Collective Dynamics

Be FAIR to Pedestrian Dynamics Data

Maik Boltes · Alica Kandler

Institute for Advanced Simulation, Forschungszentrum Jülich, Germany,
E-mail: m.boltes, a.kandler@fz-juelich.de

Received: 31 October 2023 / Last revision received: 12 April 2024 / Accepted: 12 April 2024
DOI: 10.17815/CD.2024.163

Abstract For improving the safety of people in large crowds, it is of great importance to understand the basic mechanisms of pedestrian dynamics, *e.g.* with help of experiments. The number of openly shared datasets of these experiments has increased in the last years also due to stricter requirements from journals and funders. We share our own experimental data by an open access data archive which data is widely used in the community.

However, our data and also data of other researchers in the field of pedestrian dynamics is not annotated in a systematic or semantically harmonized way, which impairs FAIRness in general and interoperability specifically. In this paper, we propose a standardized extensible metadata schema and key data structures for trajectories and geometry. The proposed metadata schema and data structures hopefully support the interoperability within the community and will assist to make data reutilization more efficient.

Our own legacy datasets are continuously annotated with essential information using this metadata schema steadily. This metadata is provided beside the converted data on our data archive and thus enhance its findability and reusability.

Keywords FAIR data principles · metadata schema · data archive · database · trajectory format · geometry format · open access · standardization

1 Introduction

Enhancing the safety and reducing the risk of injuries or even fatalities in large crowds is one of the main purposes of pedestrian research. Hereby, understanding the fundamental mechanisms that govern crowd motion or behavior is essential, *e.g.* by means of controlled experiments [1, 2].

By providing openly accessible data outside the own organization or collaborations, the possibility to reproduce findings and perform further analyses emerges, increasing the overall scientific rigor. Therefore, we share all our experimental data after scientific evaluation since 2005 and from 2017 on in our open access data archive [3].

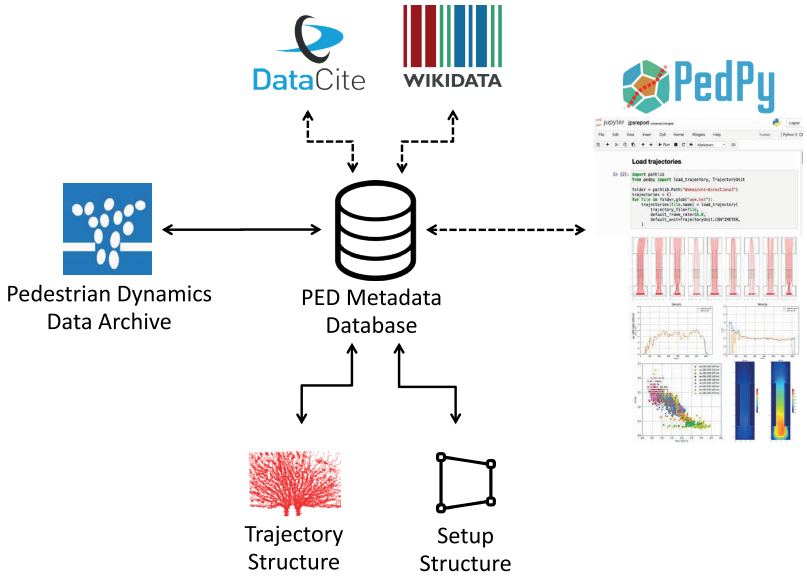

Figure 1 The embedment of the metadata database and the connection to the data archive and the usage of standardized data structures. Dashed lines show upcoming features of mirrowing to well known data repositories and the direct access from external software.

However, the data provided is described in varying degrees of detail, with lack of consistent, systematic structure, and different data structures and formats over time. Thus, in order to reuse data from the data archive, a lengthy search process of different resources, like corresponding papers, must be carried out.

For this reason, we developed a standardized extensible metadata schema to annotate our data with elaborate metadata and developed or determined standardized data structures for two of our main data types, the trajectory and the geometric setup data.

With this step we can enhance the overall data FAIRness [4] for all researchers in the field of pedestrian dynamics, which implies that data should be **f**indable, *e.g.* by a detailed description with metadata in a searchable environment, **a**ccessible, *e.g.* with easy access to data and storing the data in a long-term storage, **i**nteroperable, *e.g.* by using exchangeable formats functioning across different systems, and **r**eusable, *e.g.* with metadata including the process of data creation for easy reproduction of data.

