ACTA IMEKO
ISSN: 2221-870X
March 2023, Volume 12, Number 1, 1 - 6

A case study on providing FAIR and metrologically traceable data sets

Tanja Dorst[1], Maximilian Gruber[2], Anupam P. Vedurmudi[2], Daniel Hutzschenreuter[3], Sascha Eichstädt[2], Andreas Schütze[1]

[1] *Lab for Measurement Technology, Saarland University, Campus A5 1, 66123 Saarbrücken, Germany*
[2] *Physikalisch-Technische Bundesanstalt, Abbestraße 2-12, 10587 Berlin, Germany*
[3] *Physikalisch-Technische Bundesanstalt, Bundesallee 100, 38116 Braunschweig, Germany*

ABSTRACT
In recent years, data science and engineering have faced many challenges concerning the increasing amount of data. In order to ensure findability, accessibility, 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.

Section: RESEARCH PAPER

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

Citation: Tanja Dorst, Maximilian Gruber, Anupam P. Vedurmudi, Daniel Hutzschenreuter, Sascha Eichstädt, Andreas Schütze, A case study on providing FAIR and metrologically traceable data sets, Acta IMEKO, vol. 12, no. 1, article 5, March 2023, identifier: IMEKO-ACTA-12 (2023)-01-05

Section Editor: Daniel Hutzschenreuter, PTB, Germany

Received November 17, 2022; **In final form** February 14, 2023; **Published** March 2023

Funding: 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).

Corresponding author: Tanja Dorst, e-mail: t.dorst@lmt.uni-saarland.de

1. INTRODUCTION

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]. Providing data in a FAIR way contributes to the data curation process, which allows to maintain a high value of data over extended periods of time [2]. 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; however, they 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 metrologically sound data are a basis for the exchange of digital measurement values in research and industry [3]. Thousands of petabytes of data collected every year cannot be used to their full potential as the data are often not FAIR-compliant [4]. 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 addresses metrological properties by application of the D-SI metadata model [5].

2. USE CASE

The used test bed (see Figure 1) consists of an EMC as *Device Under Test* (DUT) and a pneumatic cylinder, which simulates an

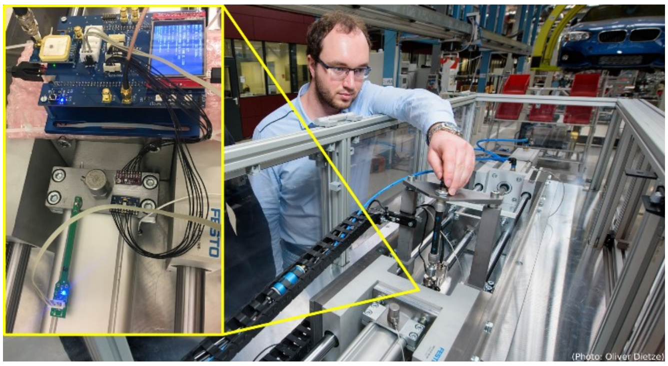

Figure 2. SmartUp Unit (small picture) used together with the ZeMA DAQ for data acquisition of the testbed for an EMC life time test (big picture).

axial load on the DUT [6]. Typically, more 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 from 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) (see Figure 1) records *Global Navigation Satellite System* (GNSS) timestamped data continuously with three *Micro-Electro-Mechanical Systems* (MEMS) sensors [7]: a 9-axis inertial measurement unit, a 3-axis accelerometer (for comparison reasons [8]), 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 [9]. It consists of numerical measurement values which can be associated to a corresponding unit of the "Système international d'unités" (SI) [10].

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. The justification of this summary is given by the detailed evaluation provided in Table 1 and further discussed in Section 4.

3.1. Open and known formats

F A I R

The use of common, well-described, and open (source) data formats provides the foundation of a FAIR data set [11]. Because the selected use case consists of large amounts of numerical data, the *Hierarchical Data Format Version 5* (HDF5) is chosen. It supports the structuring of numerical data in a filesystem-like hierarchy and allows the addition of 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 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).

Open and known formats do not contribute to the findability of the data set, as this choice only concerns the structure of the data, but not the "external visibility".

