IMEKO TC6 International Conference on Metrology and Digital Transformation
September 19 – September 21, 2022, Berlin, Germany

PROVIDING FAIR AND METROLOGICALLY TRACEABLE DATA SETS - A CASE STUDY

Tanja Dorst [a,*], *Maximilian Gruber* [b], *Anupam P. Vedurmudi* [b], *Daniel Hutzschenreuter* [b], *Sascha Eichstädt* [b], *Andreas Schütze* [a,c]

[a]ZeMA – Zentrum für Mechatronik und Automatisierungstechnik gGmbH, Saarbrücken, Germany
[b]Physikalisch-Technische Bundesanstalt, Braunschweig/Berlin, Germany
[c]Universität des Saarlandes, Lehrstuhl für Messtechnik, Saarbrücken, Germany
[*]Corresponding author. E-mail address: tdorst@zema.de

Abstract – In recent years, data science and engineering have faced many challenges concerning the increasing amount of data. In order to ensure findability, accessability, interoperability and reusability (FAIRness) of digital resources, digital objects as a synthesis of data and metadata with persistent and unique identifiers should be used. In this context, the *FAIR data principles* formulate requirements that research data and, ideally, also industrial data should fulfill to make full use of them, particularly when Machine Learning or other data-driven methods are under consideration. In this contribution, the process of providing scientific data of an industrial testbed in a traceable and FAIR manner is documented as an example.

Keywords: data set, FAIR digital objects, traceability, digital SI, research data management

1. INTRODUCTION AND STATE OF THE ART

In 2016, the FAIR principles (and their 15 subprinciples) which provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of digital resources such as code, data sets, research objects, and workflows were published [1]. Given the widespread and increasing advancement of digitalization, the main focus of the guideline is on machine-readability, i.e. the ability of computers to find, access, (inter)operate with, and reuse data with minimal or no human intervention. The FAIR principles only describe attributes and behaviors but do not provide strict rules to achieve the desired FAIRness for digital resources, i.e., they only serve as a framework for the sustainability of data. FAIR and metrological sound data are a basis for the exchange of digital measurement values in research and industry [2]. Thousands of petabytes of data collected every year cannot be used to their full potential as the data are often not FAIR-compliant [3]. Apart from the pure FAIR framework, the interoperability of measurement data requires an unambiguous machine-readable provision of its metrological properties. Essential metadata are units of measurement, types of physical quantities, and where appropriate, traceability to measurement standards in the form of measurement uncertainty provided by calibration. In this contribution, an existing data set of a lifetime test of an electromechanical cylinder (EMC) is restructured and extended with metadata in a manner that conforms to the FAIR principles and addressing metrological properties by application of the D-SI metadata model [4].

2. USE CASE

The used test bed (cf. fig. 1) consists of an EMC as device under test (DUT) and a pneumatic cylinder, which simulates an axial load on the DUT [5]. Typically, more

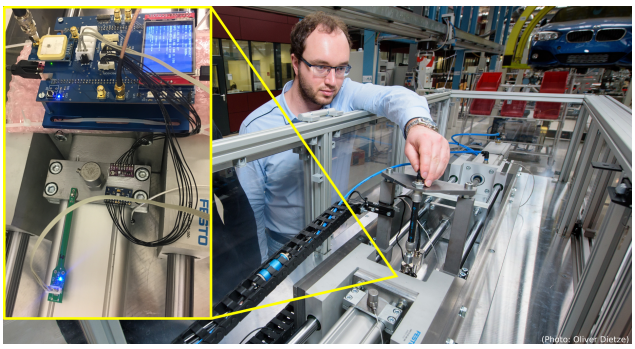

Fig. 1: SmartUp Unit (small picture) used together with the ZeMA DAQ for data acquisition of the testbed for an EMC life time test (big picture).

than 500,000 working cycles of duration 2.8 s are executed, where each consists of a forward stroke, a waiting time and a return stroke. Data are recorded with the help of two different data acquisition systems (DAQ).

