From implementation to application: FAIR digital objects for training data composition

Nicolas Blumenröhr‡, Rossella Aversa‡

‡ Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author: Rossella Aversa (rossella.aversa@kit.edu)

Academic editor: Francisco Andres Rivera Quiroz

Abstract

Composing training data for Machine Learning applications can be laborious and time-consuming when done manually. The use of FAIR Digital Objects, in which the data is machine-interpretable and -actionable, makes it possible to automate and simplify this task. As an application case, we represented labeled Scanning Electron Microscopy images from different sources as FAIR Digital Objects to compose a training data set. In addition to some existing services included in our implementation (the Typed-PID Maker, the Handle Registry, and the ePIC Data Type Registry), we developed a Python client to automate the relabeling task. Our work provides a Proof-of-Concept validation for the usefulness of FAIR Digital Objects on a specific task, facilitating further developments and future extensions to other machine learning applications.

Keywords

FAIR Digital Objects, Metadata Schemas, Vocabularies, Linked Data, Operations, Machine Learning

Introduction

In Machine Learning (ML), representative training sets (e.g. Aversa et al. 2018b) are usually composed by combining data from various sources. The data selection process includes data search, access, and evaluation. The data search is usually carried out in databases, platforms, and repositories, which may follow different data management standards mainly because of the different purposes (e.g. institutional repositories versus discipline-specific databases). A user-friendly information system to ease the search of domain repositories and databases is the NFDI4Ing Data Collections Explorer (KIT Data Manager 2022), while the Metadata Standards Catalog (HMC 2021) provides a continuously extensible list of existing metadata standards and schemas.

Once suitable data sets are found and accessed, they need to be further analyzed by the scientist in order to evaluate their usability for training a ML model on the chosen task. In particular, when dealing with supervised learning, additional preprocessing is required, e.g. the assignment of labels in image recognition methods. As the data is collected from different sources, these labels need to be aligned, e.g. according to semantically similar categories; this way, images can be grouped together and relabeled. After a sufficient amount of images has been collected and relabeled, the composed training data set can be further prepared for ML using other techniques, e.g. resizing, or rescaling. The previously described preprocessing steps may be laborious and time-consuming, as they are usually performed manually by scientists, preventing them from spending their valuable time on the actual ML task, i.e., model training and analysis of results.

A possible solution to overcome the heterogeneity of data and repositories, and to reduce the amount of manually performed actions, is to apply the FAIR principles (GO-FAIR 2018) to the data to enable its machine-actionability. This can be realized with the use of FAIR Digital Objects (FDOs) (FAIR Digital Objects Forum 2022). A FDO is a digital representation of data as a sequence of bits, identified by a globally-unique, persistent, and resolvable Identifier (PID), described by an information record and classified by a type that determines the operations a machine can perform on the data. The FDOs may contain in their information record a reference to the repository where the data is deposited. This approach enables repository-agnostic (meta)data interpretation on a common basis for machines, without making any direct changes to the original data. This abstraction layer of FDOs in turn allows different data assets to be connected and related, regardless of where they are stored.

In this work we present the application of FDOs to Scanning Electron Microscopy (SEM) images labeled with a term related to their content. In this context, we refer to metadata as either administrative (bundled in the information record and necessary to manage the data, e.g. to identify its format) or scientific (information about the data in the context of a specific scientific question, e.g. labels describing the image content, which are represented by FDOs themselves). We explain the design of the FDOs and the requirements for their implementation in this context. Finally, we discuss the benefits of representing SEM image data as FDOs to compose new training data sets for further ML applications.

FDO implementations

The FDO concept, which provides a generic framework, is not directly applicable as it covers a broad scope, requires contextual interpretation, and domain-specific expertise along with decision-making. Therefore, it must be implemented into an architecture of components that enable its use. An essential aspect that has been established in all FDO implementations is the employment of repositories as trustworthy data storage.

