Supplementary material

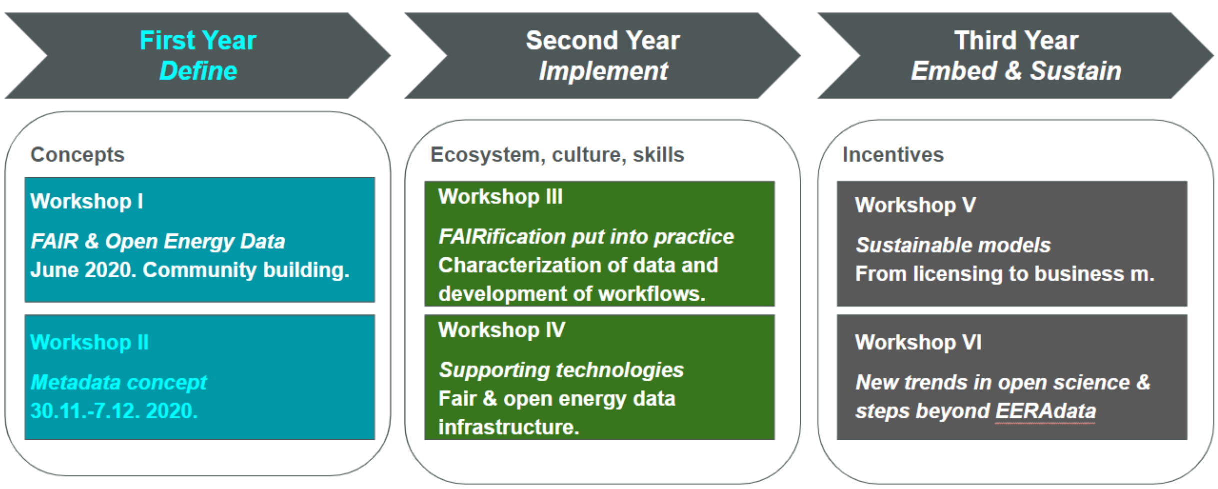
"FAIR Metadata Standards for Low Carbon Energy Research—A Review of Practices and How to Advance"

1 Workshop to engage the low carbon research community

Implementing a FAIR and open data ecosystem for low carbon energy research is dependent on the establishment of a community on the topic. Community-discussed and accepted standards, e.g., on metadata as the focus of this report is, have a better potential to lead to sustainable solutions. The method to build and engage the low carbon research community on FAIR and open data applied in the H2020 EERAdata project rests on the three pillars: (1) Designing a series of workshops that enable reflection by the community, (2) Connecting with existing networks, initiatives, and umbrella organisations, and (3) Partnership with the European Energy Research Alliance EERA.

The 3-year project EERAdata with six partner institutions (Link to materials https://www.eeradata.eu/) is developed along with the 'Index to FAIR Action Plan' as recommended by EC Expert Group on FAIR data 2018. More specifically, the testing and discussion of the FAIR/O ecosystem (incl. the community platform) are organized through a series of workshops that take place twice a year with sequential annual foci (Fig. A.1). The first year discussions created the basis for the community paper on 'Advancing FAIR and open metadata standards for low carbon energy research.

Figure S1: The workshop concept of the EERAdata project is inspired by the Expert Group of the European Commission on FAIR data 2018, FAIR Action Plan.



Each workshop takes three days. The first day is dedicated to state of the art for each workshop topic. Experts from the energy and other research communities and the energy sector are invited to present best practices and open issues. Mutual learning and exchange of knowledge are facilitated in moderated discussions following presentations. The first day advances the understanding of cutting-edge concepts. It approaches FAIR/O data (e.g., the role of metadata, data management plans, metrics, incentive mechanisms to set up FAIR and open data ecosystems, etc.). The second day uses the knowledge and understanding obtained during the first day to translate the concepts and approaches to low carbon energy research data at the selected use cases. The goal is to understand data discovery methods and synthesize metadata from domain experts' mental models. Thus, the work done during the second day takes place in parallel groups structured around use cases. Use cases in EERAdata are "Energy efficiency of buildings'', "Power transmission and distribution networks'', "Material Solutions for low carbon energy," and "Low carbon energy and energy efficiency policies''. The workshops' key participants are consortium members for each use case and the stewards of FAIR and open data principles recruited from invited experts and other interested stakeholders. The third day synthesizes, integrates, and concludes the results achieved during the first two days. First, discussed concepts and approaches for the FAIR and open ecosystem are refined. Second, the next workshop's agenda is agreed upon and specified concerning experts to be invited and topics to be addressed. All workshops are open to participants from outside of the consortium.