The *ZeMA DAQ* acquires data of eleven different sensors during each cycle. Three motor current sensors, one microphone, three accelerometers (at the plain bearing, the ball bearing, and the piston rod), and four process sensors (axial force, velocity, pneumatic pressure, and active current of the EMC motor) with different sampling rates are used in the test bed. In addition, the *SmartUp Unit* (SUU) (cf. fig. 1) records Global Navigation Satellite System (GNSS) timestamped data continuously with three Micro-Electro-Mechanical Systems (MEMS) sensors [6]: a 9-axis inertial measurement unit, a 3-axis accelerometer, and a combined pressure and temperature sensor.

The only link between both data acquisition systems is a trigger signal of the ZeMA DAQ indicating the start of a working cycle. This trigger signal is recorded with the SUU

to enable the alignment of both data sets. The aligned data set used in this publication was acquired during a lifetime test of an EMC executed in April 2021 which lasted approx. 16.5 days. The data format of this aligned data set is suitable for the use of an automated toolbox for statistical machine learning [7]. It consists of numerical measurement values which can be associated to a corresponding unit of the "Système international d'unités" (SI) [8].

3. ACHIEVING FAIR DATA

The creation of a FAIR data set for the given use case can be broken down into different aspects, each with a specific contribution to one or more of the FAIR principles. This is also summarized by the blue (contributes) or light grey (does not contribute) symbols next to the headings of the following sections.

3.1. *Open and Known Formats*

The use of common, well-described and open (source) data formats provides the foundation of a FAIR data set [9]. Because the selected use case consists of large amounts of numerical data, the *Hierarchical Data Format Version 5* (HDF5) is chosen. It provides excellent support by allowing to structure numerical data in a filesystem-like hierarchy and to add annotations for an alignment of metadata with a description of the numerical data to every node in this hierarchy. The annotation possibilities of HDF5 are utilized by storing relatively short strings in the data-interchange format JavaScript Object Notation (JSON) which represent dictionary entries of key-value pairs corresponding to the specific node. The structure of the corresponding HDF5 file is shown in figure fig. 2. It integrates two main parts corresponding to two different measurement systems each consisting of groups representing sensors and/or physical quantities. As the ZeMA DAQ is based on sensors, each measuring only a single physical quantity, no further HDF5 group is needed in this case. In contrast, the SUU sensors each measure more than one physical quantity. Thus, a further HDF5 group is inserted, which has the name of the corresponding sensor (`BMA_280` in this example).

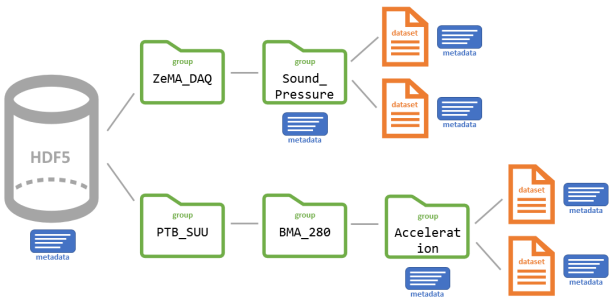

Fig. 2: HDF5 file (grey) structure containing groups (green), data sets (orange) and associated metadata (blue).

3.2. *Semantic Descriptors*

Semantically expressive concepts used to describe the metadata are an important step towards not only machine-readable but machine-interpretable data. To our knowledge, no single metadata scheme provides all the necessary concepts required for the present use case. Therefore, different ontologies and knowledge representations are used to describe specific aspects of the data set:

- Digital System of Units (D-SI) [4],
- Dublin Core (DC) [10],
- Quantity, Unit, Dimension and Type (QUDT) [11],
- Resource Description Framework (RDF) [12], and
- Sensor, Observation, Sample, and Actuator (SOSA) [13].

By using existing and semantically enriched knowledge representations as a basis, researchers, organizations, institutions, machines or algorithms are enabled to understand the data set without the expert knowledge of the initial creator. Once the meaning of the data is clear, the selection of applications for analysis and processing can become more informed.

