The balance between openness and privacy for health data collected through citizen science: various perspectives

Ria Wolkorte

Lieke Heesink

Michelle Kip

At the Citizenlab, we want to do research in which researchers and citizens work together, and in which everyone can bring in their own expertise. This is also called citizen science. This research can be done in different ways, such as:

- qualitative methodologies (such as interviews, group discussions or co-creation sessions);
- quantitative methodologies (data collection on a website, wearing activity trackers);
- a combination of the above (such as questionnaires with open and closed questions).

The citizens in the project regularly have a double role; they think along with the research (in the role of co-researcher) but also supply or collect data about themselves or about their own health or well-being (in the role of research participant).

Data

It is important that the **security and privacy of data** is ensured, especially since our research involves collecting data on people's health. It is also very important that research participants always **know what their data will or may be used for**, so that they can make a good decision **whether or not to give their consent** and thus participate in the research.

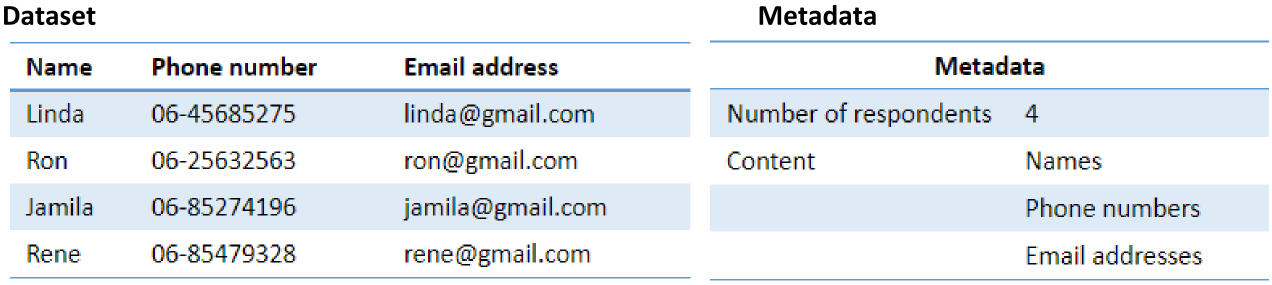
FAIR: what is FAIR data?

Within citizen science, it is considered very important to work **open and transparent**, which also means that you should be as open as possible with the data that you collect during a research project. To achieve this, we try to work in citizen science projects according to the principles of **FAIR data**. FAIR includes guidelines for describing, storing and publishing scientific data. An important goal of FAIR data is to **make scientific data suitable for reuse**, for example so that other researchers can use these data in other research. Another advantage of FAIR data is that it is possible to check whether the research has been carried out correctly. The rule is: **as open as possible, but closed where necessary**. The considerations here depend on the type of data. As a result, different choices can be made regarding the publication of data on air quality for example, compared to data on a person's health.

Metadata

Within the principles of FAIR, it is not only about data, but also about metadata. **Metadata is data about data.** A list containing telephone numbers and names is the data itself, the description of this list ("this list contains 4 telephone numbers and their corresponding names") is called metadata. But also, for example, the name of the author of the data, or the number of pages that the data consists of is referred to as metadata (see Figure 1).

Within the principles of FAIR, you can choose for example not to make the data of the research itself public, but only the metadata; then others know that the data exists and what information it contains, but others cannot view the content.



Dataset

Name	Phone number	Email address
Linda	06-45685275	linda@gmail.com
Ron	06-25632563	ron@gmail.com
Jamila	06-85274196	jamila@gmail.com
Rene	06-85479328	rene@gmail.com

Metadata

Metadata	
Number of respondents	4
Content	Names
	Phone numbers
	Email addresses

Figure *1. Example of a dataset and a part of the corresponding metadata*

Open data and storage of data

Scientific research data is increasingly stored in an **online** storage place which is called a **repository** . This is a central place, on a website, where data is stored. Here, researchers can find data collected by others which they may also be able to use for their own research. In this way, they do not have to collect the data again themselves. **Anyone can search for datasets in a repository without paying or logging in.**

In such a repository, you can place research data with different levels of openness:

- ‘**Open data**’. Open data refers to datasets that are freely accessible. This means that everyone has access to these data. Anyone can download these datasets without reason or payment.
- ‘**Access only on request**’ means 'accessible with a good reason'. This means that only the metadata of a study is published in the repository, indicating that researchers who are interested in the dataset can contact the original researchers. It can then be discussed whether these researchers will get access to the (anonymised) data and under what conditions.
- ‘**Metadata only**’. Another option is to only include the metadata of a study in the repository. This means that others never have the opportunity to view or work with the data. An important reason to choose for 'metadata only' is because the dataset itself contains personal data, whereby the dataset loses its value when these personal data are removed. In addition, this can be chosen when participants have not given their permission to share the data (publicly).

Anonymisation of data

If data of people are used in a scientific article or stored in a repository, it is important that these data are anonymous. This means that the data **cannot be traced back** to the people who participated in the research. There are several ways to anonymise data. For example, you can change a date of birth into a category (for example, the category 30-40 years, instead of 21 July 1983). Or you can omit sensitive information (such as names or medical history) from the data set.

Results of 2 group conversations

In order to decide how to deal with data in the future within the Citizenlab, we set up conversations. Four people with arthritis took part in the first meeting, all of whom had previously participated in Citizenlab research. The second meeting was attended by two **people with arthritis, a citizen science researcher, an ethicist, and a data steward** (a data steward is concerned with the proper collection, storage and processing of data). In general, the opinions of the participants were very similar, although sometimes small individual differences were mentioned.

Importance of sharing data

During both sessions, all participants agreed that sharing research data is important. This can make research more efficient; certain data collection does not have to take place again. This reduces the burden on participants and possibly the costs of research. A participant's contribution can also have a greater impact because it is used for different studies.

Importance of privacy, consent and anonymisation

According to the participants, privacy is of high priority when participating in research. Data may only be shared and/or reused with other researchers if participants have been explicitly informed about this and have given their consent. It is very important that the data is anonymised. This means that personal data, such as names, e-mail addresses, and other recognisable data are removed or replaced by, for example, a number or a general description. This means that personal and other identifiable data are never shared outside the research for which the data was initially collected.

Data within the Citizenlab

There are **different types of data**, varying from interviews in which a lot of (personal, possibly sensitive) information is shared, to continuous registration of a heartbeat (which in itself is not traceable to a person). The participants indicated that with qualitative data much attention should be paid to good anonymisation. Once the data has been properly anonymised, all types of data should be treated in the same way.

The participants indicated that they would prefer

- **publishing metadata in a repository**.
- **access to the anonymised database that can be obtained by a researcher with a legitimate research question**; it is up to the researchers to determine what a legitimate question is. Criteria could possibly be drawn up for this.

Inform in advance

According to all participants, it is very important that it is clearly explained **before participation** what will happen with the data. In this way, people can make an **informed choice** about whether to participate in a study. However, a balance must always be struck between **complete clarity** and the **readability and length** of the explanation. Various suggestions have been made for doing this. For example, the explanation could be short and concise, with a button “ ℹ ” being used to give participants the option of reading more extensive information. Another option would be a short video clip with an explanation, and the possibility of receiving more information from the researchers.

Conclusion

When storing, processing and possibly sharing the data, it is especially important to be open and transparent about the research. This allows potential participants to make a well-informed decision about whether to participate. The importance of sharing research data is recognised, but privacy must always be guaranteed.

Based on all the information obtained from these interviews and additional discussions with data stewards, privacy specialists and specialists in human subjects research, we will write a manual on the conscious use of data obtained in research and citizen science. Among other things, this manual will describe:

- what information participants should receive in advance.
- that data must be anonymised and what researchers should pay attention to in this respect.
- That, when applicable, the data is preferably translated into English, so that it can be reused/viewed by a larger group of researchers.
- which options there are with regard to sharing data when data is stored in a repository and what these options entail.
- how to describe good metadata of the dataset, so that others know what kind of data has been collected.
- How to share data safely with other researchers.

This manual can always be adapted on the basis of new insights, and of course - where desirable and substantiated - deviations from the manual can be made within projects.

For more information or participation in follow-up studies, please contact us via r.wolkorte@utwente.nl.