Workshop series to develop the community paper

The consortium initiated and facilitated discussions on advancing metadata standards for low carbon energy research through a series of community-wide workshops. The workshops aimed to build on the contributions of specialist and non-specialist participants to a community paper designed to reflect the broad perspectives of the low carbon energy research community on the state of the art of metadata practices, perceived needs, envisaged prospects of deploying heterogeneous data with different levels of granularity and recommendations to accomplish this.

The working title of the community paper 'Advancing metadata standards for low carbon energy research' underpinned all the key steps and methodology adopted in this present research and which draw their content from EERAdata workshops organized in Phase 1 of the project implementation, which is about "Define FAIR and open energy data'' (see Fig. A.1.). Two workshops were held. Workshop I titled 'FAIR principles and metadata for low carbon energy' occurred in June 2020. The workshop, II titled 'FAIR and open energy metadata,' was subdivided into various workshop sessions between Nov 30 – Dec 7, 2020.

The groundwork from Workshop I fed into online community building and provided a starting point for reaching different participants and generating interests that stimulate constructive inputs into Workshop II's community paper focus. An initial working draft was provided to participants to outline the context of Workshop II, including a list of relevant literature and projects. Workshop II started and ended with one-day writing sessions organized in breakout groups (Fig. A.1). Material to workshops can be assessed from the project website: https://www.eeradata.eu/ (Section 'Events,' EERAdata workshop No. 2).

Figure S2: Workshop series to develop the white paper on low carbon energy metadata. Framing workshops/Writing Sessions on Day 1 and Day 7. In between (Day 2-6) topical and expert sessions.

Schedule Day 1 Writing session (30/11, 10-16)	Contribution to the white paper
10-10.15 Welcome and introduction to the first draft of the planned paper (Valeria Jana Schwanitz and August Wierling, Western Norway University of Applied Sciences, short: HVL) **10.15-10.30** Discussion and collection of comments and notes. **10.30-10.45** Creation of writing teams in break out-groups (BOGs). Preferences are collected beforehand. **10.45-15** Work in BOGs. **15-16** Wrap up: Reports from BOGs, collection of feedback and comments. Suggestions of issues to take up in topical and expert workshops.	**Preparation:** read ahead and records of the preparatory workshop, commenting draft of community paper **Writing team 1:** Literature review on (meta)data for low carbon energy - gaps & needs. Documentation of work: wiki, Section 3.1, and others. **Writing team 2:** Review of existing metadata concepts in the energy domain. Compilation of work into a table. Development of evaluation criteria: Section 3.2 and 3.3.1. **Writing team 3:** Review of existing platforms in the energy domain to support metadata access. Compilation of work into a table. Input to Section 3.2 and 3.3.2.

Schedule Day 7 Writing session (7/12, 10-16)	Contribution
10.00-10.45 Welcome and input from topical and expert workshops (VJS, AW - HVL) **10.45-11.00** Guidance for writing teams **11.00-11.15** Creation of writing teams in BOGs. **11.15-15.00** Work in break out-groups. **15.00-16.00** Wrap up day 2 of the framing workshop: Reports from work done in BOGs. Collection of feedback and comments from all. Guidance on how to finalize the paper collaboratively after the workshop.	Focus on input to Section 4. **Writing team 1:** Compile and discuss lessons learned (Section 4.1) **Writing team 2:** Compile and discuss recommendations (Section 4.2) **Writing team 3:** Individually working on sections and open issues.

Schedule Day 2-6, Topical & expert sessions
Tuesday, 1/12, 10-14 **Policy use case**
10-11 Introduction to the session and description of tasks - exploring taxonomies with the software ‘gitmind’. How to organize domain knowledge at different levels of granularity? (Maria Bałazińska and Mariusz Kruczek, Central Mining Institute) **11-13** Time for individual work. Group discussions upon request. **12.45** Collecting of contributions **13-14** Presentation and discussion of work Contribution to the white paper: Section 3.2.1, state of the art (Section 3.1), and outlook (Section 4). Work compiled in a separate online document. Workshop participants also explored and extended taxonomies.