Overall, these implementations enhance the general research data management, allows to build up a consistent documentation, helps to find, use and reuse data more efficiently and prevent the loss of crucial information. The annotation, conversion and enrichment of our legacy experimental data is ongoing and provided step by step [3]. All trajectories and setups are converted to the new formats.

Furthermore, the developed metadata schema and the standardized data structures could help to establish a basis for coherence within the research community of pedestrian dynamics. This would enable the development of shared tools within the community, *e.g.* for data analysis as for example with the library PedPy [5].

Fig. 1 shows the mentioned connection between the database of metadata with the existing data archive using standardized data structures. The dashed lines show the idea of further improvement of the usibility by mirrowing to other metadata repositories and a direct access from other software through a database query.

2 Metadata

Metadata, which has been defined by the ISO Records Management Standard as "data describing the context, content and structure of records" (see section 3.12 of [6]) is of great importance, especially when trying to comply with the interoperability and reusability aspects of the FAIR data principles. Due to the fact, that experiments in the field of pedestrian dynamics can vary greatly in content, setup and recorded data, we chose to create a new metadata schema, rather than adjusting an existing one. The metadata schema for the field of pedestrian dynamics was developed based on the assumption that each performed experiment is documented individually. An experiment usually consists out of 1 or more runs which can have none, one or multiple varied parameters between them.

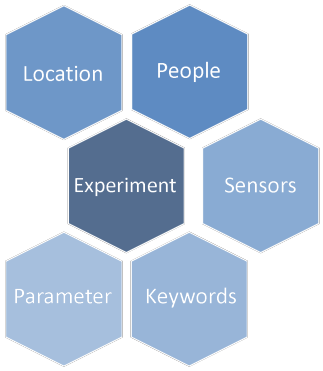

Figure 2 Main categories of metadata.

In preparation to building the schema, metadata of representative types of experimental datasets was studied and 6 main information categories were singled out (see Fig. 2). Based on these categories, general administrative information as well as specific details can be documented side-by-side in the metadata schema for experimental data in the field of pedestrian dynamics.

2.1 Schema

The metadata schema is built around the central section of general information about the experiment. This includes properties as for instance the name of the experiment, the data collection date as well as a short description about the purpose of the experiment and links to further documents describing the experiment in more detail. The *location* section contains the affiliation at which the experiment has been performed, by the use of its unique ROR ID (Research Organization Registry Identifier) [7]. Additional metadata about the location, as address and building name allow for a detailed documentation even if the affiliation does not hold a ROR ID. The *people* section documents the people responsible for the collection of the dataset, which can range from technical to content planning personnel. This section utilizes unique identifier, the ORCID (Open Researcher and Contributor Identifier) [8], which allows a distinct identification of the researchers involved. Furthermore, the ROR ID is utilized again to draw a connection between a researcher and its affiliation at the time of the experiment. The *sensor* section describes the technology used as well as its crucial settings, for instance the framerate at which the data was recorded. The *keyword* section facilitates the better findability of suitable metadata and their corresponding datasets. Lastly, a *parameter* section holds specific details of the individual runs. This includes the name of the run, the starting time, the geometry, which may change between runs, the status (*e.g.* planning, performed, data prepared, data revised, published) and a set of parameters and its values (*e.g.* bottleneck width, number of participants or instruction given).

Overall, the schema allows a structured collection of metadata for a large variety of datasets in pedestrian dynamics. Even though the schema was created through an ex-

