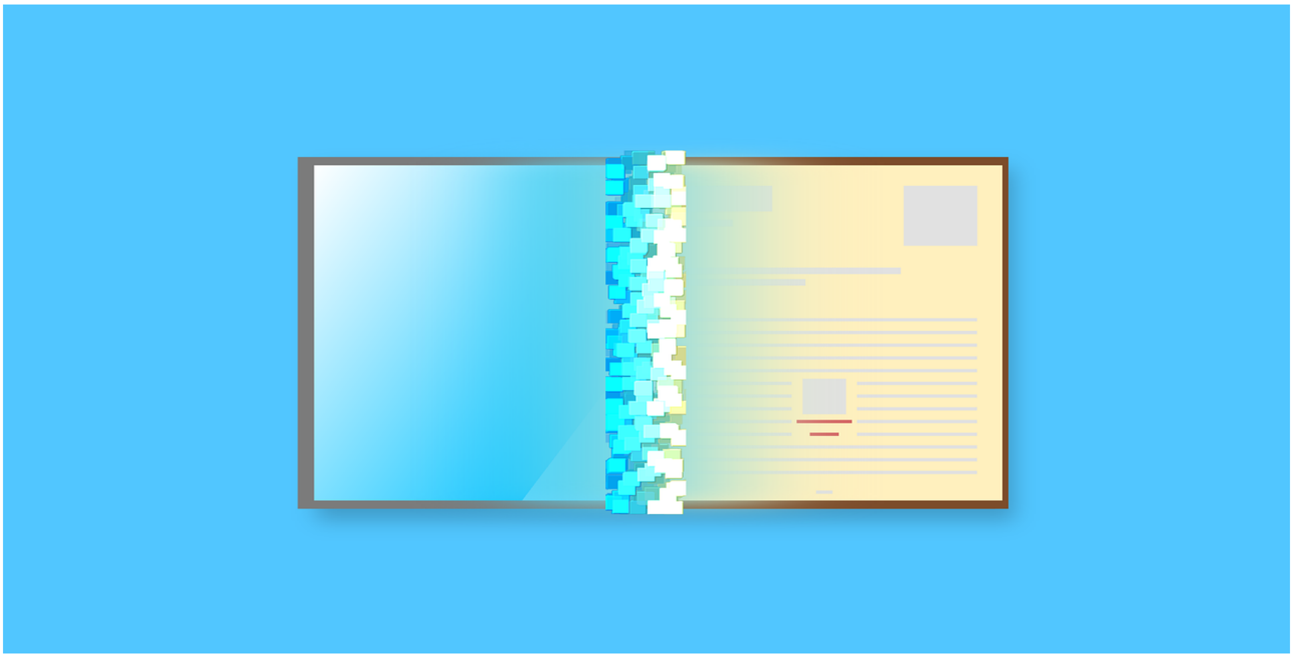


# Find, Use, Cite, Repeat: A short guide on how to cite microdata from Statistics Netherlands

*Written by Emilio Cammarata and Angelica Maineri (ODISSEI FAIR Support team, FAIR Expertise Hub)*

**Bricks for the construction of knowledge: recognizing value of the data**

Data is vital to generate insights and validate findings in the social sciences, and beyond. Yet, finding, evaluating, and reusing data sources can be time consuming for data users. Despite efforts in implementing the FAIR principles (see the ODISSEI FAIR support page), datasets are often scattered throughout many repositories, and they are often only accessible under conditions that vary significantly between different data providers. For data producers and data owners, on the other hand, it is difficult to track who uses their data and in which context, unless the (re)use of data is properly documented.

Just like there are standards to reference bibliographic sources, specifications to properly reference data are also available, for instance, those recommended by DataCite or APA. Citing data in scientific works has multiple advantages: first, it increases the reproducibility of a study and hence its validity; second, it enhances the reusability potential and the visibility of data sources; third, it increases the recognition of the authors' work behind the data. Moreover, accuracy in the report of data in the bibliography is highly desirable in the process of citation. All the previous factors are good reasons why every researcher should cite datasets in a publication, and there are no drawbacks to that.

The goal of this short article is to summarise the main principles of data citation, to apply these principles to the citation of Dutch administrative data via Statistics Netherlands (CBS) and finally to show how the ODISSEI Portal can aid in the process.

**The (recent) development of data citation standards**

Two years before the publication of the FAIR principles, in 2014, a 'Joint Declaration of Data Citation Principles'(JDDCP) was published [1]. Its aim was to align the academic community around common principles related to datasets' citation, as a lack of consensus was detected. This declaration states that 'data citation…is part of the scholarly ecosystem supporting data reuse', not less than academic publications [1].