The approach of the Research Data Alliance (RDA) (European Commission et al. 2013) community for implementing the FDO concept is built on existing systems: the PID-

Service of the Handle Registry (DONA-Stiftung 2014) is used to create, update and resolve the PID assigned to a data object in order to access it; the ePIC (European Persistent Identifier Consortium 2009) Data Type Registry (DTR) contains the Kernel Information Profiles (KIP), where the attributes of the information records are provided, defined, typed and assigned a machine-readable PID as well as a human-readable string name. Based on this structure, a minimum set of attributes (including the type of a FDO), which we call administrative metadata, required for machine-actionable decisions can be provided to clients.

Further ways to model the FDO concept have been proposed: one of them is the FAIR Digital Object Framework (FDOF). Its implementation was shown in the frame of a test case from NFDI4DS, where ML data components (e.g. images as training data, source code, and publications) were connected to each other by representing them as FDOs (Boukhers et al. 2022). Another solution is offered by the Distributed System of Scientific Collections (DiSSCo) research infrastructure, which is implementing FDO via PID records (Handle), also called FDO records, and JSON-LD representation of Linked Data. This can enable different services, such as Graph Machine learning on a biodiversity knowledge graph (Grieb et al. 2023).

Material and methods

Situation before the use of FDOs

In order to show the advantage of the FDO concept when the data is distributed in different repositories, we created two data sets starting from the NFFA-Europe – Majority SEM Dataset (Aversa et al. 2018a), which consists of JPEG images packed in a TAR archive file, recorded using the SEM measurement technique, and deposited in B2Share. The images, grouped into 10 nanoscience categories, were earlier employed to train a ML image classifier (Modarres et al. 2017). For our application case, we chose the images corresponding to the category "Biological" as our first data set. To obtain a second data set with similar image content and labels, we selected a subset of the "Biological" images, corresponding to "Neurobiology", and uploaded it to Zenodo (Blumenröhr 2023a). To avoid duplicates in the checksums of the original and the copied images, the format of the latter was converted to PNG.

Metadata curation

In the original data set (Aversa et al. 2018a), all the images belonging to a category are grouped in a folder named as the category label. However, for the purpose of our application case, the image labels need to be in a machine-interpretable format. Thus, we described each of these in a structured way for each data set using the ML-basic metadata JSON schema (Blumenröhr 2022), and we harmonized their definitions by using the UNESCO Thesaurus (UNESCO 1977). The label "Biological" of the first data set was linked to the semantically close vocabulary concept “Biology”, while the label

"Neurobiology" of the second data set was linked to the homonymous concept. The resulting JSON metadata documents were then deposited in MetaRepo (NFFA 2022), the NFFA-Europe Pilot metadata repository.

Implementation of the FDO concept

We created several FDOs to represent the data landscape, i.e., images and labels, as well as their types (Fig. 1). For the sake of clarity, we will refer to them in the text using different prefixes (i.e., image-, label-, type- FDO). We created the PIDs using the Handle Registry PID-Service. This was facilitated by a Typed PID Maker (TPM) instance (Pfeil and Jejkal 2020) that was implemented and connected to the Handle Registry and the ePIC DTR. The TPM covers validation of the FDOs information records against their KIPs during creation, and can also be used to update and retrieve them.

To fill the information record of the image- and label- FDOs, we used the Helmholtz KIP (Jejkal et al. 2022), which was developed by the community of the Helmholtz Metadata Collaboration (HMC) and refers to the RDA Kernel Type WG recommendations on KIPs (Weigel et al. 2019). It contains a set of 15 basic attributes that cover the recommended minimum, and additional information about a FDO. It is registered and defined in the ePIC DTR test instance we used for our work (Jejkal and Schweikert 2022). For our application case, we used 10 of the 15 attributes, including a reference to the location where the content is stored.

In our implementation, each image- and label- FDO has a reference (i.e., PID) to a type-FDO, which contains information such as MIME-type, related metadata schema, and version. The type- FDO, which is based on the File Type KIP (Pfeil 2022) and typed, determines the operations that can be performed on the data represented by the image- and label- FDOs. After the type- FDOs were created, their PIDs were included in the attributes of the information records for the image- and label- FDOs, together with the repository location (URI) of the data objects, and the cross-references to the related FDOs. We yielded a total of 7 PIDs, 3 for the type- FDOs (i.e., 21.11152/c7009b0f-3cbd-41fa-9e42-cecf0d1f3552, 21.11152/1bd1fab9-2d13-4710-84ff-0aed44097fee, and 21.11152/de2d965b-8941-46f1-b0f0-94e2ca41c18c) and 4 for the image- and label- FDOs (i.e., 21.11152/37833c54-1d36-42e4-858d-831447122863, 21.11152/219ca65d-a534-4f44-aec5-a98be36f9f56, 21.11152/324ef76e-25f7-4f8c-89aa-992ae6996d1a, 21.11152/b0b5de04-6e11-480b-ab66-2d4a5f42ea9e).