3.2. Semantic descriptors

F A I R

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:

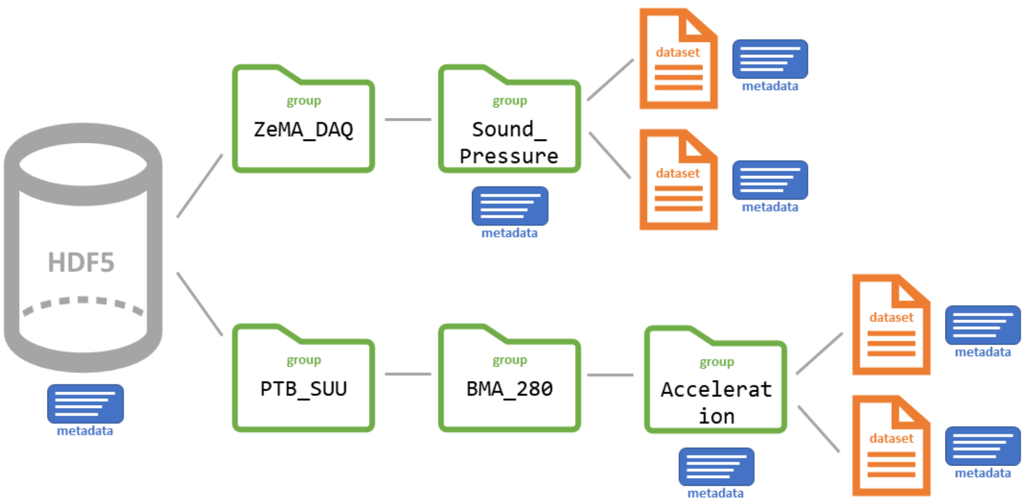

Figure 1. HDF5 file (grey) structure containing groups (green), data sets (orange) and associated metadata (blue).

- Digital System of Units (D-SI) [5],
- Dublin Core (DC) [12],
- Quantity, Unit, Dimension and Type (QUDT) [13],
- Resource Description Framework (RDF) [14], and
- Sensor, Observation, Sample, and Actuator (SOSA) [15].

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.

Semantic descriptors do not contribute to the findability of the data set, as they are only used inside the data set to describe the different data, hence not contributing to the "external visibility".

3.3. Top level metadata

One of the most important steps preceding the (re)use of appropriate data is to find them. 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 [12] 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, acknowledgementText, funding programme, fundingNumber
- 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 Listing 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 [16].

Top level metadata is primarily used to find and categorize the data set on a rough level.

3.4. Publication

In general, several aspects need to be considered: 1) choosing a suitable publishing license and an online access repository; 2) making the data set findable under a persistent identifier, and 3) providing example code to access and reuse the data set.

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 by 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 [17]. 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.

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

```
"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"
}
```

3.4.2. Online Access Repository

The finalized data set is published at the online service Zenodo [18], which specializes in Open Science. Other platforms with a similar scope also exist, e.g., the "Open Access Repository of the Physikalisch-Technische Bundesanstalt" (PTB-OAR) [19]. Only publication in suitable archives with searchable metadata ensures 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 [18]. DOI is a standard (ISO 26324:2012) widely used in the scientific community to resolve an identifier to the current storage (web)site [20]. 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 Figure 1 is associated with its own specific metadata. In Listing 2, the JSON representation of this metadata is shown for two exemplary quantities of this data set.

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

Listing 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.

```
"/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"}
}
```

the type of associated measurement uncertainty (`qudt:standardUncertainty`). SOSA provides a relation to the sensors creating the data (`sosa:madeBySensor`).

Quantity, unit, and sensor metadata do not contribute to the findability and accessibility of the data set, as this choice only concerns the "internal visibility".

3.6. Traceable measurement data

F A I R

Metrological traceability is defined as the relation of a measurement value to a reference through a chain of calibrations [21]. In practice, this is achieved by quantifying the measurement uncertainty of this measurement value. 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 is the number of cycles, the number of columns is the number of data points, and multiple pages (up to three) correspond to different directions (X, Y, and Z).

Traceable measurement data do not contribute to the findability of the data set, as this choice only concerns the measurement data, but not the "external visibility".

4. DISCUSSION OF THE FAIRNESS LEVEL

The data set is evaluated according to the FAIR data maturity model [22]. All indicators belonging to the four main FAIR data principles are qualitatively rated from 0 (not considered) to 4 (fully implemented) as proposed by [23], and the outcome is shown in Figure 3. A more detailed mapping of specific sub-indicators to the subsections of Section 3 is provided in Table 1. The evaluation of all indicators already shows a good compliance