Specifically, the declaration pointed to eight principles:
- *importance*, researchers should give to data the same importance that is given to publications themselves;
- *credit and attribution*, which should be given to the authors of the datasets, both at a legal and scholarly level;
- *evidence*, in every case there is a claim on data that should be cited;
- *unique identification*, every dataset should have a unique and persistent machine-readable identifier in order to find every identifier for a specific dataset;
- *access*, the use of data citation should lead to easier access to data and to their associated metadata;
- *persistence,* metadata and in particular unique identifiers should persist through a long period of time;
- *specificity and verifiability*, the use of data citation should include the origin and version of the data in order to verify in an easier way the specific dataset;
- *interoperability* and *flexibility*, data should be flexible in order to be used across communities, but at the same time it should be usable in conjunction with other resources/datasets.

In practice, data citation consists of reporting the most basic metadata of the dataset - i.e., information about data. The citation of a dataset should include at least the following information:

- organisation/s or person/s *author* of the dataset;
- *year of publication* of the dataset (which can be different from the year of data collection);
- *title* of the dataset;
- *persistent identifier* (e.g., a DOI (Digital Object Identifier,...);
- *distributor* of the dataset (if different from the author);
- *version* of the dataset (if available).

Five of the eight principles of the JDDCP are reflected in this basic metadata requirements:
- importance is given to the author/organisation that was involved in the creation of the dataset and with it has been given credit and attribution to it;
- evidence, unique identification, access, and specificity are supported by the use of a persistent identifier (or PID)[3]. The use of PIDs is an important step toward the findability, documentation, reproducibility and reuse of scholarly data. The value of assigning PIDs and adding them to the metadata are further discussed in this blogpost (also available on Zenodo).

The other three JDCC principles (e.g. verifiability, interoperability and flexibility) are reflected by the accessibility of the dataset itself and therefore are not directly linked to the elements included in the data citation.

Whenever a dataset is referenced in a scientific article, the PID provides access to the dataset page by making the metadata (and frequently the data) accessible and easier to find for other researchers. This process is depicted in the diagram below [Figure 1].

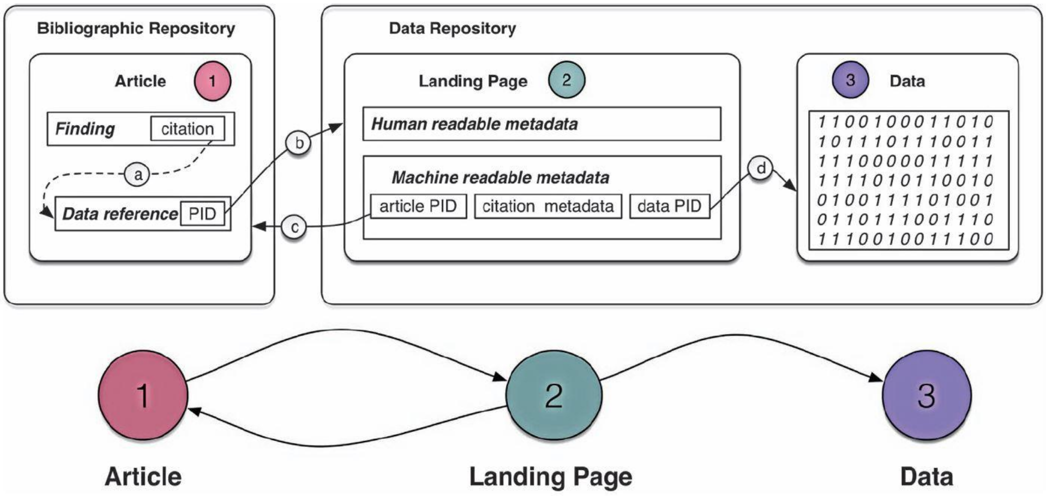


Figure 1. Image of the connection between an academic paper and its dataset [3].

Measuring datasets usage by tracking citations also helps measuring impact for those who created and published the dataset, e.g. as is done on zenodo.

### Use case: the CBS microdata

So far we have discussed data citation in a general way. In the following paragraphs we will focus on the case of data citation of Dutch administrative data held by Statistics Netherlands (CBS). CBS Microdata are data gathered at the micro level on "individuals, companies and addresses which can be made available to Dutch universities, scientific organisations, planning agencies and statistical authorities within the EU under strict conditions for statistical research"[1]. Microdata is highly sensitive and for this reason access is constrained to the CBS remote access environment. The regulations and guidelines behind the access procedure are described in a pdf available through the CBS website at this link.

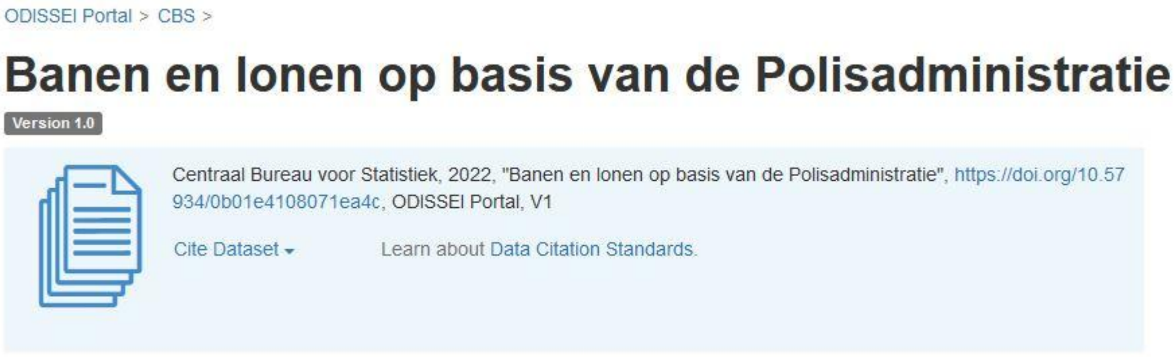
### Using the ODISSEI Portal to find and cite CBS microdata

Before citing the data, it is important to be able to locate the data and its accompanying description and documentation (e.g.., its metadata), as stated above. In the case of CBS microdata, the data are documented in pdf files available on the CBS website. However, these files are not easy to search through. For instance, the thematic focus has to be chosen before the titles of the pdfs are even visible, and it is not possible to search the content of the pdfs for specific keywords. More importantly, since the pdfs are easily readable for humans but not for machines, once a dataset is found, reconstructing a data citation from these files requires manual work, making the process burdensome for users and prone to mistakes.
To facilitate data discovery, the metadata of CBS microdata has been ingested and published in the ODISSEI Portal. While the accessibility of the microdata remains conditional to the CBS Microdata service, the metadata of the CBS micro datasets can be browsed through in the ODISSEI Portal. Users

[1] https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research

can find the metadata through the search bar or through faceting filters based on the following characteristics: the year of publication, the keyword term related to the datasets, the name of the distributor and the topic classification term.

Once the dataset of interest is found, the ODISSEI Portal enables users to cite it easily. The block “Citation metadata” under the tab “Metadata” includes all the mandatory information to build a data citation according to the principles outlined above, namely: author (Centraal Bureau voor Statistiek), year of publication, title, PID (doi). To make it even easier, the ODISSEI Portal already offers a suggested data citation (see image below) on the top of the page and offers different options to export the data citation in formats that are compatible with most reference managers.



ODISSEI Portal > CBS >

Banen en lonen op basis van de Polisadministratie

Version 1.0

Centraal Bureau voor Statistiek, 2022, "Banen en lonen op basis van de Polisadministratie", https://doi.org/10.57934/0b01e4108071ea4c, ODISSEI Portal, V1

Cite Dataset ▾

Learn about Data Citation Standards.

Figure 2. Metadata of the dataset on ‘Jobs and wages based on the Policy Administration’ [doi].

It should be noted that in addition to citing datasets, data owners may have additional requirements. For the use of CBS microdata the publication requirements include:

1) Indicating the following as a source: “Results based on calculations by [name of research institution or commissioning party] using non-public microdata from Statistics Netherlands.”
2) If a data availability statement is required, the following text should be used: “Under certain conditions, these microdata are accessible for statistical and scientific research. For further information: microdata@cbs.nl.”

Using the ODISSEI Portal offers a great advantage for the researchers. A (potential) user can easily find a CBS microdata that is relevant for their particular research question and refer to it when communicating with colleagues or when requesting access to CBS microdata. By providing the DOI of the CBS microdata record in publications and working papers, others are able to uniquely identify which dataset is referred to.

## Concluding remarks

In this blogpost, we demonstrated how easy it can be to cite data via repositories like the ODISSEI Portal, as well as the advantages for doing so. Data citation has many benefits for researchers, including more transparency and increased data findability, and for data owners, e.g. for tracking data usage and then their impact throughout the academic community. Therefore the last suggestion that we want to make is this: cite the data you use and contribute to making this a standard academic custom!

*Do you have any questions on how to cite a certain dataset, or do you have any questions related to RDM and FAIR? Reach out to fairsupport@odissei-data.nl*

**References**

[1] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
[2]  Starr et al. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJComput. Sci. 1:e1; DOI: 10.7717/peerj-cs.1 .
[3] Cousijn et al. (2018) A data citation roadmap for scientific publishers, *Scientific Data,* 5. Doi: 10.1038/sdata.2018.259.

Featured image by pixabay.com, scientific data (nature) and CBS.