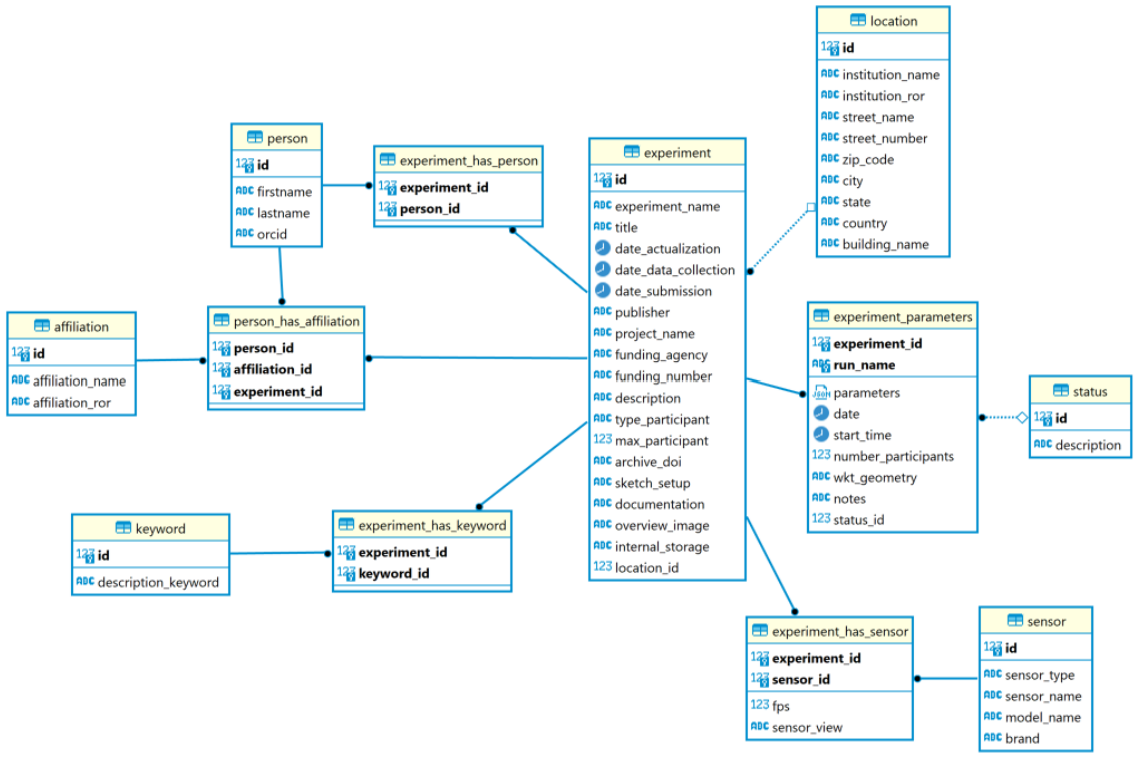

Figure 3 Entity relationship diagram of the metadata schema.

tended work process, it can be easily adapted to serve the needs of other working groups.

Examples on how to access metadata as well as a template of the metadata schema can be found on the information page of the data archive [9].

2.2 Database

To be able to store, maintain and later access the collected metadata in an orderly fashion, a database can be setup to facilitate this process. The entity relationship diagram (ERD) shown in Fig. 3 visualizes the structure of a database resulting from the metadata schema. Junction table enable many-to-many relationships like *experiment_has_kewyword* for *experiment* and *keyword*.

Experimenter are asked to already collect metadata while planning and performing new experiments. This helps keeping the data mangement plan up to date and supports uniform and directly available metadata and thus promote the FAIR data principles. Keeping track of the metadata can be done using a metadata template [9] or through direct entry in the database.

To ensure good accessibility to the collected information, a file exported in JSON format containing the metadata about an experiment, is published alongside the openly available datasets on our data archive [3]. Hereby, the reusability of our datasets is enhanced and users can make use of the datasets full potential (see Fig. 4). An example of the newly provided metadata information can be found *e.g.* for the experiment “Forward propagation of a push through a row of five people” [10].

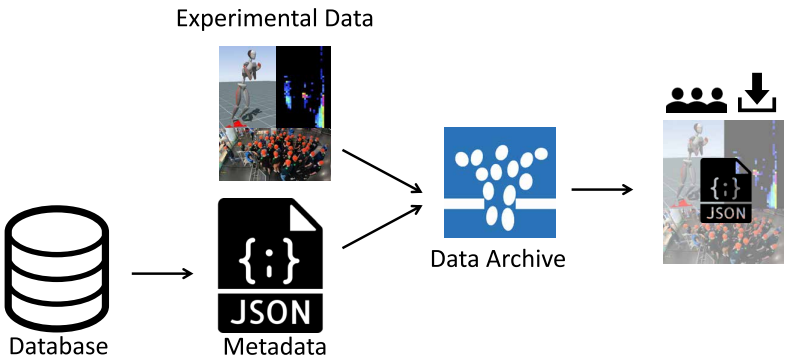

Figure 4 Experimental data and its corresponding metadata published side-by-side on the pedestrian dynamics data archive.

