Karlsruhe Institute of Technology

Steinbuch Centre for Computing

HELMHOLTZ METADATA COLLABORATION

FAIR Digital Object Concept for Composing Machine Learning Training Data

Nicolas Blumenröhr, Andreas Pfeil, Thomas Jejkal, Rainer Stotzka

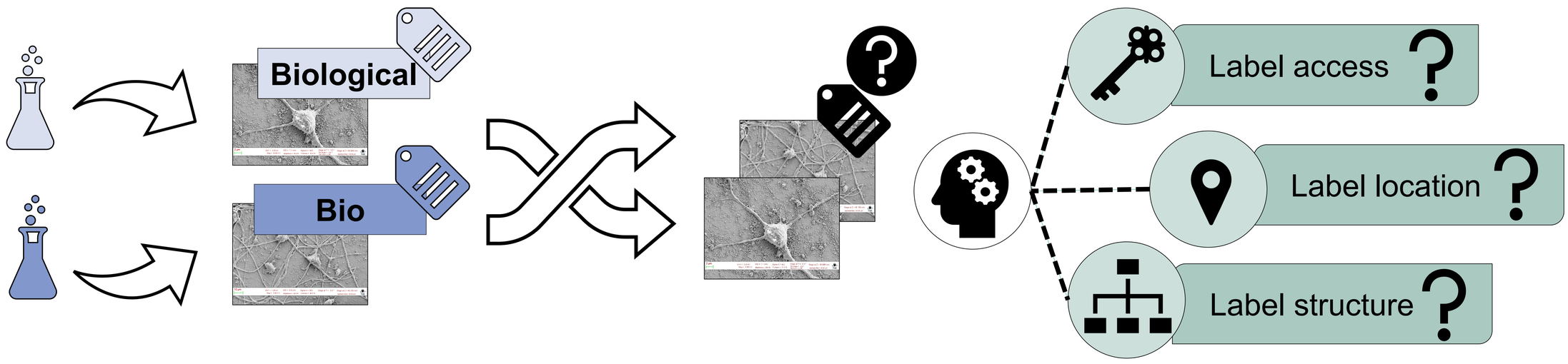
Composing Machine Learning (ML) training data sets from heterogeneous sources is laborious due to their relabelling into uniform categories.

To automate this task, the FAIR Digital Object (FAIR DO) concept can be used.

Relabelling data the classic way

Very time-consuming, because research data information needs to be searched manually:

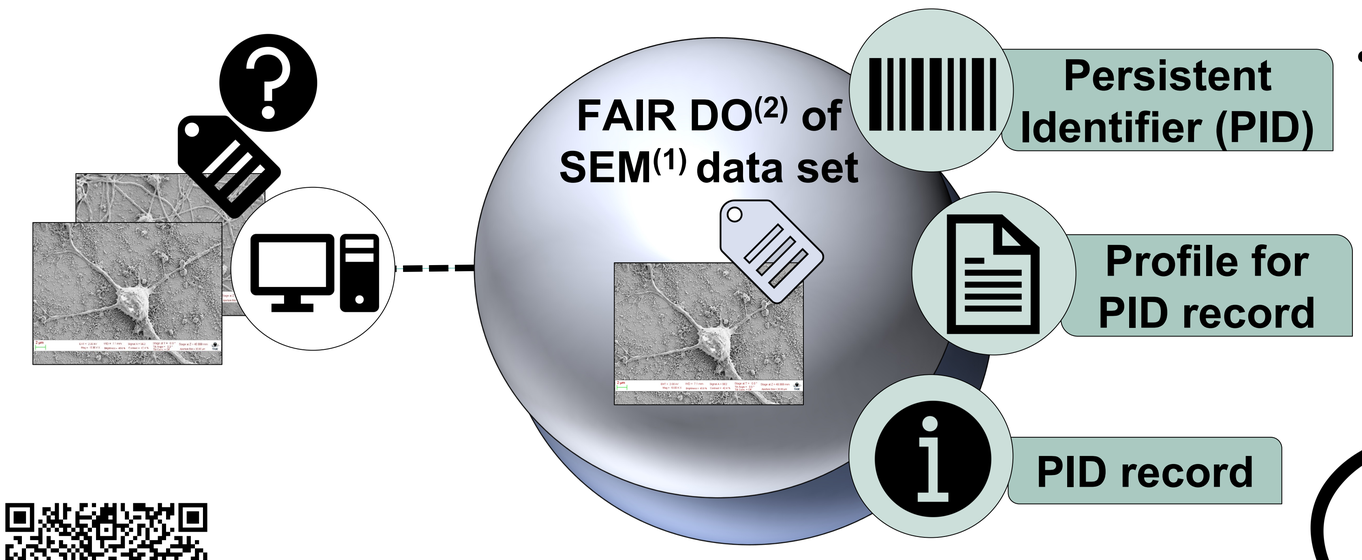
Different Laboratories create SEM(1) data sets, labelled heterogeneously. → Composing training data set for ML. → Laborious manual data relabelling.



Relabelling data represented as FAIR DO

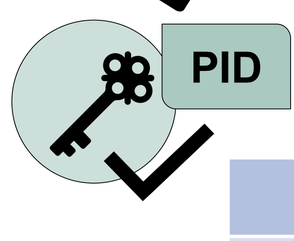
Is a representation of research data information, enabling time sparing actionability by computers in aspects of FAIR:

Composing ML training data from SEM(1) data sets represented as FAIR DOs. → Data relabelling using PID- and type-based computer actionability for FAIR DOs.



PID record of data set DO	
Profile attribute	Value
DO location	URL to SEM(1) data set
DO type	PID of data set type
DO metadata	**PID of label metadata DO**
…	…

Look up the full FAIR DO PID record at:
https://hdl.handle.net/21.11152/37833c54-1d36-42e4-858d-831447122863



PID record of label metadata DO	
Profile attribute	Value
DO location	**URL to label document**
DO type	**PID of label metadata type**
…	…

Conclusions

- If laboratories would represent their data as FAIR DOs, associated label information could be located and accessed easier.
- Clients and additional tools that are compatible with FAIR DOs can be used to support automated relabelling and other data preparation steps.
- This saves a lot of time for the ML user.

(1) **Scanning Electron Microscopy (SEM)** data set, provided by R. Aversa et. al. http://doi.org/10.23728/b2share.19cc2afd23e34b92b36a1dfd0113a89f
(2) Introduction to PIDs and FAIR DOs: https://kit-data-manager.github.io/fairdo-cookbook/about.html

Contact and feedback: nicolas.blumenroehr@kit.edu