List of standards across different disciplines:

The common ground:
ISO 19139/INSPIRE, OGC, (meta)data standards for geospatial data
ISO 19115 metadata for geographic data
ISO 19157 for geographic data quality

Aim: To Examine Data Quality Practices in Citizen Science across multiple disciplines

A multidisciplinary examination of data quality practices in Citizen Science

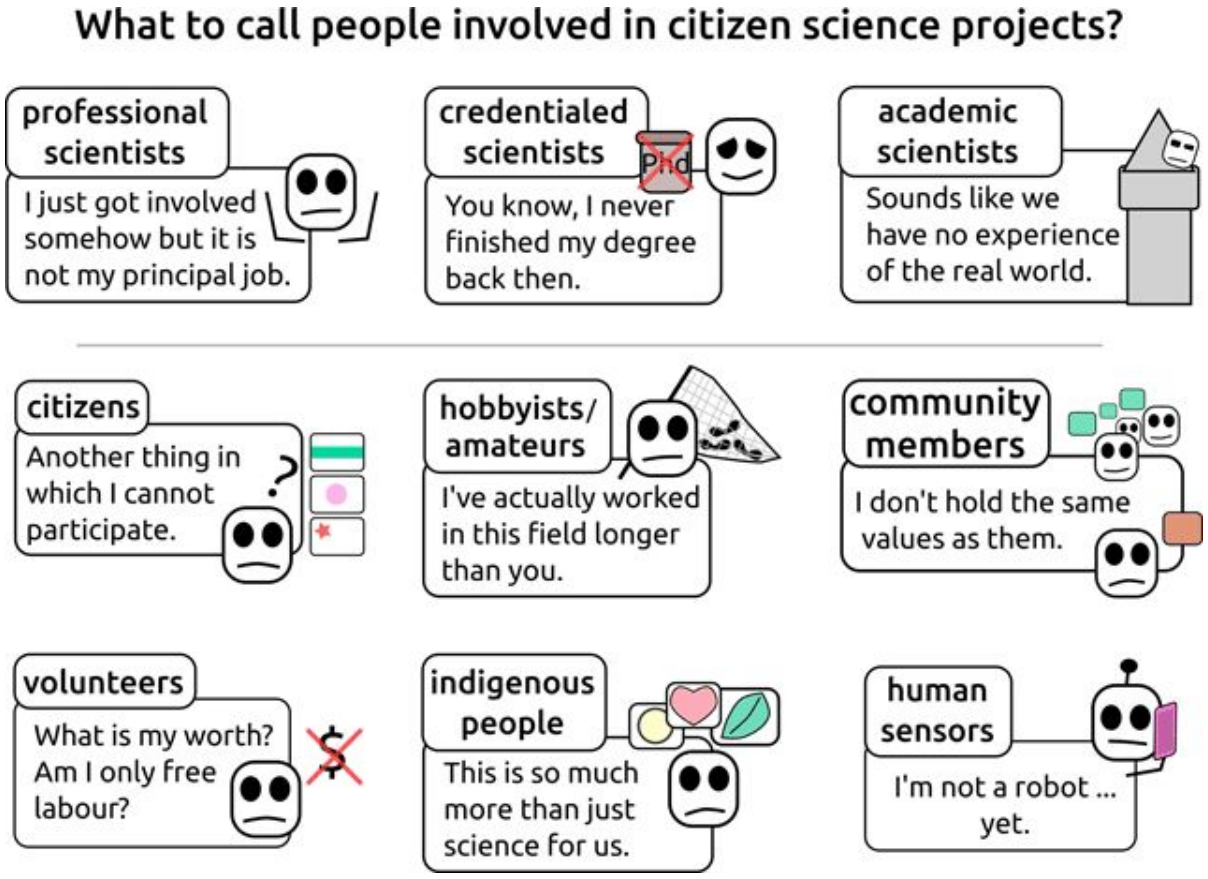
TARGET 10,000 words

1. Introduction (1000 words)

Person(s) responsible: Peter, Linda, Roland, Balint

- proliferation of citizen projects but also data from citizens more generally, i.e. VGI + give some examples of big/successful projects
- value of the data for scientific research but also by other stakeholders, e.g. government bodies, data for international agreements, mapping data for navigation
- main concern is data quality / trust of the data by users/stakeholders
- different use contexts: planning , decision making, etc.
- many papers on examining the issue of quality in different disciplines where CS/VGI is being used but often these are very specific to the project and the data being collected; cite some examples here
- there have been some papers addressing different quality elements (cite these), as well as reviews of CS/VGI data quality (cite these)
- Emphasise the point about fitness for purpose/usability as one key element that underlies the quality
- In addition, there has been the appearance of different frameworks for quality assessment (just mention these by name with references), emerging from different disciplines
- In parallel there are different standards (mention some of these), which are employed by different communities to promote open data, interoperability, sharing of data yet many projects do not record metadata, which means you need different approaches to determine quality or the data are not usable
- aim of this paper: to look across domains for different approaches to data quality assessment; to look for commonalities and differences by a) considering the range of different data elements that are observed manually and by sensors by citizens and b) the minimum data quality measures that should be satisfied within typical use cases within different disciplines; to make recommendations on how to improve data quality; to consider the future for data quality from the perspective of technology, other things ???

- Make some statements about who is this paper useful to? Is it for policy makers? It is for specific scientists? For NGOs? For large associations?



(source: https://theoryandpractice.citizenscienceassociation.org/article/10.5334/cstp.96/)