Implementation of a FDO client

We implemented a Python client (Blumenröhr 2023b) to exploit our machine-readable and interpretable FDOs in order to perform the relabeling task. The client is a semi-automatic tool that enables the user to dynamically select the attributes of the FDO's information record needed for a given operation, which is implemented as a class method locally. For operations that need to communicate with an API endpoint on the server-side, we employed the Python 'requests' package, which uses the HTTP protocol.

In general, the class methods use as input the selected attributes, defined in the ePIC DTR, and return the result of the processed attribute value. Some methods are FDO type-agnostic, whilst others are FDO type-specific. As an overview, the steps the client performs are:

1. receive the PID of a FDO
2. send a request to the Handle Registry using the TPM to resolve the PID and retrieve the information record
3. validate the PIDs of the record attributes against their profile using the ePIC DTR
4. in case of FDO type-specific methods, evaluate the FDO type and execute the corresponding task (Fig. 2).

In our specific use case of the relabeling task, we provided the PIDs of the image FDOs to the client, which then performed the sequence of operations shown in Fig. 3. It is worth noticing that we developed a FDO type-specific method to retrieve the label information from a JSON file: the method assesses similarities between label terms from different FDOs by reading the vocabulary definitions of the ML-basic schema (Blumenröhr 2022) and by using the related, broader, and narrower concepts. The latter defines how terms are connected to each other and can be easily processed by machines, as the UNESCO Thesaurus provides a JSON-LD representation of them.

Discussion of results

Situation with the use of FDOs

Without the FDO representation, the metadata attributes were seldom machine-readable or collected in metadata documents describing the images and the labels, which were only available in the metadata repository. Moreover, it was possible to reference the image location from the metadata repository, but not vice versa. Having represented images and labels as FDOs, their administrative metadata are machine-readable and -interpretable, and include attributes relating the corresponding data objects to each other, regardless of their location. With this representation, a client can perform operations on them without the prerequisite for any changes to the original data, contributing to an automated composition of a ML training data set of SEM images.

Relabeling ML data with FDOs

Our client exploits the FDO data representations to retrieve the SEM images and to relabel them, i.e., to assess the relations between different label terms, based on the machine interpretation of the UNESCO Thesaurus concept definitions. To successfully perform this task, the information record of the FDO must contain at least the following attributes: "checksum" as a means of verification, "type" to call the appropriate method, "license" to evaluate whether it is allowed to use the data, "topic" to access the relevance of the data for a given task before downloading it or further processing its corresponding

information record, "location" to access the SEM images and the JSON documents, "hasMetadata" to retrieve the information record of the label- FDO and "isMetadataFor" in turn to point to the image- FDO.

All attributes are part of, but not exclusive to, the Helmholtz KIP. Therefore, our client supports any FDOs, even based on other KIPs that contain the aforementioned required attributes. However, it must be noted that the relation assessment of the label terms was performed using the concept definitions from the UNESCO Thesaurus, an approach based on linked data. Metadata files that are based on other schemas and vocabularies will require additional implementations.

Conclusions

Our work successfully shows the feasibility of using FDOs in the context of an ML application case to automate the time-consuming relabeling task as part of the training data composition. It is worth remarking that our approach implements harmonized data descriptions using schemas and vocabularies to facilitate machine-readability and -interpretability. We decided to represent each data component, i.e., SEM images data sets and metadata documents containing the labels, as separate FDOs. This allows to use each FDO independently from this application case, and to link new FDOs to the already existing ones through their PIDs. The latter matches our scenario, where scientific metadata was added after data curation. The introduction of a type- FDO enables additional enrichment of FDO type-specific attributes in the information record, e.g., MIME-Type and metadata schema. Moreover, it facilitates a one-to-many relation, where all data components of the same type point to the same type- FDO, i.e., to the same PID.

