Using Schema-based Metadata for Image Labels Accessed with FAIR Digital Objects
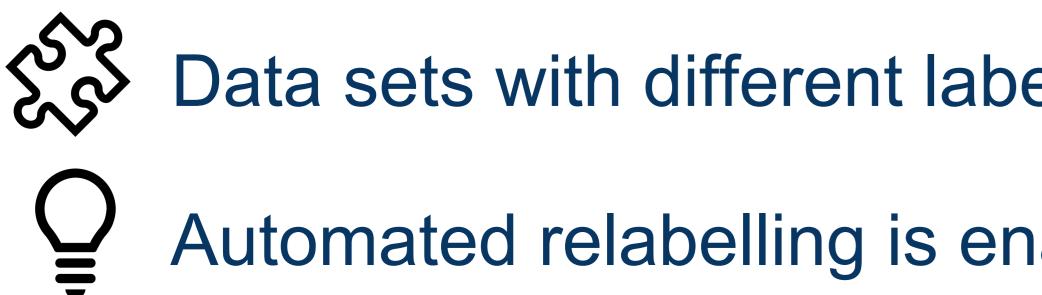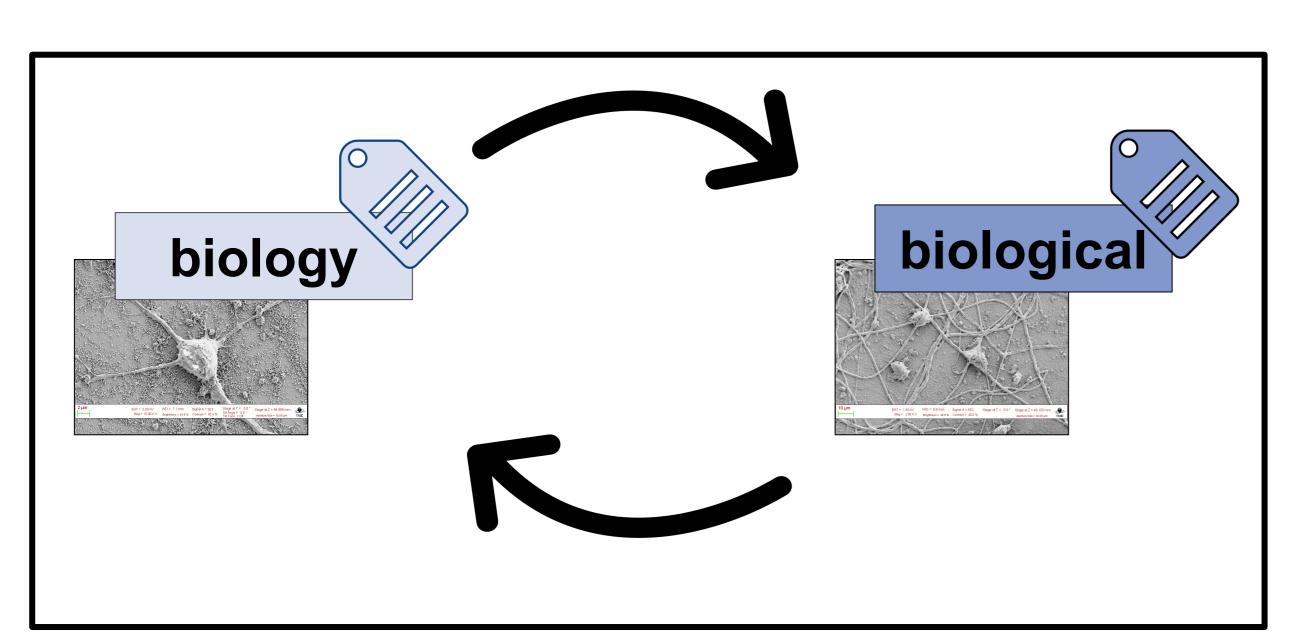
Nicolas Blumenröhr, Rossella Aversa
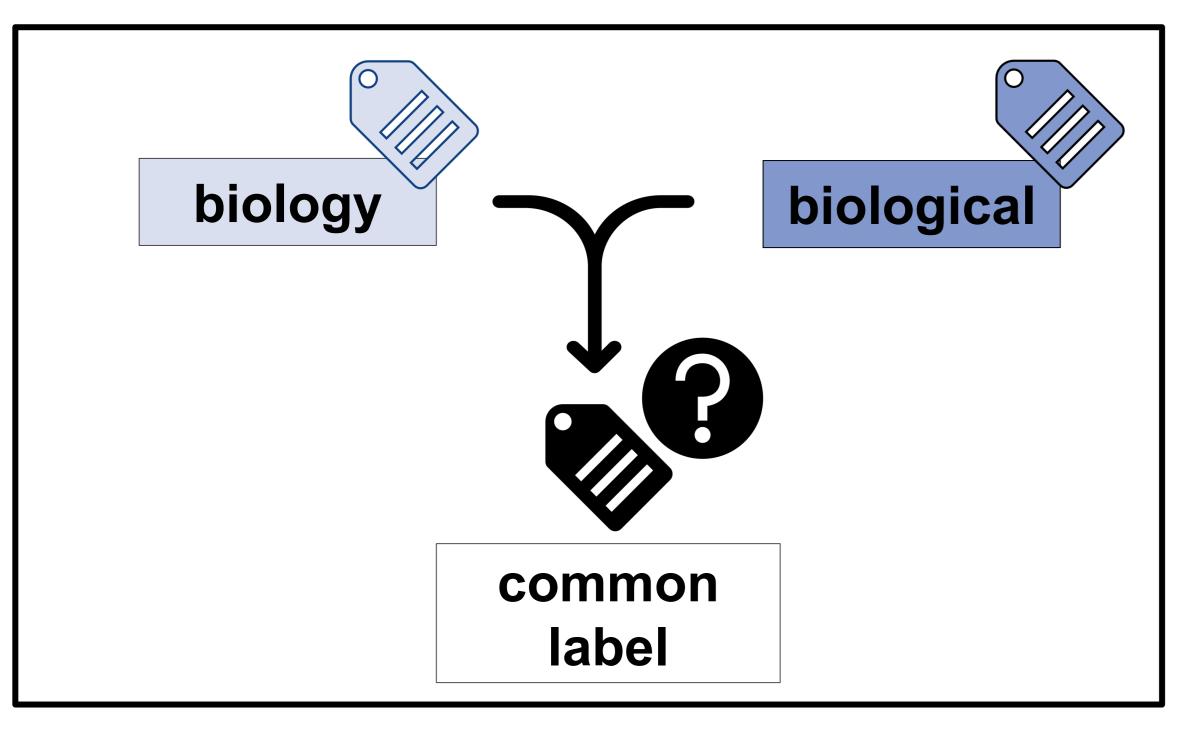
HMC HELMHOLTZ METADATA COLLABORATION