3.3. *Top Level Metadata*

One of the most important steps preceding the (re)use of appropriate data is to find it. With the help of machine-readable top level metadata, both humans and computers can easily find the digital resource.

The top level metadata are split into four groups with several subproperties. Where appropriate, identifiers from the Dublin Core ontology are used and specified by the `dc:` prefix. In the group "Project", general information is stored about the project in which the data set was generated. The creators of the data set are listed in the group "Person". In the "Publication" group, the DOI and the license can be found among other information. To assign the data set to a unique EMC test, the "Experiment" group provides information about the specific test. The following list is not complete, but provides suggestions for the properties of the four different groups as used in the EMC data set:

- Project: fullTitle, acronym, websiteLink, fundingSource, fundingAdministrator, funding programme, fundingNumber, acknowledgementText
- Person: dc:author, e-mail, affiliation
- Publication: dc:identifier, dc:license, dc:title, dc:type, dc:description, dc:subject, dc:SizeOrDuration, dc:issued, dc:bibliographicCitation
- Experiment: date, DUT, identifier, label

Top level metadata in JSON are added as shown in list. 1. In this contribution, only the most important top level metadata are shown. The complete list of all top level metadata for this EMC data set is included in the published data set itself [14].

3.4. *Publication*

For publication, several aspects need to be considered.

3.4.1. *Publishing License*

The selection of a suitable publishing license enables the reuse of existing data not only by the creators, but also other researchers. In general, the license defines who can reuse the data for which purposes and how the usage should be reported. Guidance for such a decision is given, e.g., in [15]. For the EMC data set, a *Creative Commons Attribution 4.0 International (CC-BY-4.0)* license is chosen to maximize reusability, as it places no restrictions on any entities using the data as long as the original creators are credited.

```
"Project": {
  "fullTitle": "Metrology for the Factory of the Future",
  "funding programme": "EMPIR",
  "fundingNumber": "17IND12"
},
"Person": {
  "dc:author": ["Tanja Dorst", "Maximilian Gruber", "Anupam Prasad Vedurmudi"],
  "e-mail": ["t.dorst@zema.de", "maximilian.gruber@ptb.de","anupam.vedurmudi@ptb.de"],
  "affiliation": ["ZeMA gGmbH", "Physikalisch-Technische Bundesanstalt", "Physikalisch-Technische Bundesanstalt"]
},
"Publication": {
  "dc:identifier": "10.5281/zenodo.5185953",
  "dc:license": "Creative Commons Attribution 4.0 International (CC-BY-4.0)",
  "dc:title": "Sensor data set of one electromechanical cylinder at ZeMA testbed (ZeMA DAQ and Smart-Up Unit)",
  "dc:subject": ["measurement uncertainty", "sensor network", "MEMS"],
  "dc:SizeOrDuration": "24 sensors, 4776 cycles and 2000 datapoints each"
},
"Experiment": {
  "date": "2021-03-29/2021-04-15",
  "DUT": "Festo ESBF cylinder",
  "identifier": "axis11"
}
```

List. 1: JSON code for the most important top level metadata.

3.4.2. *Online Access Repository*

The finalized data set is published at the online service Zenodo [16], which specializes in Open Science. Other platforms with a similar scope also exist, e.g., PTB-OAR [17]. Only publication in suitable archives with searchable metadata ensure that others are able to find and access the created data set.

3.4.3. *Persistent Identifier*

To make the data set addressable under a globally unique and persistent identifier, a Digital Object Identifier (DOI) is generated within the Zenodo service [16]. DOI is a standard (ISO 26324:2012) widely used in the scientific community to resolve an identifier to the current storage (web)site [18]. Even if the data is no more available online, the DOI still resolves to some information about the data set and its repository.

3.4.4. *Example Access Code*

Alongside the publication at an online service, basic scripts, for the EMC data set in Python and MATLAB®, are provided to facilitate the file opening.