2. Review of what is available across several disciplines (4000 words)

People Responsible: Jaume, Linda, Cidalia, Andelina, Falko

2.1 Range of activities across disciplines [800 - 1000 words]

- This section is not discipline specific - it is an overview of the disciplines.
- provide some stats on number of projects, e.g. from SciStarter, and the number of disciplines + review we did last year, which also breaks VGI projects into disciplines
- could also do a quick Scopus query on citizen science, VGI, quality and then it is automatically broken down by major discipline in the outputs from Scopus - this gives us some idea of the distribution or research across domains
- Give an idea of the flavour and spectrum of disciplines.
- (wiki link: https://en.wikipedia.org/wiki/List_of_citizen_science_projects)
- Many scientific disciplines collect much the same type of data but they do it in many different ways. There is no “one-size-fits-all” approach.
- Subsequently the concept of data quality is a multi-dimensional one consisting of may non-exclusive metrics or elements. Some of these are task dependent while others focus on the design of projects or the data management practices used within those projects.

- different approaches and frameworks (see StateoftheArtFrameworksDQ.doc)
- end with the Veiga paper and make the point that this framework is applicable to any discipline
- starting point is use cases so this is the approach taken in this paper

2.2 Approaches to data quality (specific to disciplines)
the logic of these texts is story - actor - data (Jane is developing a GIS map of ...)

TARGET 600 - 800 words for each of these.

Environmental Monitoring (600-800 words)-> Wim

- State of the art
- Pocock et al 2016, Roy et al. 2012, Chandler,...

GIS/VGI/Mapping (600-800 words)→ (Linda, Cidália, Andelina)

- ISO 19157 + book chapter by Fonte et al. which covers paper by Antoniou and Skopeliti
- Goodchild and Li (2012)
- quality framework from
- summary on reviews of quality of OSM

Natural history/BioDiv Observation (600-800 words)→ Roland, Falko
State of the art

- refer to GEO BON activities
-

- using databases such as http://www.buergerschaffenwissen.de/, EUMON (http://eumon.ckff.si/), http://www.citizen-science.at/, http://biodiversity.eubon.eu/web/citizen-science/data-providers to outline the landscape of projects
- maybe apply a classification/"clustering" according to data quality/sampling effort

Harmful Species Monitoring - Jaume

- State of the Art

Existing projects for vector borne diseases mapping : Mosquito alert (Spain) (http://www.mosquitoalert.com/en/), Muggenradar (Netherlands), ZanzaMapp (Italy), MosquitoWEB (Portugal)

Other (?)

