ing.grid preprints

RESEARCH ARTICLE

The Applicability of the Persistent Identification of Instruments (PIDINST) Metadata Schema for complex compound instruments

Sac Nicte Medina [1], Federico Guillermo Diaz Capriles [1], Christian Langenbach [1], Christian Odenthal [2], Darwin Schlenk [3], Matthias Sperl [4], Christina Pätzold [3], Witold Arndt [1]

1. Scientific Information, German Aerospace Center, Cologne.
2. Institute of Engineering Thermodynamics, German Aerospace Center, Cologne.
3. Facility of Flight Experiments, German Aerospace Center, Braunschweig.
4. Institute of Material Physics in Space, German Aerospace Center, Cologne.

Abstract. The Persistent Identification of Instruments (PIDINST) schema, developed by the Research Data Alliance (RDA), provides a standardized framework for globally unique, persistent identifiers to scientific instruments. In this paper, we explore the applicability of the PIDINST metadata schema to three experimental facilities from different research areas of the German Aerospace Center (DLR). We aim to evaluate the degree of applicability of the schema on different use cases to highlight its strengths and identifying possible areas for improvement. The methodology consisted of conducting a survey. For each facility, the instruments were described using the PIDINST schema. The value for each property was assigned a category and subsequently counted and averaged. The results of our study suggest that the PIDINST schema is around 70 % applicable to our three use cases. The remaining 30 % requires a deeper analysis due to the limitations of our method.

Date Submitted:
2025-04-14

License:
This work is licensed under CC BY 4.0

Keywords:
Scientific instruments, PIDINST, metadata schema, standardization

Data availability:

Software availability:

Corresponding Author:
Sac Nicte Medina
sac.medina@dlr.de

1 Introduction

High-quality metadata is essential for the sustainable use and reusability of data [1] generated by scientific instruments. For that reason, automated, machine-readable, and schema-compliant metadata descriptions are indispensable for the efficient and reproducible replication, reproduction, and re-use. All of which are relevant to third parties, colleagues, or even the original researchers themselves after years have passed [2]. In data-intensive fields such as engineering and natural sciences, the precise and comprehensive description of scientific instrument data through metadata is crucial for ensuring the integrity and usability of research outputs. Therefore, our project *inst.dlr* [3] aims to develop and commission a central and persistent database with accompanying services. This database serves both as a source and repository for metadata of scientific instruments and facilities connecting their measured data. Lastly, it demonstrates extensive post-use possibilities of these measured data. The first step towards this goal is to identify a general metadata schema for scientific instruments, which serves as the foundation for creating standardized and interoperable metadata across various scientific disciplines.

The Persistent Identification of Instruments[1] (PIDINST) metadata schema [4] developed by the Research Data Alliance (RDA) Working Group (WG) Persistent Identification of Instruments (RDA WG PIDINST)[2] is an excellent choice to test if one can describe scientific instruments from various fields in an effective manner. Firstly, the schema aims to provide a globally unique and unambiguous identification of scientific instruments, ensuring precise referencing. Secondly, it allows linking of data to the instruments that generated them, facilitating data provenance and contextual understanding. Moreover, the schema enhances interoperability and open data sharing by standardizing metadata across different systems and disciplines. It also improves the discoverability and visibility of instruments and their data, making it easier for researchers to find and use relevant data. Additionally, the schema supports equipment logistics and mission planning by offering detailed information about the instruments. As a community-driven solution, it is widely accepted and recognized by the RDA [5], suggesting it as an ideal option for ensuring the precise, comprehensive, and standardized description of scientific instruments and their associated data. To summarize, the purpose of this schema is to provide globally unique, persistent, and resolvable identifiers for scientific instruments. In particular, measuring instruments, defined as devices used for making scientific measurements, alone or in conjunction with one or more supplementary devices [6]. By ensuring that each instrument can be unambiguously identified across various networks and infrastructures, the schema enhances the traceability, discoverability, and interoperability of instrument-related data. Therefore, the PIDINST schema is also a FAIR [1] implementable way to persistently identify measuring instruments and contextualize the data gathered by them.

The PIDINST metadata schema is the results of an empirical and iterative approach, among RDA WG PIDINST members and interested stakeholders. It was developed by first collecting use cases and then identifying commonly defined metadata properties through a schema that was iterated to obtain community feedback. The PIDINST metadata schema [5] includes essential properties grouped in the following 13 categories: *identification*, *schema version*, *landing page*, *name*, *owner*, *manufacturer*, *model*, *description*, *instrument type*, *measurable variable*, *date*, *related identifier*, and *alternative identifier*. Nearly every category contains more properties, resulting in a total of 33 properties listed in the table 1. These properties facilitate the linking of related resources to compile extensive information about a given instrument. One particular characteristic of the PIDINST schema is that properties are built on commonly used characteristics across 15 collected use cases, however, the majority (approx. 60 %) of them are related to Earth Sciences according to [5]. Also, just a few properties can be considered common because in only five use cases (50 %) they were. Therefore, testing its applicability in more disciplines is still necessary to demonstrate its practical viability for general instrument description.

In this paper, we present an applicability test of the PIDINST metadata schema on three distinct scientific experiments across three research infrastructures within the German Aerospace Center(DLR)[3]. We provide a detailed description of the scientific experiments in section 2 to expand the coverage of disciplines. In section 3 we describe the methodology that we use to evaluate

1. https://www.pidinst.org/
2. https://www.rd-alliance.org/rationale/persistent-identification-instruments-wg/rev-002/
3. https://ror.org/04bwf3e34

the applicability of the PIDINST metadata schema while the results are presented in section 4. Finally, in section 5 we give a summary of our results and highlight our main findings.

	ID	Property	Occ
1	1	Identifier	1
2	1.1	IdentifierType	1
3	2	Schema version	1
4	3	LandingPage	1
5	4	Name	1
6	5	Owner	$1 - n$
7	5.1	OwnerName	1
8	5.2	OwnerContact	$0 - 1$
9	5.3	OwnerID	$0 - 1$
10	5.3.1	OwnerIDType	1
11	6	Manufacturer	$1 - n$
12	6.1	ManufacturerName	1
13	6.2	ManufacturerID	$0 - 1$
14	6.2.1	ManufacturerIDType	1
15	7	Model	$0 - 1$
16	7.1	ModelName	1
17	7.2	ModelID	$0 - 1$
18	7.2.1	ModelIDType	1
19	8	Description	$0 - 1$
20	9	InstrumentType	$0 - n$
12	9.1	InstrumentTypeName	1
22	9.2	InstrumentTypeID	$1 - 0$
23	9.2.1	InstrumentTypeIDType	1
24	10	MeasuredVariable	$0 - n$
25	11	Date	$0 - n$
26	11.1	DateType	1
27	12	RelatedID	$0 - n$
28	12.1	RelatedIDType	1
29	12.2	RelationType	1
30	12.3	RelatedIDName	$0 - 1$
31	13	AlternateID	$0 - n$
32	13.1	AlternateIDType	1
33	13.2	AlternateIDName	$0 - 1$

Table 1: The PIDINST properties as in [5]. In the number of occurrences indicated in the column labelled as "Occ", 1 means a mandatory property that must appear once. An "Occ" value of 0–1 is an optional property that may appear at most once. 1–n is a mandatory property with potentially multiple values, i. e. the property may appear one or more times in a single record. Finally, 0–n is a multivalued optional property that may appear zero or more times in a record.

2 Pilot scientific instruments of three research areas

As mentioned previously, we applied the PIDINST metadata schema on three distinct scientific experiments to evaluate its applicability to instrument description. Although the majority of the PIDINST schema properties were earth sciences (60 %) [5] use cases, the schema was designed to be field-agnostic. Therefore, the selection of different scientific environments is to also measure adoption in other disciplines. The examined experiments (see figure 1) are described as follows:

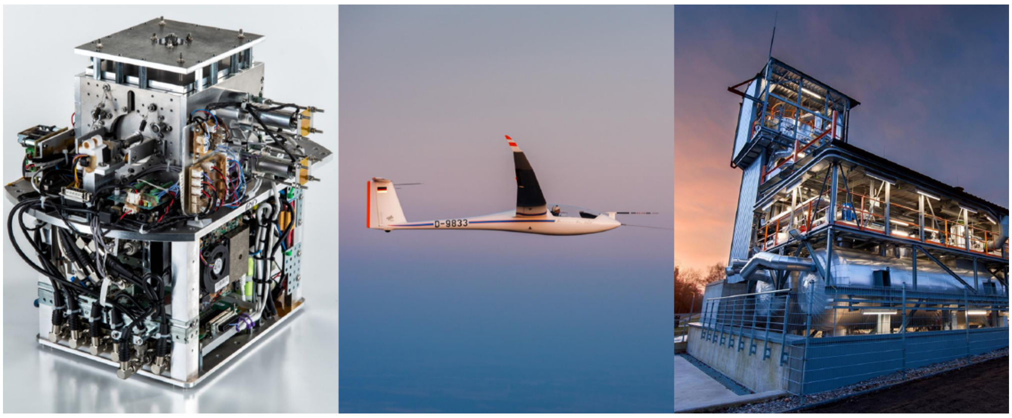

Figure 1: The three examined experiments from the German Aerospace Center. On the left the MEGraMa experiment is shown, used by the Institute for Materials Physics in Space. The next image depicts the Discus-2c research glider associated with the Flight Experiments facility. Lastly on the right, the TESIS facility is shown which is associated with the Institute of Engineering Thermodynamics. Image credits: DLR CC BY-NC-ND 3.0.

- **MEGraMa**: the orbital experiment MEGraMa (**M**agnetically **E**xcited **Gra**nular **Ma**tter) [7] is used to research so-called granular gases in low gravity environments and was developed for drop tower experiments and parabolic flights on planes or sounding rockets. The experiment has flown successfully several times on the aforementioned platforms. It was developed by the Institute for Materials Physics in Space at DLR. The MEGraMa device has five main components: Sample sphere, high speed cameras, magnets, illumination and batteries.
- **Discus-2c**: the Discus-2c[4] [8] is a research glider based on the Schempp-Hirth series model of the same name which was modified for use as a research and experimental aircraft. 48 strain gauges and 22 measuring points using fibre Bragg grating are built into wings and fuselage to determine aerodynamic loads in different flight states. The Discus-2c DLR features magnetometers and accelerometers in various locations, deflection sensors on all control surfaces, a combined Global Navigation Satellite System (GNSS) and inertial measurement unit inside the fuselage and a nose boom. The nose boom is equipped with a five-hole-probe to detect airflow angles and speeds. The experimental autopilot allows precise and accurate control of surface deflections with exact repeatability. This can be used, for example, to examine flight mechanical characteristics with high accuracy. The Discus-2c is equipped with its own data acquisition system, which continuously records sensor readings as well as the actuating variables of the experimental autopilot. This advanced sensor system is used in a wide range of areas, including aerodynamic studies, flight mechanics, aeroelasticity, measurement techniques, and human-machine interaction. It also plays a role in meeting certification requirements and supports research in digitization.
- **TESIS**: the test facility for thermal energy storage in molten salts (TESIS) [9] consists of two sub-facilities: TESIS:store and TESIS:com. Both of which have been operated as large-

4. http://s.dlr.de/XjCK8

scale process plants by the Institute of Engineering Thermodynamics without interruption (24 hours a day) since 1st January, 2019. The facilities are used for the development of new types of single-tank heat storage (TESIS:store) and the clarification of process engineering issues and qualification of salt components (TESIS:com). The instruments utilized within the TESIS facility are mainly several hundred temperature measurement devices (thermocouples and thermoresistors). Common measurements of the experiment include pressure measurement, flow measurement valve positions, filling levels and pump rotational speeds which require measurement devices.

Our analysis is produced by the utilization of the PIDINST schema on individual instruments of each experiment. From the MEGraMa experiment, the Institute of Materials Physics in Space selected a high speed camera; the Mikrotron EoSens mini1. The Flight Experiments Facility considered the aircraft of the Discus-2c experiment as one single instrument. Lastly, the Institute of Engineering Thermodynamics used an individual thyristor power flow heater stage 1 from the TESIS facility. Expanding the use case variability for the PIDINST schema enhances its adaptability to diverse scenarios, ensuring broader applicability and interoperability[5]; see for example the SCHOlarly LInk eXchange (Scholix) Framework [10]. Therefore, by proving that its properties are applicable to our three individual instruments, we are able to test the flexibility of its proposed standard capability. The following sections detail the evaluation process of the PIDINST schema performance for each instrument.

3 Methodology

To evaluate the applicability of the PIDINST metadata schema, each institute applied the schema to the instruments described in section 2. Initially, we identified the 33 schema properties (see table 1) for each instrument, where applicable. Subsequently, we examined how each institute utilized the schema to describe their instruments. Finally, we counted the properties used and calculated an average percentage to determine the schema's overall applicability.

The full process of PIDINST applicability evaluation involved the following steps:

- **Data collection**: each scientist and engineer participants filled out the properties of the PIDINST schema in a corresponding table like table 1 for the selected instrument of their experiment (see section 2).

- **Data ranking**: by asking the participants involved to use the PIDINST schema to describe their instruments, we aim to count the applicability of the properties in their instrument. We evaluated the answers by compared the information provided by the participants with the definitions given by [5] and we ranked the answers as follow:

  C Correctly used: The information was correctly provided.

  $\approx C$ Almost correctly: When only part of the information was provided correctly.

  $\notin C$ Incorrectly used: The information was incorrectly provided.

  N/A Unavailable: When there was not information provided.

5. https://docs.pidinst.org/en/latest/adoption/index.html

- **Completeness calculation**: we quantify, for each instrument, the total number of filled properties for each ranking as T_C for the properties ranked as *Correctly used*, $T_{\approx C}$ for properties ranked as *Almost correctly used*, $T_{\notin C}$ for properties ranked as *Incorrectly used*, and finally, $T_{N/A}$ as properties that are *Unavailable*.
- **Average and percentage calculation**: we computed the average value of filled properties among the three instruments. For example, for the properties ranked as *Correctly used* we used $p_C = \frac{\Sigma T_C}{n}$ to obtain the proportion of the responses, where n is the total number of entries, which in our case is 99. To obtain the corresponding percentage value, the result is multiplied by 100. We proceed similarly for the other rankings.

  In order to estimate the accuracy of our results, we computed the statistical standard error (SE) [11] for each ranking using the formula:

$$\mathrm{SE} = \sqrt{\frac{p_i(1-p_i)}{n}},$$

  where p_i are the response proportion for each ranking. Results are also multiplied by 100 to obtain the percentage value.

This method allowed us to quantitatively determine whether the schema could be applied. Therefore, providing valuable insights into its practical utility and effectiveness. The method involves merely filling in the requested information and counting the properties used. However, it does not completely account for whether the participants fully understood the schema, except for counting almost correct and incorrect answers, there is no tracking of comprehension of the schema.

4 Results

In this section, we present the results of applying the PIDINST metadata schema to three experimental instruments to test whether all the proposed properties in table 1 can be applied to the instruments described in section 2. The filled properties of each experiment are found in table 2. The column *PIDINST property* refers to the PIDINST schema properties [5], while the rest correspond to the answers of each institute. As mentioned in section 3, the relevant information is if the properties are applicable through the fact that they can be used either correctly, almost correctly, or even incorrectly. Therefore, to estimate the applicability performance we considered these three as one to indicate when a PIDINST property is applicable, and *Unavailable* to denote when it is not applicable. In table 3, the frequency of use of the properties in each instrument is listed.

The TESIS experiment filled 27 properties correctly, the Discus-2c glider had 12 properties, and the MEGraMa instrument 14 properties. Therefore, an average of 53.54 ± 5.01 % of the properties were filled out as *Correctly used*. The properties that all the three institutes applied correctly are *Name*, *Owner*, *OwnerName*, *Manufacturer*, *InstrumentType*, *RelatedID*, *RelatedIDType*, and *AlternateIDType*. A 8.08 ± 2.74 % were filled out as *Almost correctly*, being only the *Description* property consistently in 2 of the 3 instruments. Only 7.07 ± 2.58 % were filled out as *Incorrectly* and again just one property, *IdentifierType*, was the persistent cases in 2 of the 3 instruments.

PIDINST Property	TESIS	Discus-2c	MeGraMa
Identifier	C	$\notin C$	N/A
IdentifierType	C	$\notin C$	$\notin C$
Schema version	C	C	$\notin C$
LandingPage	C	C	N/A
Name	C	C	C
Owner	C	C	C
OwnerName	C	C	C
OwnerContact	N/A	$\approx C$	C
OwnerID	N/A	$\approx C$	C
OwnerIDType	N/A	$\approx C$	C
Manufacturer	C	C	C
ManufacturerName	C	N/A	C
ManufacturerID	C	N/A	$\approx C$
ManufacturerIDType	C	$\notin C$	N/A
Model	C	C	N/A
ModelName	C	N/A	N/A
ModelID	C	N/A	N/A
ModelIDType	N/A	N/A	N/A
Description	$\approx C$	$\approx C$	N/A
InstrumentType	C	C	C
InstrumentTypeName	C	N/A	C
InstrumentTypeID	C	N/A	N/A
InstrumentTypeIDType	C	N/A	N/A
MeasuredVariable	C	N/A	C
Date	$\approx C$	$\notin C$	N/A
DateType	C	N/A	$\approx C$
RelatedID	C	C	C
RelatedIDType	C	C	C
RelationType	C	N/A	N/A
RelatedIDName	C	N/A	N/A
AlternateID	C	C	$\notin C$
AlternateIDType	C	C	C
AlternateIDName	C	N/A	N/A

Table 2: Rank of the PIDINST properties answers from the three institutes as defined in section 3. For each filled property we ranked the answer as C for correctly used, $\approx C$ for almost correctly used or caused confusion, $\notin C$ for incorrectly used, and N/A when one cannot tell because it was empty/unavailable.

Ranking	TESIS	Discus-2c	MEGraMa	Average [%]
C	27	12	14	53.54 ± 5.01
$\approx C$	2	4	2	8.08 ± 2.74
$\notin C$	0	4	3	7.07 ± 2.58
N/A	4	13	14	31.31 ± 4.66

Table 3: Total counts of each category for every experiment as described in section 3. Additionally, a percent average is added to qualitatively show the amount of “confusion” for all experiments.

Therefore, approximately 70 % of the properties were applied. This level of utilization indicates significant applicability.

On the other hand, 31.31 ± 4.66 % of the properties were not filled out and therefore *Unavailable*, which indicate that they may not have been applicable. From this properties, only the *ModelIDType* is consistent in all 3 cases, while *ModelName*, *ModelID*, *InstrumentTypeID*, *InstrumentTypeIDType*, *RelationType*, *RelationIDName*, and *AlternativeIDName* appear consistent in 2 of the 3 cases. The approximately 30 % of unused entries might suggest the elements are not relevant to the specific context of the instruments or that the property descriptions lack sufficient clarity. In particular, the clarity regarding the importance and type of information required, leading to their omission. As mentioned in section 3, our response counting method does not have tracking of comprehension of the schema. We can deduce some clues from the answers *Almost correct* and *Incorrect*, as they may show a deficiency in the understanding of their meaning. However, as the properties that were evaluated in these rankings appear to be random in all 3 cases, no further information can be favourable. Consequently, as we cannot provide a definitive answer about the not applied PIDINST properties, further studies are required. For example, repeating the experiment with many more instruments and different research groups would help generalize our data for such a conclusion on user confusion. Additionally, repeating this experiment after training researchers on the schema may result in more favourable outcomes and shine a light on understanding the schema from their perspective.

Our results show that around 70 % of the properties in the schema were utilized, highlighting its potential to describe scientific instruments. Furthermore, these results seem to show that the difference in the scientific areas do not play a role for the PIDINST schema not being used because the participants did not report inability to use them. Therefore, the development of a prototype central and persistent database will use the PIDINST schema as a first approach for metadata of instruments. Interestingly, the 31.31 ± 4.66 of the properties that were not used will play a crucial role in refining and testing the next steps in the *inst.dlr* project. By understanding and addressing these limitations, we can better evaluate the potential improvements. The next phase will focus on rigorous testing and evaluation to verify the most plausible explanation for not using the 31.31 ± 4.66 properties. It is important to acknowledge that while our findings provide support, they do not constitute conclusive evidence for the transferability of the PIDINST schema to other research fields.

5 Discussion and Conclusion

In this paper, we evaluate the applicability of the PIDINST schema by analysing the properties used to describe three distinct scientific instruments. Our findings reveal that 70 % of the schema properties were utilized for these three instruments independent of the research field. This rate of utilization suggest usefulness of the schema to address general scientific instruments description. The fact that a significant majority of the properties were used suggests that the schema is aligned with the practical requirements and operational contexts of the diverse scientific instruments involved.

However, the 31.31 ± 4.66 of unused entries highlight an area for further examination. This gap may indicate that certain elements within the schema are not relevant in different scientific context of the instruments, this suggests a refinement of the PIDINST schema. One should note, it is important to assess that is not clear whether these unused elements are inherently irrelevant or if their omission is due to complexity or vagueness in the property descriptions. Ambiguities in the description of properties could lead to misunderstandings about the type and importance of the information required; resulting in these entries being disregarded. With our results, we are not able to measure the level of comprehension by the users.

Another important aspect of this work was to increase the scientific areas of testing for the PIDINST schema. Since, according to [5], it was developed with geoscience use cases in mind. Our results are independent of the scientific area, however our methodology demonstrates its limitations in obtaining further information on this possible dependency. This limitation should be considered in future studies.

Due to the significance of the applicability of the PIDINST schema, we will work on creating a central and persistent prototype database software to manage metadata describing scientific instruments in DLR. The results of this paper will support a more efficient development of our software by providing a comprehensive and clear guidance on the use of each entry. We aim to enhance the schema adaptability and comprehensiveness by ensuring that all entries are clearly defined and highlighting the significance of each element. We also plan to improve the user input to identify more precisely the possible limitations of the PIDINST schema.

6 Acknowledgements

We would like to acknowledge Dr. Peidong Yu from the Institute for Materials Physics in Space at DLR for his valuable contribution as a data provider for the description of the high speed camera Mikrotron EoSens mini1 from the MEGraMa experiment.

7 Roles and contributions

Sac Nicte Medina: Formal Analysis, , Methodology, Project administration, Writing –- original draft, Writing –- review & editing

Federico Guillermo Diaz Capriles: Software, Writing –- review & editing

Christian Langenbach: Supervision, Writing –- review & editing

Christian Odenthal: Resources, Data provider, Writing –- review & editing

Darwin Schlenk: Resources,Data provider, Writing –- review & editing

Matthias Sperl: Resources, Writing –- review & editing

Christina Pätzold: Resources, Data provider, Writing — review & editing

Witold Arndt: Conceptualisation, Funding acquisition, Writing –- review & editing

References

[1] M. D. Wilkinson et al., “The fair guiding principles for scientific data management and stewardship,” *Scientific Data Journal*, vol. 3, 2016. DOI: https://doi.org/10.1038/sdata.2016.18.

[2] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press, 2015, ISBN: 9780262327862. [Online]. Available: https://doi.org/10.7551/mitpress/9963.001.0001.

[3] S. N. Medina and F. G. Diaz Capriles, "Inst.dlr - semantic instrument database," *RDA Deutschland Tagung 2024, Potsdam, Deutschland*, 2024, Zenodo. DOI: 10.5281/zenodo.10781959. [Online]. Available: https://doi.org/10.5281/zenodo.10781960.

[4] M. Stocker et al., "Persistent identification of instruments," *Data Science Journal*, vol. 19, 2020. DOI: https://doi.org/10.5334/dsj-2020-018.

[5] R. Krahl et al., *Metadata schema for the persistent identification of instruments*, version 1.0, Mar. 2022. DOI: 10.15497/RDA00070. [Online]. Available: https://doi.org/10.15497/RDA00070.

[6] "International vocabulary of metrology. basic and general concepts and associated terms. vim 3rd edition, 2012. "[Online]. Available: https://www.bipm.org/en/publications/guides/#vim.

[7] P. Yu et al., "Magnetically excited granular matter in low gravity," *Review of Scientific Instruments*, vol. 90, no. 5, p. 054 501, May 2019, ISSN: 0034-6748. DOI: 10.1063/1.5085319. eprint: https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/1.5085319/14698025/054501_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.5085319.

[8] E. Braun, "Dlr's latest research sailplane - discus-2c dlr," *Congress Proceedings*, pp. 23–26, 2021, XXXV OSTIV Congress. [Online]. Available: https://elib.dlr.de/143095/.

[9] N. Breidenbach, C. Martin, H. Jockenhöfer, and T. Bauer, "Thermal energy storage in molten salts: Overview of novel concepts and the dlr test facility tesis," *Energy Procedia*, vol. 99, pp. 120–129, 2016, 10th International Renewable Energy Storage Conference, IRES 2016, 15-17 March 2016, Düsseldorf, Germany, ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2016.10.103. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610216310645.

[10] A. Burton et al., "The scholix framework for interoperability in data-literature information exchange," *D-Lib Magazine*, vol. 23, no. 1/2, 2017, ISSN: 1082-9873. DOI: 10.1045/january2017-burton. [Online]. Available: http://www.dlib.org/dlib/january17/burton/01burton.html.

[11] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th. W. W. Norton & Company, 2007.