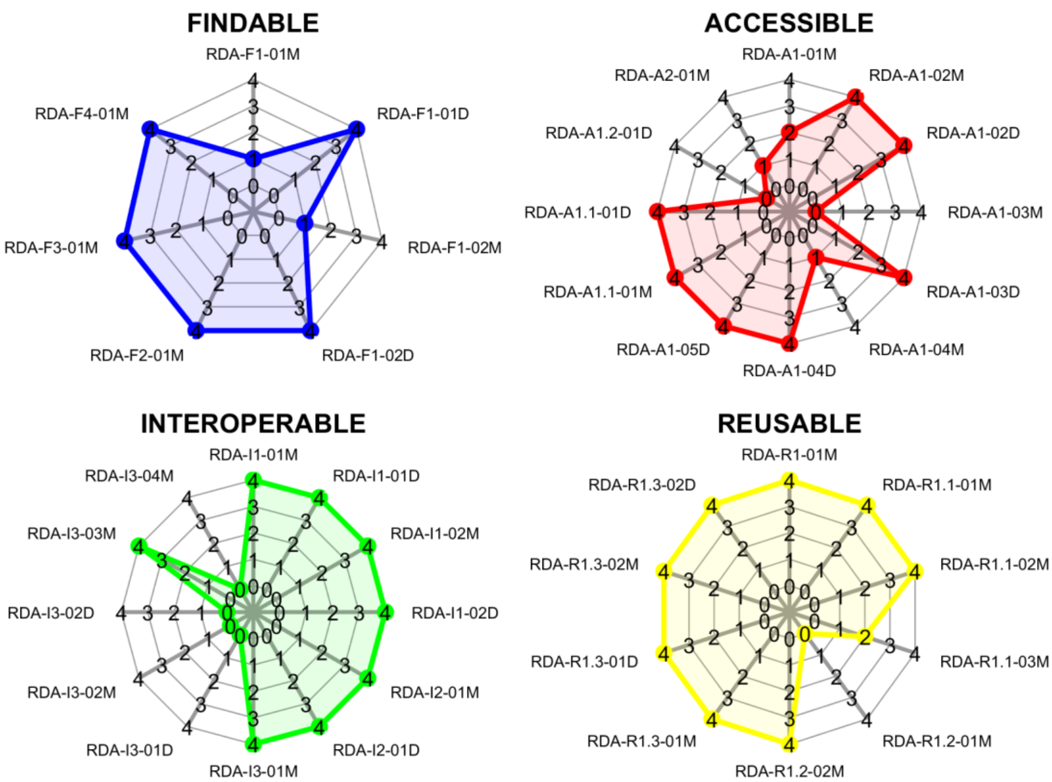

Figure 3. Detailed rating of the sub-indicators of the FAIR data maturity model.

Table 1. FAIR data maturity model indicators evaluated by the subsections of Section 3 in this paper. A checkmark indicates that the specific subindicator is fulfilled by a specific aspect. The different aspects identified as relevant for the data set are explained in Section 3.

		Open and Known Formats	Semantic Descriptors	Top Level Metadata	Publication	Quantity, Unit and Sensor Metadata	Traceable Measurement Data
F1	RDA-F1-01M						
F1	RDA-F1-01D				✓		
F1	RDA-F1-02M						
F1	RDA-F1-02D			✓	✓		
F2	RDA-F2-01M			✓			
F3	RDA-F3-01M			✓			
F4	RDA-F4-01M			✓	✓		
A1	RDA-A1-01M				✓		
A1	RDA-A1-02M	✓					
A1	RDA-A1-02D	✓					✓
A1	RDA-A1-03M						
A1	RDA-A1-03D				✓		
A1	RDA-A1-04M						
A1	RDA-A1-04D	✓			✓		✓
A1	RDA-A1-05D	✓			✓		✓
A1.1	RDA-A1.1-01M	✓	✓		✓		
A1.1	RDA-A1.1-01D	✓			✓		✓
A1.2	RDA-A1.2-01D						
A2	RDA-A2-01M						
I1	RDA-I1-01M		✓			✓	
I1	RDA-I1-01D	✓					✓
I1	RDA-I1-02M		✓			✓	
I1	RDA-I1-02D	✓					✓
I2	RDA-I2-01M		✓			✓	
I2	RDA-I2-01D						✓
I3	RDA-I3-01M		✓			✓	
I3	RDA-I3-01D						
I3	RDA-I3-02M						
I3	RDA-I3-02D						
I3	RDA-I3-03M		✓			✓	
I3	RDA-I3-04M						
R1	RDA-R1-01M		✓			✓	✓
R1.1	RDA-R1.1-01M				✓		
R1.1	RDA-R1.1-02M				✓		
R1.1	RDA-R1.1-03M						
R1.2	RDA-R1.2-01M						
R1.2	RDA-R1.2-02M		✓				
R1.3	RDA-R1.3-01M	✓	✓			✓	
R1.3	RDA-R1.3-01D	✓					✓
R1.3	RDA-R1.3-02M	✓	✓			✓	
R1.3	RDA-R1.3-02D	✓					✓

between the data set and the FAIR data principles. However, certain aspects need further improvement. Also, it should be considered that even if the maximum value for all indicators is reached, it is not guaranteed to represent meaningful data.

5. CONCLUSIONS

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. Future research should further investigate the semantic model of the metadata structure. Moreover, the feasibility of an automated evaluation of the FAIR data maturity model is another topic of interest.

ACKNOWLEDGEMENT

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).

REFERENCES

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton (+48 more authors), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3(1) (2016) art. No. 160018.
DOI: 10.1038/sdata.2016.18

[2] S. A. Thomas, F. Brochu, A framework for traceble storage and curation of measurement data, Measurement: Sensors 18 (2021)

art. No. 100201.
DOI: 10.1016/j.measen.2021.100201
[3] CIPM Task Group on the "Digital-SI", Online [Accessed 26 March 2023] https://www.bipm.org/documents/20126/46590079/WIP+Grand_Vision_v3.4.pdf/aaeccfe3-0abf-1aaf-ea05-25bf1fb2819f
[4] S. Stall, L. Yarmey, J. Cutcher-Gershenfeld, B. Hanson, K. Lehnert, B. Nosek, M. Parsons, E. Robinson, L. Wyborn, Make scientific data FAIR, Nature 570 (2019), pp. 27–29.
DOI: 10.1038/d41586-019-01720-7
[5] D. Hutzschenreuter, F. Härtig, W. Heeren, T. Wiedenhöfer, A. Forbes (+18 more authors), SmartCom digital system of units (D-SI), Zenodo (2020).
DOI: 10.5281/zenodo.3816686
[6] 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
[7] S. Eichstädt, M. Gruber, A. P. Vedurmudi, B. Seeger, T. Bruns, G. Kok, Toward smart traceability for digital sensors and the industrial Internet of Things, Sensors 21(6) (2021).
DOI: 10.3390/s21062019
[8] B. Seeger, T. Bruns, Primary calibration of mechanical sensors with digital output for dynamic applications, Acta IMEKO 10 (2021) 3, pp. 177–184.
DOI: 10.21014/ acta_imeko.v10i3.1075
[9] T. Dorst, Y. Robin, T. Schneider, A. Schütze, Automated ML toolbox for cyclic sensor data, Proceedings of Mathematical and Statistical Methods for Metrology MSMM, Turin, Italy, 31 May – 1 Jun. 2021, pp. 149–150. Online [Accessed 26 March 2023] http://www.msmm2021.polito.it/content/download/245/1127/file/MSMM2021_Booklet_c.pdf
[10] BIPM, Le système international d'unités (si), 9th ed. 2019. Online [Accessed 26 March 2023] https://www.bipm.org/documents/20126/41483022/SI-Brochure-9.pdf
[11] P. Groth, H. Cousijn, T. Clark, C. Goble, FAIR Data Reuse - The Path through Data Citation, Data Intelligence 2(1-2) (2020), pp. 78–86.
DOI: 10.1162/dint_a_00030
[12] DCMI Metadata Terms, Dublin Core Metadata Initiative, DCMI Recommendation (2020). Online [Accessed 26 March 2023] http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/
[13] QUDT.org, QUDT CATALOG - Quantities, units, dimensions and data types ontologies (2022). Online [Accessed 26 March 2023] http://www.qudt.org/2.1/catalog/qudt-catalog.html
[14] W3C, RDF Schema 1.1 (2014). Online [Accessed 26 March 2023] https://www.w3.org/TR/rdf-schema/
[15] W3C, Semantic sensor network ontology (2021). Online [Accessed 26 March 2023] https://www.w3.org/TR/vocab-ssn/
[16] T. Dorst, M. Gruber, A. P. Vedurmudi, Sensor data set of one electromechanical cylinder at ZeMA testbed (ZeMA DAQ and Smart-Up Unit) (2021).
DOI: 10.5281/zenodo.5185953
[17] GitHub Inc., Choose an open source license. Online [Accessed 27 March 2023] https://choosealicense.com/
[18] CERN Data Centre & Invenio, Zenodo. Online [Accessed 26 March 2023] https://zenodo.org/
[19] Physikalisch-Technische Bundesanstalt, Open Access Repository of the Physikalisch-Technische Bundesanstalt. Online [Accessed 26 March 2023] https://oar. ptb.de/
[20] J. Liu, Digital Object Identifier (DOI) and DOI Services: An Overview. Libri 71 (2021) 4, pp. 349–360.
DOI: 10.1515/libri-2020-0018
[21] BIPM JCGM WG2, JCGM 200:2012 International Vocabulary of Metrology - Basic and General Concepts and Associated Terms (VIM). 2012. Online [Accessed 27 March 2023] http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf
[22] C. Bahim, C. Casorrán-Amilburu, M. Dekkers, E. Herczog, N. Loozen, K. Repanas, K. Russell, S. Stall, The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments, Data Science Journal 19(1) (2020), pp. 1–7.
DOI: 10.5334/dsj-2020-041
[23] Zenodo FAIR Data Maturity Model Working Group, FAIR Data Maturity Model. Specification and Guidelines, 2020.
DOI: 10.15497/rda00050