- State of the art

e.g. spatial planning → https://maptionnaire.com/

3. Identify common aspects and a more generic framework

TARGET 800 Words
→ Table on data elements with ranking of importance
Explain the table - what it is, how it is annotated, how it should be used etc.

- introducing stories/applications/use cases

Disconnect data quality from strategy/design! Focus on the quality of the observation.
Provide minimal data requirements (overview existing standards for the identified data elements) without specifying the scope/research objectives in too much detail. As project design is deliverable 2 which can include more detail on various levels of DQ requirement which depend on the design.

Create a table with data elements/types versus disciplines - level of mandatory to optional. Data elements included should have standards behind it.
(see Elements-Grid.xls)

- be consistent across the stories.
- choose a good set of common attributes for use cases. These can be extracted from the 'Elements Grid' which we have created.
- Think about the type of data and what happens to these data. Consider the data object which must be generated as part of the observation in the stories.

TARGET PER STORY (600 WORDS)
Story - Actor Scenario: Natural History/Biodiversity. (Roland, Falko)

Case 1 (Falko): The story is that you need the old (legacy) data and the new (current) data to study what has happened over time. (time series analysis) Citizens are involved in new data capture and in transcription / classification of legacy data. Comparable data quality is important to merge both types of data.
Actors: NH staff, agencies, citizens

Case 2 (Roland): NGOs and their nature conservation strategies/actions (low requirement on data quality), risks: actual presence/absence, false positive reports, credibility of NGOs

Story - Actor Scenario: Environmental Monitoring

Case 1 (Wim, Luigi): Water or air quality using manual or sensor data - what, where, when, who, ideally ´how´ focus on quality observation (e.g. focus on those common elements (described in section 3) that are extremely important for environmental modeling (so no difference between for example water or air pollution monitoring), take in account that manual and sensor might need different requirements but not necessary independent)

Scenario (the data-quality issues need to be highlighted)
Sultan is a university student at the Istanbul Technical University - Environmental

Engineering Department. She and her friends are aware of the air and water quality problems in their environment. They need sensors which are able to monitor water and air quality in their district (Başakşehir) and in the Balıkçı Adası Parkı marina. According to their initial research, there is no such a monitoring system functioning in Başakşehir district and in the Balıkçı Adası Parkı marina. They are also planning to create a business around the implementation of this monitoring system.

Needs and concern

Sultan and her neighbours are concerned about the air and water quality in their urban neighbourhood and in a marina nearby and they learn through inquiries to their regional environmental-protection agency's office that air and water quality monitoring stations are not granular enough to represent their neighbourhood and the marina.

Current situation and challenges

In the current situation, Sultan purchases an air and water quality sensor she finds online and recruits neighbours to place the sensor outside their homes and in the marina. Some neighbours participate and deploy the sensors. These sensors collect data for six weeks. According to Sultan's research, the data demonstrate a violation of the national air- and water-quality standards. She and her concerned neighbours share these data with their regional environmental-protection agency's contact who informs them that the data cannot be used because of quality-assurance issues. They are frustrated and left wondering if the regional environmental-protection agency's is hiding something.

Enabling infrastructure and future scenario

Through a quick online search on a citizen-science platform, Sultan finds a local project supported by the regional environmental-protection agency in need of help. The citizen-science platform makes it easy for them to join, and the "tools" section of the platform makes it easy for them to read reviews of low-cost, DIY instruments, including those generating data accepted by the regional environmental-protection agency's standards, and to build, borrow or buy the tools. Sultan and her neighbours regularly meet up for training, and to share their concerns with the regional environmental-protection agency's office. The citizens and the regional environmental-protection agency develop mutual respect for each other and work together to discover and address community concerns.