3 Data format

Another important aspect to make the data more FAIR are uniform data formats. Thus for two of the main data types standardized data structures have been developed or determined. Both formats base on established data formats.

The unification enables a better data exchange and an easier development of shared tools within the community, *e.g.* the library PedPy [5] already supports the new data formats.

3.1 Trajectory format

The focus was laid on developing a new standard data structure to store trajectories of people's head movement in a uniform and consistent way. Even though the developed structure is based on an available standard, an option was implemented to allow the user to enrich the trajectories with information about a persons static (e.g. height, gender, age) as well as further dynamic information (e.g. shoulder rotation, step sequence).

For the deployment of the new trajectory format, the established HDF5 format [11] was chosen, while a fixed structure serves the purpose of unification, which improves the FAIRness of the data. HDF5 is a recognized, widely supported and open source file format for storing large scientific data and it supports grouping of different data types in one file and the embedment of metadata.

The fixed structure within the HDF5 file for the trajectory format consists of the main group *trajectory*, which is complemented by the *personal details* and the *other data* groups (see Fig. 5). Hereby, the *other data* group stands for any additional data groups that might be of interest alongside the trajectory data (*e.g.* motion capturing, heart rate or electrodermal activity data). The trajectory group contains the fundamental structure of *id, frame, x, y, z* and can be complemented with other entries of dynamic information like the shoulder rotation at time *frame*. Since the HDF5 format enables each dataset to have its own describing metadata, the information supplied by the metadata schema can be added to the dataset directly and therefore, accessed by the users without further ado. In addition to that, additional metadata can be singled out and added to the highest level of the file, exemplary the file version, data origination and geometry. Examples on how to access the data as well as metadata in the HDF5 files can be found on the information page of the data archive [9].

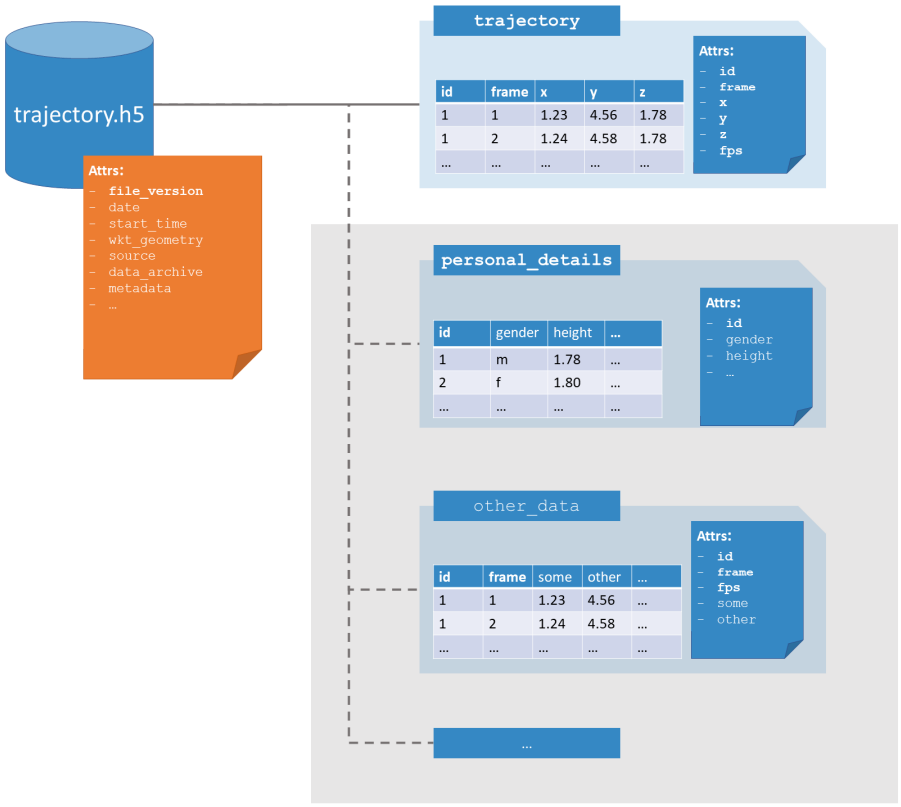

Figure 5 Structure of the novel HDF5 trajectory format.