Wednesday, 2/12, 10-12 **Buildings efficiency**
Introduction: What are the aims of the "Buildings Efficiency" use case of EERADATA? Mehmet Efe Biresselioğlu, Izmir University of Economics. **Presentations and panel discussion:** **(I)** Standardized Flexibility: FAIR data principles in H2020 projects ECHOES and ENCHANT databases, Jens Olgard Dalseth Røyrvik, Norwegian University of Science and Technology. **(II)** Achieving Fair Principles: An Overview of the Hotmaps Project and Hotmaps Building Stock EU28 Dataset, Simon Pezzutto, EURAC. **(III)** openENTRANCE CS1 residential energy demand response - the importance of nomenclature and data transparency for open source energy system model platform Ryan O'Reilly, Energieinstitut, Johannes Kepler University. **(IV)** Exceed database: lessons learned and continuous improvement, Daniele Antonucci, EURAC. Contribution to the white paper: State of the art (Section 3.1) and outlook (Section 4).
Thursday, 3/12, 10-12 **Aligning metadata and workflows**
10.00-10.15 Aims and intended role of the platform in the community (Manfred Paier, Austrian Institute of Technology, short: AIT) **10.15-10.30** Status quo (initial mock-ups, Community Forum) (Astrid Unger, AIT) **10.30-11.30** Experiences from UC4, UC1, other Use Cases, and Gap Analysis: 1. Searching for data, 2. Linking existing datasets. General discussion. Critical article on ENTSO-E transparency platform? What are the proper search channels for whom? What is the platform concept? **11.30-11.45** Top-down metadata, bottom-up metadata (Michael Barber, AIT) **11.45-12.00** for designing and implementing the platform (Manfred Paier, AIT) **Contribution to the white paper:** Results from a questionnaire on metadata user stories, Section 3.3.2, state of the art (Section 3.1), and outlook (Section 4).
Friday, 4/12, 15-17, **M4M Workshop: Making domain-relevant machine-actionable metadata at scale**
Session at the International FAIR Convergence Symposium. **Part I** Review of previous M4M workshops. An introduction to M4M by Erik Schultes, GO FAIR, followed by presentations of M4M.4 - VODAN Africa (Mirjam van Reisen, Leiden University); M4M.5-6 - DeiC - AnaEE & NEST (Nikola Vasiljevic, Technical University of Denmark); M4M.7-13 - ZonMW - COVID Program (Barbara Magagna, Environmental Agency of Austria); short Q&A. **Part II** Brainstorming the next M4M workshop: EERAdata consortium. How to set up an M4M Workshop, the EERA case? - (E.S.) Introduction to the EERA case - (Valeria Jana Schwanitz, Western Norway University of Applied Sciences). Panel discussion. **Contribution to the white paper:** Section 3.2.2, state of the art (Section 3.1), and outlook (Section 4). The idea was to discuss how to plan domain-specific M4M workshops with and for EERAdata.

2. Evaluating existing metadata practices in the energy domain

Two hermeneutic approaches on how to define metadata are commonly found in the literature. First is the direct way states '*metadata is …*' (NISO, 2007; Smiraglia, 2005; Pomerantz, 2015; Hill, 2016; Gregg, 2019). The second approach defines metadata by its function, i.e., '*metadata do …*' (Day, 2001; Gilliland, 2016; Haynes, 2018). For example, Day's model of metadata purposes distinguishes between the resource description, resource discovery, administration and management of resources, record of intellectual property rights, documenting of software and hardware environments, preservation management of digital resources, and providing information on context and authenticity (Day, 2001). A synthesis of both ways is the definition by Jane Greenberg, 2003, who defines metadata as "*structured data about an object that supports functions associated with the designated object*." A consequence of that definition is that data are inevitably linked to their metadata, and without metadata, data loses usefulness (Kahn and Wilensky, 1995). Furthermore, Pomerantz, 2015; Gilliland, 2016, and Wilkinson et al., 2016 emphasize the role of metadata in knowledge representation and information reduction.

While the above definitions relate to metadata in general, we see the need for an extension if metadata is linked to data resulting from research activities; regarding this publication, they concern low carbon energy research results. The functional approach is the starting point because we apply criteria to assess the quality of current metadata practices in the energy domain. The need for extension of the above metadata definitions can be further reasoned by science's responsibility to ensure reproducibility and valid answers to the research questions raised. In practical terms, this implies the need to capture an observation by its data and record the entire workflow from data creation to scientific publication. Otherwise, the reproduction of results is impossible. The validation of research results goes a step further by also qualifying the answers given as valid for purpose. Metadata for research, thus, serves the additional function to also deliver the necessary information for reproduction and validation.

We expand on the practical approach towards defining metadata for low carbon energy research through the functions metadata should serve. As discussed in Section 3.1 (main text), metadata supports the searching for information for different user groups. Typically, a range of standards is implemented to provide these functionalities. While various measures to deliver the desired functionality are denoted as a metadata concept, we call the implementation of such a vision a metadata practice. To understand the extent of performance, we analyze existing metadata practices in the low carbon energy research community and apply evaluation criteria (see Box 1, main text). The criteria to evaluate metadata practices were derived from reviews and discussions. It is important to note that they do reflect not only human-readability but also machine-actionability of metadata. The criteria were tested against the extent to which they are in line with standards proposed by the following references:

- **Metadata 2020 principles** are suggested by a library/publishers backed consortium (Mitchell, Counsell, 2018; Schneider 2018). Accordingly, high-quality metadata should feature compatibility (content for machines & people that is open, interoperable, parsable, machine-actionable, and human-readable); completeness is being reflected and comprehensive; credibility that is being discoverable and of longevity; reusability and institutionally curated data. These principles are also well in line with the criteria of 'metadata security,' as described in Haynes 2018.
- **NISO principles** are suggested by the National Information Standards Organization (NISO 2007), a non-profit association accredited by the American National Standards Institute (ANSI). They recommend metadata to show usefulness that conforms to community standards, interoperability, and object-orientation; that is, "*good metadata records are objects*

themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification."