(Case 2 (Falko): Polluted beaches. Biodiversity and pollution monitoring on beaches.)
Actors: citizens (i.e. tourists) gather and identify objects on the beach. These are biotic and abiotic things (biodiversity vs. pollution)
KdUINO is a DIY instrument for water transparency (developed in the CITCLOPS project). Although its relative low cost, it provide haigh quality measurements. See http://www.mdpi.com/1424-8220/16/3/373 Figure 13.

Maybe better alternative: EyeOnWater-Colour (CitClops funded mobile app)

DIY air quality sensor: http://luftdaten.info/

Story - Actor Scenario: VGI / Mapping

Case 1 (Land Cover/Land Use): Actor: National mapping agency wants to use citizen data to update and correct their maps. This includes opportunistic information (e.g. from OSM and social media) and directed (through a specific application developed by the agency).
Type of data needed (from the table): Explicit location data; photographs; vector data; time; classifications; user confidence; user id; collection methods (heads up digitization or smartphone GPSS) .

Case 2 (Disaster Mapping/Response): Actor: Civil Protection Entity or Mitigation institutions. The data is required to assist the civil protection in identifying the needs for emergency response (and mitigation activities) in the case of an extreme event.
Type of data: georeferenced photographs, videos, posts, tweets.
Two different aspects may be of importance: 1) Data about the occurring event or 2) Data about the region where the event occurs.
For the first aspect time of capture is important, as only data that reports the event or conditions surrounding it are important. Therefore time of data capture (at a day or even hour level may be relevant) is an element of data choice and identification. Geolocation is also fundamental, preferably with explicit location data (coordinates). If implicit geolocation is available, the data may not be that relevant, however, if it a restricted location can be associated to it (such as a street) it may be relevant.
For the second aspect, data about the region surrounding the location of the event may be useful to anticipate what secondary hazards may occur (for example location of petrol stations in case of a fire) or communities that need special actions (schools, nursing homes, immigrant shelters). For these data temporal information is relevant, but less precise data is useful (the time granularity required is much wide - data of a particular year may be useful).
Validation approaches used: expert validation + ????

FINAL STORY IS INTERDISCIPLINARY
Consider a story which works across the disciplines in the Elements Grid.
Story - Actor - Scenario: **Peter (suggestions welcome)**

An oil spill at the beach that has consequences to biodiversity, environment, land use …

Case TBD

Story - Actor - Scenario: Harmful species monitoring (Jaume) John is interested in developing a system for reporting the occurrence of harmful species (vector diseases, crop damaging, …). He is taking into account that the system will be focused on specific target species and also that in many cases there will be sensitive information. He designs an app to report the observations: the app includes information about how to identify the target specie and also a questionnaire on the key features to spot. This will help in evaluating the credibility and the quality of the observation. The app also offers the possibility to validate observations from other volunteers

Other stories -
Scenario: Biomedical literature annotation (Federica) Biomedical literature represents one of the largest and fastest growing collections of unstructured biomedical knowledge. Finding critical information buried in the literature can be challenging. To extract information from free-flowing text, researchers need to: 1. identify the entities in the text (named entity recognition), 2. apply a standardized vocabulary to these entities (normalization), and 3. identify how entities in the text are related to one another (relationship extraction). With web-based application Mark2Cure (http://mark2cure.org), it was demonstrated that NER tasks also can be performed by volunteer citizen scientists with high accuracy. Data quality issue was addressed by using replication across multiple participants, having participants evaluate established control items, and using a corpus of text that already had been expertly reviewed as a benchmark (Tsueng, G. et al., (2016). Citizen Science for Mining the Biomedical Literature. Citizen Science: Theory and Practice. 1(2), p.14.)

4. Foresee the future about DQ assessment in citizens Science (1000 words)
People Responsible: Peter, Balint - all can suggest points to include in the vision for the future.

Start with a paragraph or two summary of the contents of the paper - revisiting what we have promised at the beginning of the paper namely “to look across domains for different approaches to data quality assessment; to look for commonalities and differences”.