3.2 Geometry format

Since the geometry of the experimental setup (*e.g.* the position of barriers or walls) is often used for the analysis of trajectory data, it is an important data source when considering reusability and interoperability of data. With a standardized format describing the geometric setup an additional step towards data FAIRness is done.

The variety of formats describing the geometric representation of pedestrian facilities is large (*e.g.* DXF, DWG). These formats and corresponding software are often very complex and the generation cost and time consuming and thus for a diverse research community too laborious.

Most analysis of pedestrian dynamics is done in 2D and all walls or obstacles can be approximated by a polygon. Thus, the polygon object of the conventional well-known text (WKT) format [12] is chosen to describe the experimental setup, which can easily be handled by the library shapely [13]. The setup can now be described as a polygon, a non-zero area, that can be complemented by negative spaces, representing any present barriers or obstacles. The geometry must be provided in the same global coordinate system as all other spatial data such as the trajectories.

In the future, when considering, *e.g.* stairs or varying wall heights, the format has to be enhanced or changed, but with this unification the transition to a new format can be automated easily.

4 Outlook

We will continue improving the FAIRness of our data and to push forward the idea of open science within our community. The dashed lines in Fig. 1 point out the idea of further improvement of the usibility by replicating data to other metadata repositories and allowing direct data retrieval by other software via database queries. The metadata information of the experiments shall be synchronized to other repositories like DataCite and WikiData and enriched with identifier pointing directly to the data so that the linked data can straightly be accessed by external software.

Acknowledgements We thank Tobias Schrödter for developing the data formats described in this paper. The work has been financed as a special project of the Helmholtz Metadata Collaboration (HMC) of the Helmholtz Association of German Research Centres (HGF) within the framework of the Research Data Management (RDM) Challenge of Forschungszentrum Jülich (FZJ).

Ethics Statement All ethical aspects were taken into account.

Author Contributions Maik Boltes: project administration, conceptualization, writing - original draft / Alica Kandler: investigation, software, data curation, writing - original draft.

References

[1] Shi, X., Ye, Z., Shiwakoti, N., Grembek, O.: A state-of-the-art review on empirical data collection for external governed pedestrians complex movement (2018). doi:10.1155/2018/1063043

[2] Feng, Y., Duives, D., Daamen, W., Hoogendoorn, S.: Data collection methods for studying pedestrian behaviour: A systematic review. Building and Environment **187**, 107329 (2021). doi:10.1016/j.buildenv.2020.107329

[3] Forschungszentrum Jülich, Institute for Advanced Simulation: Data archive of experiments on pedestrian dynamics. doi:10.34735/ped.da

[4] Wilkinson, M. D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**(1), 160018 (2016). doi:10.1038/sdata.2016.18

[5] Schrödter, T.: PedPy - Pedestrian Trajectory Analyzer. Forschungszentrum Jülich (2022). doi:10.5281/zenodo.10814490

[6] International Organization for Standardization (ISO): Records Management Standard 15489-1:2016. URL https://www.iso.org/obp/ui/#iso:std:iso:15489:-1:ed-2:v1:en. Accessed: 2023-10-31

[7] Research Organization Registry (ROR). URL https://ror.org

[8] Open Researcher and Contributor ID (ORCID). URL https://orcid.org

[9] Forschungszentrum Jülich, Institute for Advanced Simulation: Information about Trajectory and Metadata File Format. URL https://ped.fz-juelich.de/da/dataFormat

[10] Feldmann, S., Adrian, J., Boomers, A.K., Boltes, M., Čamernik, J., Ernst, M., Kandler, A., Lügering, H., Schrödter, T., Seyfried, A., Sieben, A.: Forward propagation of a push through a row of people (2022). doi:10.34735/ped.2022.2

[11] The HDF Group: Hierarchical Data Format, version 5 (1997). URL https://www.hdfgroup.org/HDF5/

[12] International Organization for Standardization (ISO): Information technology – database languages – sql multimedia and application packages – part 3: Spatial 13249-3:2016. URL https://www.iso.org/standard/60343.html. Accessed: 2023-10-31

[13] Gillies, S.: Shapely: Manipulation and analysis of geometric objects (Polygon) (2007). URL https://shapely.readthedocs.io/en/stable/reference/shapely.Polygon.html