatives

Processes of assurance and controll

- **Assuring data quality**: preemptively restricting inputs
  - **profiling** - initially assess the data to understand its quality challenges
  - **standardization** - ensure that data conforms to quality rules
  - **Autocorrect** geocoding of address data
  - **Matching** or Linking - similar, but slightly different records can be aligned
- **Controlling data quality**:
  - **Triangulate** - combine multiple methods to ensure quality (Wiggins et al. 2011)
  - **Monitoring** - keeping track of quality over time and reporting variations
  - Use **protocols** and standards for consistency
  - Create **compatible** information systems, provide long-term storage, curate and archive
  - Use ISO 8000 as an international **standard** for quality
  - **release data** under open science principles, open-access licence

Who are your stakeholders in data curation?



Challenges for Data Quality

- **Multiple goals** of CS projects → Varied legitimacy problems around CS center in data quality, varied conventions of providing legitimacy.
- **Early stage in the development of data quality standards** for citizen science -- Literature tends to be very project specific: no clues how to transform to a more general guidance.
- **Several factors combined** makes structuring and forming the focus of Data Quality discussions in Citizen Science very challenging.

Challenges for Data Quality

- Data Quality in Citizen Science – very long spectrum
  - Data quality created on the project level but problems rarely shared
  - Methods of data generation/capture/etc.
  - The potential end users and end-use applications and purposes of the data
  - Expectations of quality (accuracy, temporality, etc)
- Possible output: **Data Quality Review Tool, a harmonized approach to data quality assurance across different citizen science projects**


EnvirMon
BioDiv
GIS/VGI/Mapping
e.g. Health & Crime
State of the art
Stories/case studies
strategies/recommendations
Synthesis of disciplines / conclusion
balazs.balint@essrg.hu
https://www.cs-eu.net/wgs/wg1/