The strength of our design stays in its high flexibility: the attributes in the FDO information records are standardized and reusable, being defined in a DTR; other metadata schemas and vocabulary specifications can be implemented to extend the FDO content; a client with modified features can be easily realized to perform different tasks with respect to the one presented in this work.

As a future perspective, interesting aspects can be explored: which particular attributes of the FDO information record are required to enable machine-actionable decisions in order to fulfill a given task? What is the most efficient level of granularity to represent the data in a given application case? Our FDO design can surely pave the way for further development and applications to support the answers to these questions and beyond.

Acknowledgements

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. This project has received funding from the European Union’s

Horizon 2020 research and innovation programme under grant agreement No. 101007417 within the framework of the NFFA-Europe Pilot (NEP) Joint Activities. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Aversa R, Modarres MH, Cozzini S, Ciancio R (2018a) NFFA-EUROPE - SEM Dataset. 1.0. NFFA-EUROPE Project. Release date: 2018-2-19. URL: https://b2share.eudat.eu/records/19cc2afd23e34b92b36a1dfd0113a89f
- Aversa R, Modarres MH, Cozzini S, Ciancio R, Chiusole A (2018b) The first annotated set of scanning electron microscopy images for nanoscience. Scientific Data 5 (1). https://doi.org/10.1038/sdata.2018.172
- Blumenröhr N (2022) mldata_basic_schema. 2.0.0. KIT. Release date: 2023-2-07. URL: https://metarepo.nffa.eu/api/v1/schemas/mldata_basic_schema?version=2
- Blumenröhr N (2023a) NFFA-EUROPE - Sub-SEM Dataset. 1.0.0. NFFA-EUROPE Project. Release date: 2023-2-07. URL: https://zenodo.org/record/7615060#.Y-oky-zMJGw
- Blumenröhr N (2023b) FAIR DO Client Implementation. 1.0.0. Release date: 2023-2-06. URL: https://github.com/kit-data-manager/FAIR-DO-Client-Implementation
- Boukhers Z, Beyan O, Koumpis A (2022) A test case for applying a fast track approach for FDOs in the health/data science domain. Presentation at the FDO Conference 2022 in Leiden. URL: https://docs.google.com/presentation/d/1Uoq_dN7nYIRr9lPvsPAzOpb-IcuXbE0T/edit#slide=id.p1
- DONA-Stiftung (2014) Handle.Net Registry (HNR). https://www.handle.net/. Accessed on: 2023-2-13.
- European Commission, United States Government's National Science Foundation, Australian Government's Department of Innovation, National Institute of Standards and Technology (2013) Research Data Alliance. https://rd-alliance.org/. Accessed on: 2023-2-13.
- European Persistent Identifier Consortium (2009) ePIC. https://www.pidconsortium.net/. Accessed on: 2023-2-13.
- FAIR Digital Objects Forum (2022) https://zenodo.org/communities/fdoforum/?page=1&size=20. Accessed on: 2023-7-13.
- GO-FAIR (2018) FAIR Principles. https://www.go-fair.org/fair-principles/. Accessed on: 2023-2-13.
- Grieb J, Addink W, Leeflang SIS, Weiland C (2023) Leverage FAIR machine-actionable andcrowd-sourced analysis of biodiversity data. URL: https://drive.google.com/file/d/18xiPWAqLEs6wEhbThE6yzn97NbNbAb9e/view?usp=share_link
- HMC (2021) Metadata Standards Catalog. https://msc.datamanager.kit.edu/. Accessed on: 2023-2-21.

- Jejkal T, Schweikert j (2022) HelmholtzKIP. http://dtr-test.pidconsortium.eu/#objects/21.T11148/b9b76f887845e32d29f7. Accessed on: 2023-2-13.
- Jejkal T, Pfeil A, Schweikert J, Pirogov A, Barranco PV, Krebs F, Koch C, Günther G, Curdt C, Weinelt M (2022) A basic Helmholtz Kernel Information Profile for machine-actionable FAIR Digital Objects. Karlsruher Institut für Technologie (KIT) https://doi.org/10.5445/ir/1000151434
- KIT Data Manager (2022) NFDI4Ing Data Collections Explorer. https://rxp.datamanager.kit.edu/. Accessed on: 2023-2-13.
- Modarres MH, Aversa R, Cozzini S, Ciancio R, Leto A, Brandino GP (2017) Neural Network for Nanoscience Scanning Electron Microscope Image Recognition. Scientific Reports 7 (1). https://doi.org/10.1038/s41598-017-13565-z
- NFFA (2022) MetaStore Frontend for NFFA EU Pilot. https://metarepo.nffa.eu/frontend/schema-management.html. Accessed on: 2023-2-13.
- Pfeil A, Jejkal T (2020) Typed PID Maker. 1.0.2. KIT Data Manager. URL: https://github.com/kit-data-manager/pit-service
- Pfeil A (2022) File Type. http://dtr-test.pidconsortium.eu/#objects/21.T11148/2c3cafa4db3f3e1e51b3. Accessed on: 2023-2-13.
- UNESCO (1977) UNESCO Thesaurus. https://vocabularies.unesco.org/browser/thesaurus/en/. Accessed on: 2023-7-13.
- Weigel T, Plale B, Parsons M, Zhou G, Luo Y, Schwardmann U, Quick R, Hellström M, Kurakawa K (2019) Recommendation on PID Kernel Information. Research Data Alliance. https://doi.org/10.15497/rda00031

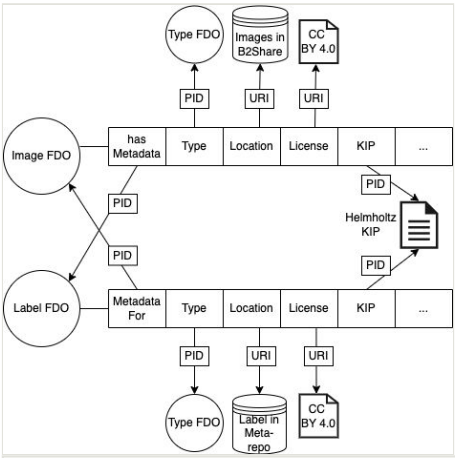

Figure 1.

The design of the FDO represents SEM images with their labels. The arrows that go from the PID record attributes point to the different objects either via URI, or PID, i.e., to the KIP, the types, the locations, the license, or another FDO.

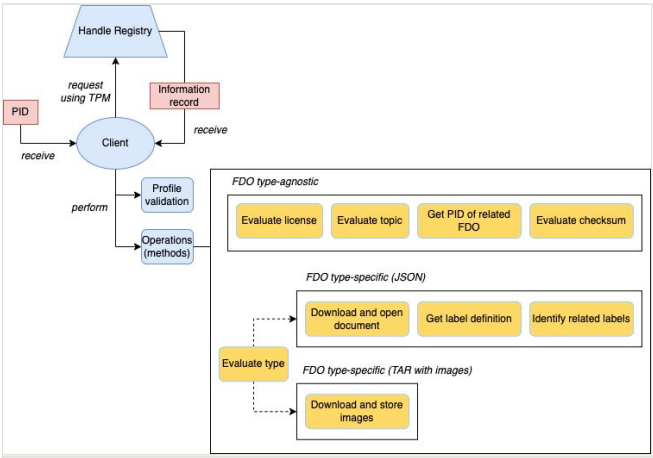

Figure 2.

The architecture of a client that performs operations on the FDO representation of the SEM images to carry out the relabeling task.

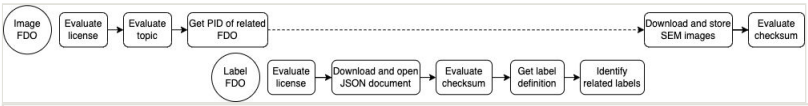

Figure 3.

The sequence (i.e., workflow) in which the operations shown in Fig. 2 are performed by the client. For simplicity, we only show the scenario where the first PID provided to the client corresponds to the image FDO.