- We have provided an overview of “fit for purpose” DQ for various domains (in the review), that hopefully solve some of the commons issues within domains but also across domains (i.e. added value this review of various frameworks, and identification common data quality aspects)
- Highlight apparent fields/concerns that require attention (and propose solutions?):
    - where technological innovation (automated methods etc) is required to solve some outstanding issues for the various domains,
    - solve data accessibility/ownership issues, for example open data licenses are sensitive or tricky topics for each domain where every domain handles them differently
    - Can the users of these data by key drivers in pushing the need for interdisciplinary work on these issues? It is difficult to get agreement across domains or disciplines. However we are seeing more needs for the integration of data which is taken from different domains.
    - DQ requirements in terms of communication increase of trust by users and stakeholders
    - take into account citizen science data peculiarities and citizens as sensors
    - create CS specific DQ standards or adapt existing ones
    -

To use and integrate diverse types data it would be useful to have a common ground regarding the data quality assurance.

From the "CSA 2017 Data Quality Workshop":

How can practitioners communicate data quality to establish credibility?

Communication depends on audience. Funders, academics, and community groups all require different types messaging.

- Data quality practices can be communicated through training, workshops, and demonstrations as well as text.
- To ensure broad understanding, a plain language format is important.
- A range of publication mediums may include social media, success stories, and scholarly publications. Peer reviewed publications are helpful from the perspective of peer validation, but time consuming to produce.
- Data sets and visualizations can also be published and shared.
- Researchers and practitioners can also cite other citizen science data sources and publications to further validate the quality inherent in those results.
- Federal agencies and other authorities can also work with their agency communications offices, for example to release press releases.

- Convincing audiences about data quality will be easier if we can point to standard protocols for data collection and analysis, and use standard language to describe practices.
  - These could be developed, shared, and verified or validated by a data quality working group or other authority.
  - In general, project protocols should be published and open to commenting. Projects should establish QA/QC procedures, including around data verification or validation, and discuss how these procedures impact the results.
  - It's important for the researchers associated with a project to know their QA/QC plan and be able to demonstrate it in action.
  - Researchers should also publish comparison studies when possible and appropriate.
- In addition, protocols and individual participants might have more credibility in the context of a validation, or a "seal of approval."
  - This could look take the form of starred reviews or a "seal of approval." Depending on the desired level of formality, different authorities could grant this:
    - A single organization could lead and manage this process.
    - Within CSA, there could be a citizen science data quality working group that manages peer review.

            - Projects could give feedback to volunteers and ask for their review.
        - Metrics for ranking could include data quality, user reviews, citation of data.
        - Case studies are always helpful.
        - Organizations like the Encyclopedia of Life (EOL) and iNaturalist are already looking at or including data verification.
        - Publication is an important form of communication. Data may be published in repositories; research may be published in peer-reviewed publications; etc.
- There is value to transparency of protocols and data, including transparency about the limitations of data.
    - In line with the ideals of open science, the process of citizen science data collection should be “visible” and “transparent,” as with professional research.
    - From the early stages of design, projects should work with recipients of data to increase confidence in the data collection process. Projects should also work with their volunteers to ensure full participation of all possible data beneficiaries.
    - Projects should explain fitness for purpose, or how data types are sufficient to answer a particular data question (and/or other questions).
    - In addition to establishing credibility and trust, communicating data practices and intended use can help projects relate to other citizen science community members by identifying shared issues and concerns.
    - Data could be made accessible to everyone in its post-validation state.
- Sharing previous success stories can be helpful in communicating data quality.
    - Success stories also demonstrate a successful return on investment (ROI).
    - It may be best to start with least controversial cases first, and expand from there.

Bibliographies:

Mooney, Peter, et al. "Towards a Protocol for the Collection of VGI Vector Data." *ISPRS International Journal of Geo-Information* 5.11 (2016): 217.