KIT Karlsruhe Institute of Technology

Steinbuch Centre for Computing (SCC)
Data Exploitation Methods (DEM)

Data sets with different label terms are not directly reusable for Machine Learning (ML) applications.

Automated relabelling is enabled by using a vocabulary-based schema and FAIR Digital Objects (FAIR DOs).

Usually different label terms are used to describe the same category, i.e. have the same meaning.

Example with *Scanning Electron Microscopy (SEM) image data. Provided by R. Aversa et. al. http://doi.org/10.23728/b2share.19cc2afd23e34b92b36a1dfd0113a89f
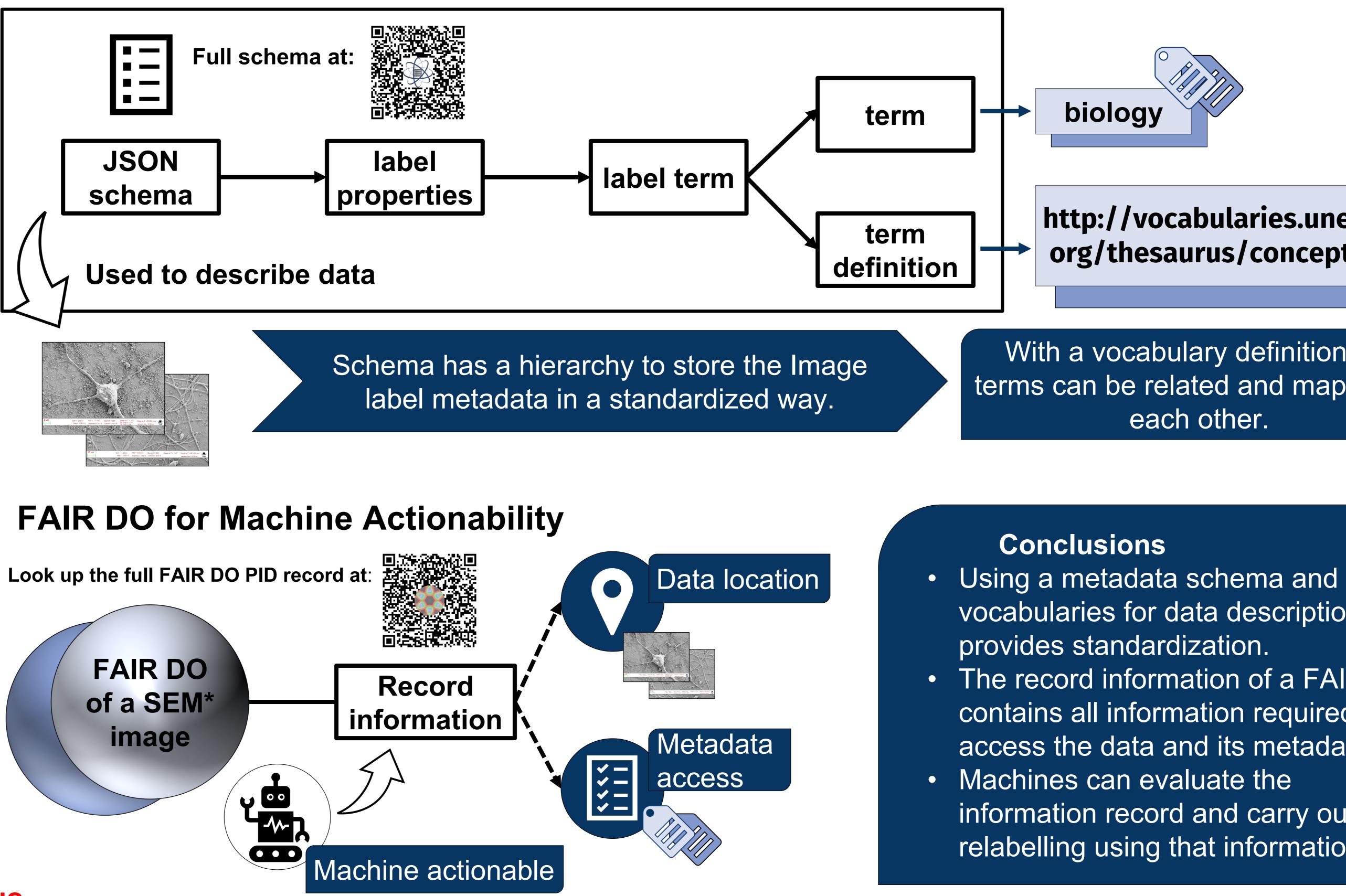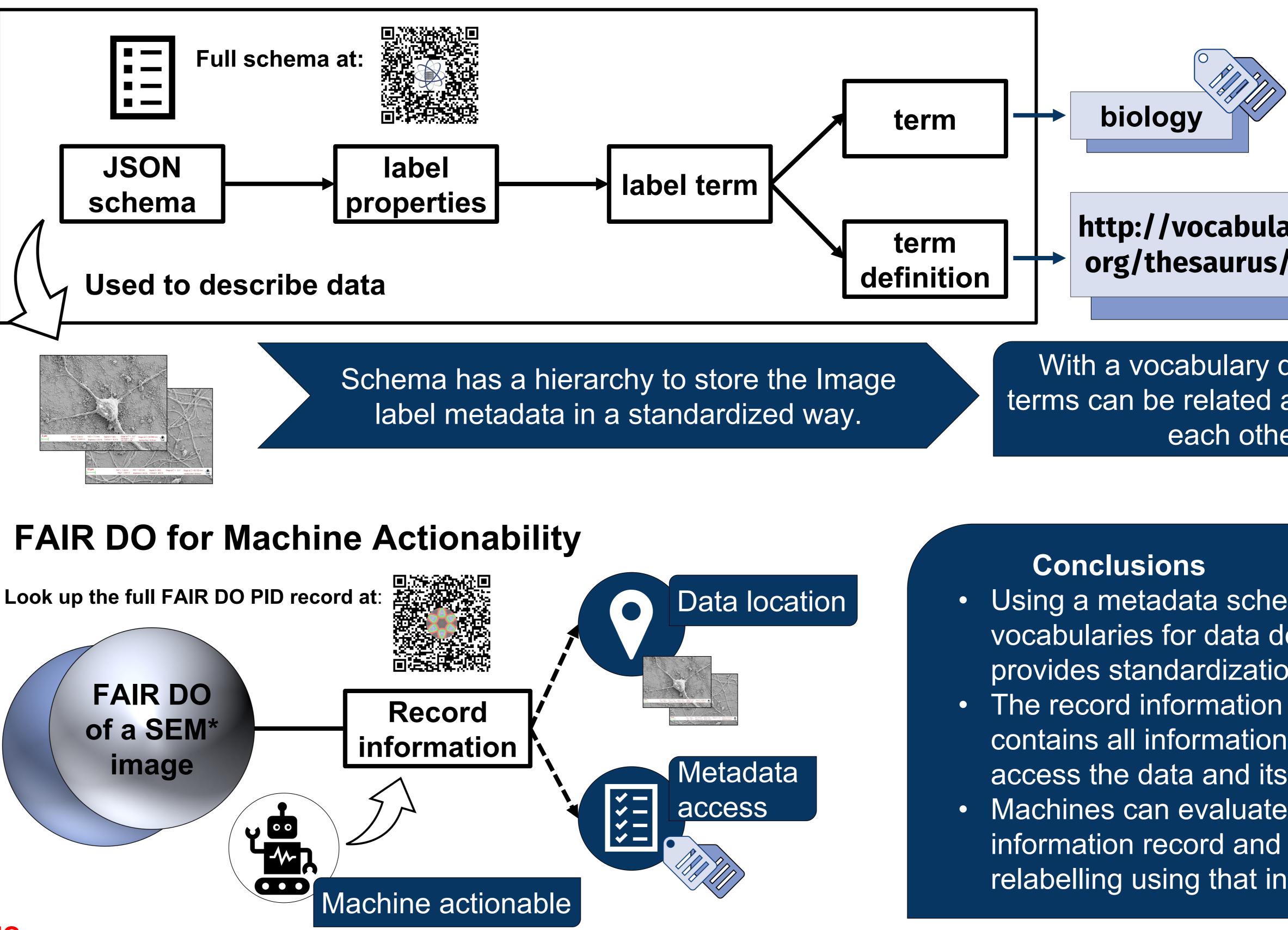
Labels can be mapped to each other, i.e. relabelled.

How to evaluate if and how the labels need to be relabelled?

Metadata Schema for Label Information

Full schema at:

JSON schema → label properties → label term → term / term definition → biology / http://vocabularies.unesco.org/thesaurus/concept219

Used to describe data

Schema has a hierarchy to store the Image label metadata in a standardized way.

With a vocabulary definition the terms can be related and mapped to each other.

FAIR DO for Machine Actionability

Look up the full FAIR DO PID record at:

FAIR DO of a SEM* image → Record information

Machine actionable

Data location

Metadata access

Conclusions
- Using a metadata schema and vocabularies for data description provides standardization.
- The record information of a FAIR DO contains all information required to access the data and its metadata.
- Machines can evaluate the information record and carry out the relabelling using that information.

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES