Towards FAIR Data for Low Carbon Energy - Current State and Call for Action

Valeria Jana Schwanitz (✉ valerias@hvl.no)
  Western Norway University of Applied Sciences

August Wierling
  Western Norway University of Applied Sciences

Mehmet Biresselioglu
  İzmir University of Economics

Massimo Celino
  National Agency for New Technologies, Energy and Sustainable Economic Development

Muhittin Demir
  İzmir University of Economics

Maria Balazinska
  Global Mining Institute

Mariusz Kruczek
  Global Mining Institute

Manfred Paier
  Austrian Institute of Technology

Demet Suna
  Austrian Institute of Technology

Research Article

Towards FAIR data for low carbon energy - current state and call for action

Valeria Jana Schwanitz[1,†,*]**, August Wierling**[1,†]**, Mehmet Efe Biresselioglu**[2]**, Massimo Celino**[3]**, Muhittin Hakan Demir**[2]**, Maria Bałazińska**[4]**, Mariusz Kruczek**[4]**, Manfred Paier**[5]**, and Demet Suna**[5]

[1]Western Norway University of Applied Sciences, Department of Environmental Sciences, Sogndal, 6856, Norway
[2]Izmir University of Economics, Sustainable Energy Division, Izmir, 35330, Turkey
[3]ENEA, Lungotevere Thaon di Revel, 76, 00196 Roma, Italia
[4]Central Mining Institute, Katowice, 40-166, Poland
[5]AIT Austrian Institute of Technology, Vienna, 1210, Austria
[*]corresponding author(s): Valeria Jana Schwanitz (valerias@hvl.no)
[†]these authors contributed equally to this work

ABSTRACT

With the continued digitization of the energy sector, the problem of sunken scholarly data investments and forgone opportunities of harvesting existing data is exacerbating. It adds to the problem that the reproduction of knowledge is incomplete, impeding the transparency of science-based evidence for the choices made in the energy transition. We comprehensively test FAIR data practices in the energy domain with the help of automated and manual tests. We document the state-of-the art and provide insights on bottlenecks from the human and machine perspectives. We propose action items for overcoming the problem with FAIR and open energy data and suggest how to prioritize activities.

Introduction

This study is a response from the energy domain to a call for action by Wilkinson et al. 2016[15]. The authors urge 'all data producers and publishers to examine and implement ... (the FAIR) principles and actively participate with the FAIR initiative ...'. We respond to this call by documenting the state of the art on FAIR data practices in the energy research domain and suggest action items, drawing from an examination of 80 databases representative for data flows in the energy system. Our assessment follows the recommendations ('features that should be reflected') of[16] The FAIR Principles[15] have widely been acknowledged as the way forward for improving the Findability, Accessibility, Interoperability and Reusability of data across different sources and disciplines[3,7,8]. Various research communities are currently discussing and testing how to implement these guiding principles. In the energy domain, very few initiatives currently exist to progress the state-of-the-art (e.g.,[14,2,4]). Energy science faces a triple challenge. One is to meet the needs of a broad range of data stakeholders, who include researchers from social sciences to engineering, energy and other industries, policy- and decision-makers, funding and publishing agencies, and the general public. While domain experts need data at high granularity, other stakeholders require information at an aggregate level. Data needs also differ from stakeholder to stakeholder. For example, utility companies rely on high-resolution electricity demand data, while policy planners are more interested in aggregated trends. Energy researchers utilize a broad range of data, covering technical specifications to societal and environmental impacts. Notably, data are not only supporting knowledge building and validation but they are an important input to decide on pathways for the transition to a low carbon energy system. Second, data in the energy system cover large scales in time and space, respectively ranging from picoseconds to geological age (e.g., technical dispatch vs. the formation of energy resources) and nanoscale to the planning horizon of humanity (e.g., unit-level control of the electricity grid vs. long-term planning of secure access to pivotal resources in respect of planetary boundaries). Thirdly, a new type of agent beyond humans emerges in the energy system: Automatized decision- and control systems (machines) support human activity in supervising the energy infrastructure. This, in turn, requires that data need to become machine-actionable. Machine-actionability means that machines can be programmed so that they find, access, and process data without further human interaction. The implications of this third challenge lead to a new perspective on an energy system with human and machine agents at the center. The new perspective on the energy system is visualized in Fig. 1.

Three layers show how the energy system is controlled and steered by the two types of agents, enabling the provision of energy services in demand. Bidirectional flows of data provide the foundations for humans and machines to manage the
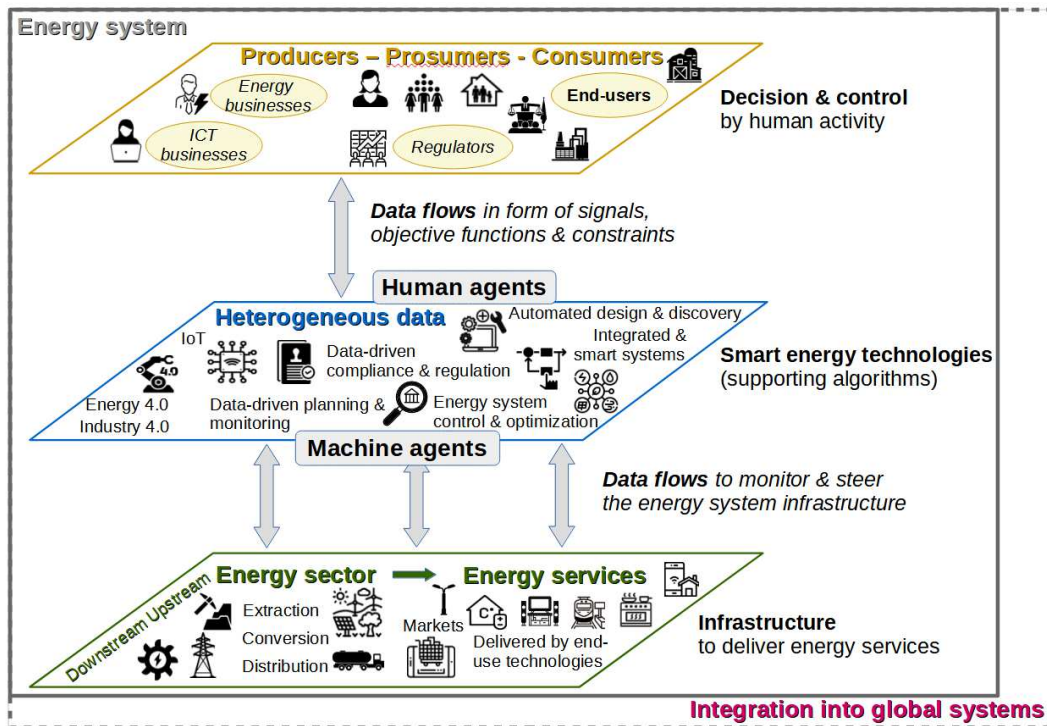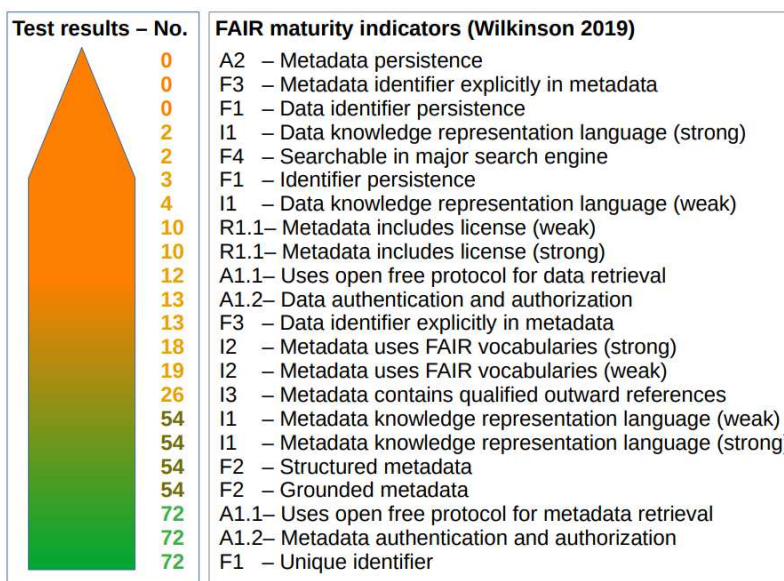
Figure 1. The energy system with human and machine agents at the center. The top layer details human actors in the energy sector, engaged in the production, distribution, and/or consumption of energy services. Their decisions and behaviors define the objectives and constraints of the energy system. This information is delivered through bilateral heterogeneous data bundles, that are taken up by smart energy technologies to monitor and steer the energy system infrastructure (bottom layer).

necessary energy infrastructure. The top layer, 'Decision & control by human activity', depicts the decision and control activities in the energy system. The classical boundaries between producers on the one hand and consumers on the other hand are thereby disappearing, making room for 'prosumers'. The top layer also emphasizes the new, significant role of ICT businesses. The middle layer complements human control of the energy system with machine control by the help of 'smart energy technologies'. The support from machines to steer the energy infrastructure (3rd layer) depends on enabling technologies such as data-driven regulatory systems with feedback and adaptive behavior, infrastructure control, flagging of alerts, real-time monitoring, data-driven compliance and regulation. The importance of data is expected to further grow in the future with the continued digitization of the energy sector, in particular with the broad introduction of smart and AI-based technologies in support of decision-making and real-time system adaptation[10, 12]. Moreover, algorithm-based strategies to identify technological solutions are increasing. An example is the automatized material selection and design without the need of expensive and/or risky experiments[13]. Consequently, market opportunities for sharing data are expected to grow tremendously, provided that the bottlenecks for doing so are removed.

The main task of machine agents is to support the infrastructure needed to deliver energy services (3rd layer). This includes the extraction and harvesting of energy resources, the conversion between different forms of energy to useful energy, the distribution of fuels, as well as the operation and maintenance of the energy equipment across time and spatial scales. Data streams flowing between the top and the middle layer are input to machines in the form of signals, objective functions, and constraints. These include taxes on energy fuels, R&D programs, energy security targets, health and sustainable development goals as well as data security and privacy requirements. The third layer exchanges data with smart energy technologies to provide the foundations for humans and machines to manage the necessary energy infrastructure.

Given the above rationales to support the sharing of energy data between layers of the energy system, the task ahead for the energy research community is to find a domain-specific way forward to implement the FAIR and open data guiding principles. The way forward depends on an agreement in the community about the costs and benefits of FAIRifying energy data. It starts with the recording of the status quo to spark discussions, which is the purpose of this paper. In line with Wilkinson et al. 2019[16], the point of the evaluation of the current level of FAIR implementation in the energy domain is to identify 'opportunities for improvements' instead of seeing scores as a goal in themselves. Fig. 1 serves as the starting point for the assessment. With its help, we choose a representative sample of 80 energy databases that cover the current and emerging energy system and

Test results – No.	FAIR maturity indicators (Wilkinson 2019)
0	A2 – Metadata persistence
0	F3 – Metadata identifier explicitly in metadata
0	F1 – Data identifier persistence
2	I1 – Data knowledge representation language (strong)
2	F4 – Searchable in major search engine
3	F1 – Identifier persistence
4	I1 – Data knowledge representation language (weak)
10	R1.1– Metadata includes license (weak)
10	R1.1– Metadata includes license (strong)
12	A1.1– Uses open free protocol for data retrieval
13	A1.2– Data authentication and authorization
13	F3 – Data identifier explicitly in metadata
18	I2 – Metadata uses FAIR vocabularies (strong)
19	I2 – Metadata uses FAIR vocabularies (weak)
26	I3 – Metadata contains qualified outward references
54	I1 – Metadata knowledge representation language (weak)
54	I1 – Metadata knowledge representation language (strong)
54	F2 – Structured metadata
54	F2 – Grounded metadata
72	A1.1– Uses open free protocol for metadata retrieval
72	A1.2– Metadata authentication and authorization
72	F1 – Unique identifier

Figure 2. Number of databases complying with FAIR maturity indicators that test 13 of 15 FAIR principles. The results are based on machine-actionability tests for 80 databases that are representative of data flows in the energy sector. None of the tested databases achieves persistence of metadata and data identifiers. Internal linking of metadata with the help of identifiers is equally problematic.

connected data flows (see the Method Section and the Supplementary Material for details).

Results

We assess a corpus of 80 databases that is representative of data flows that are pivotal for the low carbon energy transition. The selection of these databases is guided by Fig. 1. In addition, an ontology developed from the Global Energy Assessment Report (GEA 2011) was used. The proof of representativeness is described in the Supplementary Material. We assess the compliance of the selected databases with the FAIR guiding principles using an automated assessment tool[16]. Fig. 2 summarizes the results, sorting the compliance of databases with the 22 FAIR maturity indicators implemented by Wilkinson in the 'FAIR maturity evaluation service'[16]. We find that most of the energy databases allow the authentication and authorization of metadata, use open free protocol for metadata retrieval, and incorporate unique identifiers. 72 out of 80 databases are compliant, see bottom of Fig. 2. At the same time, none of the tested databases achieves persistence of metadata and data identifiers (top of Fig. 2). A general observation is also that the majority of databases poorly comply with the FAIR maturity indicators. Two thirds of the databases do not fulfill 15 out of 22 indicators. This highlights the urgency to improve the FAIR state of energy (meta)-data and stresses the general lack of machine-actionability. Without machine-actionability, opportunities in harvesting data for society will not materialize.

In addition to machine assessments, 30 of the 80 databases were also evaluated manually using the ARDC self-assessment tool[1], among others. As documented in the Supplementary Material, we detect a large spread of results in the manual assessments of the same databases by different researchers. The largest spread was found for the assessment of the interoperability of data. This underlines, first, a strong degree of subjectivity in the assessments, originating from the different disciplinary background and data governance proficiency of the analyst. Secondly, a shared understanding about what makes data FAIR is lacking. The comparison of automated and manual assessments allows us furthermore to contrast the machine- and human perspectives on FAIR evaluations in the energy domain. To this end, the original weighing of answers to the ARDC assessment questions has been transferred to the machine-actionable FAIR maturity test (refer to the Supplementary Material for the details). Fig 3 shows the stylized results of this comparison for each of the FAIR guiding principles (findability, accessibility, interoperability, and reusability). We abstain from reporting assessment scores to focus on observed gaps and the room for improvement (see[16]).

A tendency is that machine assessments score lower than manual assessments with the exception of the interoperability criteria where the results are mixed. A reason for the overall lower scoring with machines is that the assessment is strictly binary - either the test is fully compliant or not at all. In contrast, the manual assessment allows for nuances. But they are subject to interpretation by the user. We also find that metadata do not point to and identify the data they are describing. Most websites are designed to solely cater to a human data selection process. Moreover, many providers of data offer interfaces to
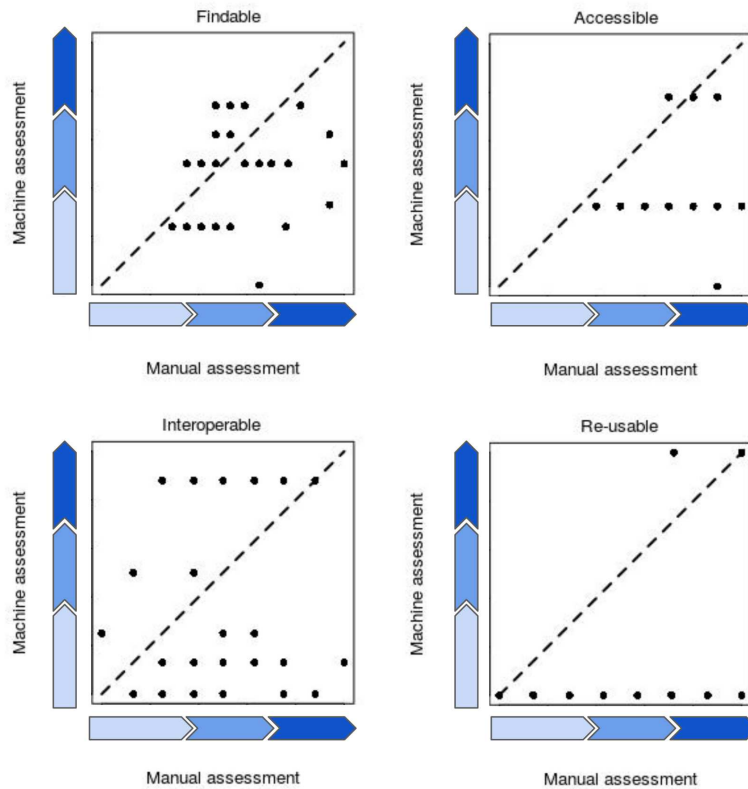
Figure 3. Stylized comparison of manual vs. machine assessments.

data and not the data themselves. In these cases, the design is not suitable for machine access. Examples include drop-down menus or hover boxes for value selection. Accessibility to (meta-)data is mostly impeded because metadata are not persistent. Among the bottlenecks for Findability are missing metadata pointers, long-term and stable access to data, and searchability of (meta-)data. At the same time, we do not observe lower scores for both machine and manual assessments. The same observation holds for the Accessibility criteria. The simple reason behind it is a selection bias - we study databases that are findable and accessible - at least through a website. This would change if a scalable, automated test, as suggested by[16], existed (or even crawling websites to find assessment candidates). Reusability is an issue because machines do not find license information, even if available for humans (hence the binary scoring results for machines). For Interoperability to work, data would need a much better standardized description of what they are about. A good example is the approach proposed for the smart grid by[9]. However, we also find that Interoperability is assessed most differently from the human and machine perspective (Fig. 3). A lack observed from a human point of view is that only islands of standardized knowledge representation and terminology exist; even less often they are interlinked, which would allow for the navigation of data and metadata from one field of expertise to the next. From the machine perspective, a standardization of vocabularies along with the pointers to its place of definition is indispensable. For example, while "Kilowatt hour" bears a meaning to users of energy data, machines need a semantic definition as, e.g., provided by QUDT, namely 'unit:KiloW-HR; URI: http://qudt.org/vocab/unit/KiloW-HR'. The problem is that semantic definitions are yet scarce, little known, and even less often implemented. Finally, we report that energy databases differ greatly in content, size, layout, and formats. Databases can store instrument readings such as metering data, data on price developments in the energy markets, and material composition of the electricity grid. This heterogeneity of energy data presents a grand challenge for scaling up database assessments. The current routines and tools are not up to the task, making the energy domain an excellent test-bed for improvements in this direction.

Discussion

The results disclose the difficulty of translating the FAIR guidance principles into domain-specific applications, as current FAIR data practices in the energy domain are still its infancy. Although the low carbon energy community has started efforts of FAIRifying energy data, platforms and tools are not yet fit to be integrated into the workflows of research teams. Most importantly, machine-actionability is not given at large.

This study is the first to assess and document FAIR and open data practices in the energy domain. We test 80 databases that are representative of data flows in the energy system with the help of manual and machine-based assessments. The comparison offers several novel insights, suggesting how to move forward in and with the community. We recommend the following action items for the energy domain (in order of priority):

1. Create institutions which are responsible for defining domain-specific, machine-actionable standards and which coordinate the future energy data space. This data space should serve as an entry point to FAIR data tools, workflows, and semantic web-services specific to the energy domain, besides ensuring interoperability with data spaces of other domains.

2. Approach the overall lack in understanding of how to implement machine-actionability through demonstrated use cases in the energy domain. Using a simple structured dataset, a blueprint can be developed to show how to enable machine-actionability. The use case illustrates how to assign persistent identifiers to (meta)data, link to existing standards, and assign licenses and access rights. The encouragement of peer-reviewed publications of such blueprints also addresses the incentive problem for investing into FAIR research data.

3. Harvest low hanging fruits by placing emphasis on the implementation of persistent identifiers for (meta-)data. Several repositories are offering these services.

4. Promoting and educating FAIR energy data stewards. The technical expertise and the resources needed to FAIRify energy data is out of reach for energy researchers. Even if assessment tools are available to support self-assessment of research data, the cycle of developing, assessing, and improving data documentation is out of scope for daily activities. In particular, the task of FAIRifying data connected to research publications should not be outsourced to the researchers.

5. Reverse the trend to prioritize the development of (graphical) user interfaces that prohibit the access to raw data. The assessments revealed that these interfaces are designed for human users only and are hardly machine-actionable.

Methods

Assessing the state of FAIRness accross low carbon energy research data relies on two basic steps: 1) to define a corpus of relevant data and II) to apply a FAIR evaluation methodology to this body. While the methodology can be to some extent developed independently of the energy domain, the systematic and comprehensive compilation of the data corpus is a domain-specific task. Using the corpus of relevant energy databases, assessment tools were applied to understand the overall compliance with the FAIR guiding principles as well as issues concerning each of the four principles. In our assessment of energy databases, we follow the 9 features of the community-driven approach as suggested in[16]: We carry out a number of assessment approaches, including automated assessments, assessments informed by the crowd (by drawing from a series of discussions about databases and FAIR gaps observed in the energy community, see[14]), and through intensive discussions within the group of authors of this publication. We also use a range of tools to assess the same databases. Most importantly, we obstain from a fixation on the assessment cores, which is why we do not report any score in the final figures. Instead, we embrace the idea that 'an intrinsic value' of scores is absent. Rather, assessment should be used as a guidance to draw conclusions for the way forward in improving the status quo of FAIR implementation in the domain.

We compile and select 80 databases representative of data flows in the energy system (Fig. 1). We test how representative our choice is with the help of an ontological concept based on Fig. 1, reflecting the importance of data flows in the energy system. The ontology draws from the Global Energy Assessment Report (GEA 2012) and established classification schemes for energy data, such as the Standard International Energy Product Classification by UNSTATS and IRES (SIEC), the Global Change Master Directory Keywords (GCMD 2020), JEL classification Codes (JEL 2021), and the European Science Vocabulary (EuroSciVoc 2020)). Table 3 in the supplementary material presents the 80 databases vis-a-vis key concepts. When choosing the databases, additional care has been taken to ensure a wide spread across hosts of databases (incl. general purpose repositories, institutional repositories - public and private, single databases and data sets, incl. datasets published as supplementary material to scientific publications). The uptake of the FAIR principles has been rapid, leading at the same time to manifold interpretations and, consequently, various assessment frameworks and metrics. Indeed, it is a very active area of research. Given the aforementioned role of automated services in the future energy system, an assessment of the machine-actionability of databases is of particular interest. Naturally, this can be best tested with an algorithmic framework. Indeed, the plethora of FAIRness claims and assessment tools led to state the FAIR principle more precisely on the one hand[11] and the development of automated tools on the other hand[6, 16].

We review available FAIR data assessment tools for manual and machine use. We select the ARDC DAIR data self-assessment tool[1] for manual assessment and the FAIR indicator maturity test[16] as one of the two available machine-actionable test. The other one is the rapidly evolving F-UJI test[6]. The rationale for our choice of the ARDC FAIR data assessment tool

is that it aligns with the FAIR principles, has a good balance between technical and non-technical questions (22 questions). The test allows scoring comparable to the machine test (12 tests) and was the only one available when this study was started. Although both tools have their own set of test questions, a mapping between is possible at the level of each of the FAIR principles. The supplementary material details the approach and connected scores (Table 5).

30 assessments were carried out manually, while 80 tests are machine-based. The rationale for this is that we were already able to identify systematic patterns and more would not have led to more. The number of assessed databases has been decided with emergence of generalizable patterns for the state-of-the-art. Note also that manual assessment was carried out before and after a briefing on how to assess, with the intention to detect the amount of subjectivity of tests. Fig. 3 shows the example for the interoperability criteria, for others see the Supplementary material.

The supplementary material provides further methodological details (Section 2), results from manual tests (Section 3.2), results from machines test (Section 3.3, spreadsheet) and aggregate scores for the comparison (Section 3.1).

Data Availability

Supplementary Material - Machine tests (spreadsheet)
https://doi.org/10.5281/zenodo.5577964.
Supplementary Material - detailed description of the adopted methodology, results of manual and machine tests (Text document)
https://doi.org/10.5281/zenodo.5578111

Code availability

We use the software for machine-assessments of databases with the help of FAIR maturity indicators (Wilkinson et al. 2019). The software is available at https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/about.

References

1. Australian Research Data Commons. FAIR self assessment tool. Available at: https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/. Accessed .

2. Booshehri, M., Lukas Emele, Simon Flügel, Hannah Förster, Johannes Frey, Ulrich Frey, Martin Glauer, Janna Hastings, Christian Hofmann, Carsten Hoyer-Klick, Ludwig Hülk, Anna Kleinau, Kevin Knosala, Leander Kotzur, Patrick Kuckertz, Till Mossakowski, Christoph Muschner, Fabian Neuhaus, Michaja Pehl, Martin Robinius, Vera Sehn, Mirjam Stappel, Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis, Energy and AI, Volume 5, 2021, 100074, ISSN 2666-5468, https://doi.org/10.1016/j.egyai.2021.100074.

3. Borycz, J. and Carroll, B. (2020) Implementing FAIR Data for People and Machines: Impacts and Implications - Results of a Research Data Community Workshop. Information Services & Use 40, 71–85.

4. COM(2020). Critical Raw Materials Resilience: Charting a Path towards greater Security and Sustainability

5. Davis, R., Shrobe, H.E., Szolovits P. (1993) What is a knowledge representation? AI Mag., 14, 17–33, DOI:11310.1609/aimag.v14i1.1029

6. Devaraju, A., Huber, R., Mokrane, M., Cepinskas, L., Davidson, J., Herterich, P., L'Hours, H., de Vries, J., White, A. (2020) FAIRsFAIR Data Object Assessment Metrics, https://doi.org/10.5281/zenodo.3934401.

7. EOSC HLEG (2018). Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC). Prompting and EOSC in practice. Available online: https://publications.europa.eu/en/publication-detail/-/publication/5253a1af-ee10-11e8-b690-01aa75ed71a1 (accessed: 18.07.2019).

8. EC Expert Group on FAIR data (2018). Interim FAIR Data Report: Turning FAIR into reality, https://doi.org/10.5281/zenodo.1285272.

9. Huang, B., Bai, X., Zhou, Z., Cui, Q., Zhu, D., Hu, R. (2017) Energy informatics: Fundamentals and standardization. ICT Express, 3, 76-80, DOI: 10.1016/j.icte.2017.05.006.

10. IEA (2017), Digitalisation and Energy, IEA, Paris, https://www.iea.org/reports/digitalisation-and-energy.

11. Jacobsen, A., de Miranda Azevedo, R., Juty, N., et al. (2020), FAIR Principles: Interpretations and Implementation Considerations, Data Intelligence 2, 10-29.

12. Jin, D., Ocone, R., Jiao, K., Xuan, J., 2020. Energy and AI. Energy and AI 1, 100002.

13. Stein, Helge S. and Gregoire, John M. (2019), Chem. Sci. 10, 9640-9649.

14. Wierling, A., Schwanitz, V.J., et al. (2021) Manuscript entitled 'Advancing FAIR metadata standards for low carbon energy research', submitted to Energies.

15. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018.

16. Wilkinson, M.D., Dumontier, M., Sansone, SA. et al. (2019) Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci Data 6, 174.

17. Zangeneh, P., McCabe, B. (2020) Ontology-based knowledge representation for industrial megaprojects analytics using115linked data and the semantic web. Advanced Engineering Informatics, 46, 101164, DOI: 10.1016/j.aei.2020.101164

Acknowledgements

Author contributions statement

V.J.S. and A.W. equally contributed to: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, vizualization, writing - original draft, writing - review & editing. All co-authors supported: funding acquisition, formal analysis, methodology, writing - review & editing.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Schwanitzetalscorescalculationmachineassessment.xlsx
- SupplementaryMaterialtothemanuscriptTowardsFAIRandopendataforlowcarbonenergycurrentstateandcallforaction.pdf