3.5. *Quantity, Unit and Sensor Metadata*

At its core, the data set provides numerical measurement data from different sensors. To foster a clear understanding of the numerical data, it is required to add metadata with suitable machine-readable descriptions of the underlying quantities, units and sensors. Each group of actual measurement data (the "leaf-folders" of fig. 2) is associated with its own specific metadata. In list. 2 the JSON representation of this metadata is shown for two exemplary quantities of this data set.

```
"/ZeMA_DAQ/Sound_Pressure": {
  "sosa:madeBySensor": "G.R.A.S.46BE",
  "rdf:type": "qudt:Quantity",
  "si:unit": "\\pascal",
  "qudt:hasQuantityKind": "qudt:SoundPressure",
  "qudt:value": {
    "si:label": "Sound pressure",
    "misc": {
      "raw_data": False,
      "comment": "Converted from ADC values based on appropriate conversion."
    },
  },
  "qudt:standardUncertainty": {
    "si:label": "Sound pressure uncertainty"
  }
},
"/PTB_SUU/BMA_280/Acceleration": {
  "sosa:madeBySensor": "BMA 280",
  "rdf:type": "qudt:Quantity",
  "si:unit": "\\metre\\second\\tothe{-2}",
  "qudt:hasQuantityKind": ["qudt:Acceleration", "qudt:Acceleration", "qudt:Acceleration"],
  "qudt:value": {
    "si:label": ["X acceleration", "Y acceleration", "Z acceleration"]
  },
  "qudt:standardUncertainty": {
    "si:label": ["X acceleration uncertainty", "Y acceleration uncertainty", "Z acceleration uncertainty"]
  },
  "misc": {"interpolation_scheme": "cubic"}
}
```

List. 2: JSON code for the metadata of the sound pressure sensor G.R.A.S.46 BE of the ZeMA DAQ and the 3-axis accelerometer BMA 280 of the SUU.

According to the recommendations from the D-SI metadata model `si:unit` introduces a mandatory machine-readable definition of the unit of measurement based on the SI system. Elements from the QUDT ontology add

information and semantics describing the measured data (e.g., `qudt:value`), and the type of associated measurment uncertainty (`qudt:standardUncertainty`). SOSA provides a relation to the sensors creating the data (`sosa:madeBySensor`).

3.6. *Traceable Measurement Data* FAIR

The actual numerical measurement data for a single quantity are provided by two multidimensional arrays of equal dimensions. One array `qudt:value` stores the measured values in the unit specified by the metadata. Another array `qudt:standardUncertainty` stores the standard uncertainty of the measured value in the same unit. The uncertainty information can be derived from datasheets or from calibration measurements. Both arrays are named in the same way as their corresponding metadata entries by design. The number of rows are the number of cycles, the number of columns are the number of datapoints and multiple pages (up to three) correspond to different directions (X, Y, and Z).

4. DISCUSSION OF THE FAIRNESS LEVEL

The data set is evaluated according to the FAIR data maturity model [19]. All indicators belonging to the four main FAIR data principles are rated and the mean values for each main principle are shown in fig. 3. The evaluation of all indicators already shows a good agreement, however, certain aspects should still be further improved. It should be kept in mind that even if the maximum value for all indicators is reached, it is not guaranteed to represent meaningful data.

5. CONCLUSION

In this contribution, the extension of an existing data set for an EMC lifetime test with metadata according to the FAIR principles has been demonstrated. The data set including all metadata can be downloaded at Zenodo (`https://zenodo.org/record/5185953`) in the HDF5 file format.

ACKNOWLEDGMENTS

Research for this paper has received funding within the projects 17IND12 Met4FoF and 17IND02 SmartCom from the EMPIR program co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation program and from the Federal Ministry of Education and Research (BMBF) project FAMOUS (01IS18078).

