FAIR Digital Object Application Case for Composing Machine Learning Training Data

Nicolas Blumenröhr[‡], Thomas Jejkal[‡], Andreas Pfeil[‡], Rainer Stotzka[‡]

‡ Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author: Nicolas Blumenröhr (nicolas.blumenroehr@kit.edu)

Abstract

The application case for implementing and using the FAIR Digital Object (FAIR DO) concept (Schultes and Wittenburg 2019), aims to simplify the access to label information for composing Machine Learning (ML) (Awad and Khanna 2015) training data.

Data sets curated by different domain experts usually have non-identical label terms. This prevents images with similar labels from being easily assigned to the same category. Therefore, using them collectively for application as training data in ML comes with the cost of laborious relabeling. The data needs to be machine-interpretable and -actionable to automate this process. This is enabled by applying the FAIR DO concept. A FAIR DO is a representation of scientific data and requires at least a globally unique Persistent Identifier (PID) (Schultes and Wittenburg 2019), mandatory metadata, and a digital object type.

Storing typed information in the PID record demands a prior selection of that information. This includes mandatory metadata and a digital object type to enable machine interpretability and subsequent actionability. The information provided in the PID record refers to its PID Kernel Information Profile (PIDKIP), defined or selected by the creator of the FAIR DO. A PIDKIP is a standard that facilitates the definition and validation of the mandatory metadata attributes in the PID record. This information acts as a basis for a machine to decide if the digital object is reusable for a particular application. Part of that is also the digital object type, which enables a machine to work with the data represented by the FAIR DO. If more information is required, the data itself or other associated FAIR DOs need to be accessed through references in the PID record.

Specifying the granularity of the data representation, and the granularity of the metadata in the information record is not a fixed task but depends on the objective. Here, the FAIR DO concept is used for representing image data sets with their label metadata. Each data set contains multiple images, which refer to the same label term. One data set associated with a particular label is represented as one FAIR DO. A type that provides information about this entity covers the packaged format of the images and the image format itself. Further information about the label term and other metadata associated with the

data set is provided or accessed through references in the PID record. For the PIDKIP, the Helmholtz KIP was chosen, following the RDA Working Group recommendations on PID Kernel Information (RDA 2013). This profile includes mandatory metadata attributes, used for machine-actionable decisions required for relabeling. Information about the data labels is not directly provided in its PID record, but in another PID record of an associated image label FAIR DO. This one represents a metadata document, containing label information about the data set. Its PID record is based on the same PIDKIP, i.e. the Helmholtz KIP. Both FAIR DOs point to each other. Thus, the image label FAIR DO is accessed via the reference in the PID record of the data set FAIR DO and vice versa. Its PID record contains information about the labels, which are relevant to the relabeling task. Accessing data label information that way means the user does not have to look up each data set, analyze its content and search for its labels. (Fig. 1)

The automated procedure for relabeling then looks as follows: A specialized client that can work with PIDs, resolves the PID of a FAIR DO which represents an image data set, and fetches its record. Analyzing its type, the client validates the data usability for composing a ML training data set. Furthermore, the referenced PID of the image label FAIR DO in the record is resolved the same way. By analyzing its PID record, the client identifies that it is relevant for getting information about the labels. The document represented by the image label FAIR DO is accessed via its location path provided in the PID record. To work with its content, a specialized tool is required that is compatible with its format and schema, i.e. its type. This tool identifies and analyzes the label term of the data set for mapping it to corresponding label terms of other image data sets.

This specification of FAIR DOs enables the relabeling of entire image data sets for application in ML. However, the current granularity of data representation is insufficient for other machine-based decisions and actions on single images. Another aspect in this regard is to increase the information in the PID record to enable more machine-actionable decisions. This requires reconsideration of the granularity of metadata in the PID record and needs to be balanced with the aim of fast record processing. Changing the content of the PID record also leads to deriving a new PIDKIP, or extending existing ones. Metadata tools applied in conjunction with the FAIR DO concept that uses the label information in the document of the metadata FAIR DOs need further specification. One requirement for their implementation is a standardized data description for the metadata document, using schemas and vocabularies.

Using the machine actionability of FAIR DOs described above, enables automation for relabeling data sets. This leaves more time for the ML user to concentrate on model training and optimization. Software development of FAIR DO-specific clients and metadata mapping tools are the subject of current research. The next step is to implement such software, for carrying out the proposed concept on a large scale.

This work has been supported by the research program 'Engineering Digital Futures' of the Helmholtz Association of German Research Centers and the Helmholtz Metadata Collaboration Platform (Helmholtz-Gemeinschaft Deutscher Forschungszentren 1995).

Keywords

Persistent Identifier, Metadata, Image Data, Label

Presenting author

Nicolas Blumenröhr

Presented at

First International Conference on FAIR Digital Objects, presentation

Conflicts of interest

References

- Awad M, Khanna R (2015) Machine Learning. Efficient Learning Machines. https://doi.org/10.1007/978-1-4302-5990-9_1
- Helmholtz-Gemeinschaft Deutscher Forschungszentren (1995) Helmholtz Metadata Collaboration (HMC) Platform. https://www.helmholtz.de/forschung/challenges/information-data-science/helmholtz-metadata-collaboration-plattform-hmc/. Accessed on: 2022-7-09.
- RDA (2013) PID Kernel Information Profile Management. https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg. Accessed on: 2022-7-09.
- Schultes E, Wittenburg P (2019) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. Data Analytics and Management in Data Intensive Domains (pp.3-16). https://doi.org/10.1007/978-3-030-23584-0_1

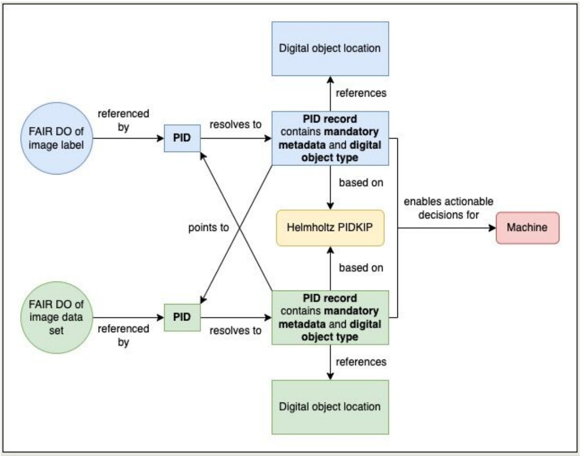

Figure 1.

Concept of how the FAIR DOs of an image data set and its label provide machine-actionable information for relabeling. Copyright information: CC BY 4.0.