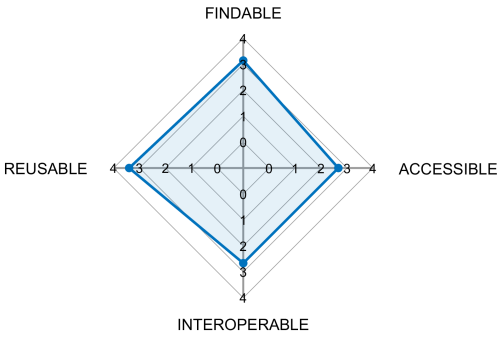

Fig. 3: FAIR data maturity model.

REFERENCES

[1] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

[2] CIPM Task Group on the "Digital-SI". [Online; accessed March 11, 2022]. URL: https://www.bipm.org/documents/20126/46590079/WIP+Grand_Vision_v3.4.pdf/aaeccfe3-0abf-1aaf-ea05-25bf1fb2819f.

[3] S. Stall et al. "Make scientific data FAIR". In: *Nature* 570 (2019), pp. 27–29. DOI: 10.1038/d41586-019-01720-7.

[4] D. Hutzschenreuter et al. *SmartCom Digital System of Units (D-SI)*. Tech. rep. July 2020. DOI: 10.5281/zenodo.3816686.

[5] N. Helwig. "Zustandsbewertung industrieller Prozesse mittels multivariater Sensordatenanalyse am Beispiel hydraulischer und elektromechanischer Antriebssysteme". PhD thesis. Dept. Systems Engineering, Saarland University, Saarbruecken, 2018. DOI: 10.22028/D291-27896.

[6] S. Eichstädt et al. "Toward Smart Traceability for Digital Sensors and the Industrial Internet of Things". In: *Sensors* 21.6 (2021). ISSN: 1424-8220. DOI: 10.3390/s21062019.

[7] T. Dorst et al. "Automated ML Toolbox for Cyclic Sensor Data". In: *Mathematical and Statistical Methods for Metrology* (2021).

[8] BIPM. *Le Système international d'unités (SI)*. 9th ed. 2019. URL: https://www.bipm.org/documents/20126/41483022/SI-Brochure-9.pdf.

[9] P. Groth et al. "FAIR Data Reuse - The Path through Data Citation". In: *Data Intelligence* 2.1-2 (2020), pp. 78–86. ISSN: 2641-435X. DOI: 10.1162/dint_a_00030.

[10] DCMI Usage Board. *DCMI Metadata Terms*. DCMI Recommendation. Dublin Core Metadata Initiative, 2020. URL: http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/.

[11] QUDT.org. *QUDT CATALOG - Quantities, Units, Dimensions and Data Types Ontologies*. Tech. rep. 2022. URL: http://www.qudt.org/2.1/catalog/qudt-catalog.html.

[12] W3C. *RDF Schema 1.1*. Tech. rep. 2014. URL: https://www.w3.org/TR/rdf-schema/.

[13] W3C. *Semantic Sensor Network Ontology*. Tech. rep. 2021. URL: https://www.w3.org/TR/vocab-ssn/.

[14] T. Dorst et al. "Sensor data set of one electromechanical cylinder at ZeMA testbed (ZeMA DAQ and Smart-Up Unit)". In: (2021). DOI: 10.5281/zenodo.5185953.

[15] GitHub Inc. [Online; accessed March 4, 2022]. URL: https://choosealicense.com/.

[16] CERN Data Centre & Invenio. [Online; accessed March 11, 2022]. URL: https://zenodo.org/.

[17] Physikalisch-Technische Bundesanstalt. [Online; accessed March 11, 2022]. URL: https://oar.ptb.de/.

[18] J. Liu. "Digital Object Identifier (DOI) and DOI Services: An Overview". In: *Libri* 71.4 (2021), pp. 349–360. DOI: 10.1515/libri-2020-0018.

[19] C. Bahim et al. "The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments". In: *Data Science Journal* 19.1 (2020), pp. 1–7. DOI: 10.5334/dsj-2020-041.