- **Metadata for machine principles**. These are taken from the current software implementation of Wilkinson et al. 2019. The criteria include tests whether metadata has unique and persistent identifiers; is being structured & grounded (e.g., having a hierarchical organisation); is linked via persistent identifiers to data; is searchable (e.g., by search engines), retrievable (e.g., open & free protocol, authentication & authorization, persistency); resolves applicable content (including knowledge representation, community standards, and vocabulary); and is licensed.
- **Metadata Quality Assurance Framework**: Király (2015) and Stiller & Király (2017) propose a conceptual framework for the metrics based on Bruce and Hillmann (2004) and Ochoa and Duval (2009) Ariadne project's computations. They add quantitative metrics for their criteria, which are: completeness (quantity of available information, weighted value option); accuracy (correctness of information); conformance to expectations (the degree to which metadata fulfills the requirements of a given user community); logical consistency and coherence (w.r.t. standards in the domain); accessibility (machine actionability and human readability); timeliness (currency); and provenance (knowledge of the source and its reputation).
- **Other:** Kemp, Dean & Chodacki (2018) suggest that valuable metadata are those which map interconnections, balance consistency, and flexibility, and provide a shared, distributed resource for all.

In the next step, we test existing metadata practices in the energy domain with the principles summarized in Section 3.2.3 (main text), i.e., the richness of metadata, the level of consensus on them in the community, the accessibility & transparency of metadata, the extent of links to other metadata as well as their functional implementation. To this end, Table S1 shows the evaluation results for examples of metadata practices (4th column). The table also specifies these practices by their application scope, the standards applied, the support services offered, and references. A discussion of the findings is presented in Section 3.3.1 in the white paper. Note that the compilation of Table S1 and the evaluation of the practices as part of the community workshop create input to the White Paper.

Table S1: Existing metadata practices in the energy domain. The evaluation is based on applying criteria (Section 3.2.3: A - Richness, B - Consensus: C - Accessibility & transparency, D - Linked metadata, E - Functional implementation).

Metadata practice	**Scope of applicatio n**	**Standards used**	**Evaluation**	**Support services**	**References**
Metadata String by Open Energy Platform	Energy Systems Modelling	Dublin Core	**A:** High, 80 different definitions for all types of metadata. **B:** Medium, limited to the open modeling community. **C:** High, open platform. **D:** Low, ontology in development. **E:** High, open license, JSON as	Semantic search with dbpedia databus under development (Q1/2021)	dbpedia databus Github: OEP Metadata & Metadata Description

			data format.		
Standard International Energy Product Classification (SIEC); Single Integrated Metadata Structure (SIMS)	Energy statistics covering the statistical production process	Own standard	**A:** High, covering different aspects from administrative to work-flow and content information, modular content information. **B:** High, UN institutional back-up. **C:** Medium. Accessible for humans, machine-actionability limited. **D:** Low-medium. classification available. **E**: Does not apply.	Statistical Data and Metadata Exchange (SDMX) guidelines; Energy Statistics Compilers Manual; standardized energy indicators	SIEC (link) by International Recommendations for Energy Statistics (IRES)
Energy Identification Code scheme (EICs)	Identification of electricity and gas market participants and other entities active within the Energy Internal European Market	ENTSO-E standard	**A:** High for the field of application **B:** High, approved by ENTSO-E and EU regulations **C:** Medium, limited accessibility for machines, **D:** None-low, **E:** Excel sheet and pdf manual only	Manuel. List and backup by taxonomy. Online access to the list of participants	Link
CityGML Energy Application Domain Extension	Representation of 3D urban objects (buildings & infrastructure), supporting energy simulation	City Geography Markup Language (CityGML)	**A:** Medium, the granularity of the energy extension is relatively low (e.g., not resolving all house appliances). **B:** High, international working group; formal process to extend catalog, referencing ISO standards, **C:** Medium-High, accessibility for machines but navigation for humans difficult, **D:** High, linked through its implementation, **E:** High, class diagram and XML Schema available	Standardized and interlinked terminology based on XML; online feature catalog	CityGML Energy ADE
Common Information Model, by	Exchange of data and	Standards: IEC 61970, IEC 61850	**A:** High, capturing many technical features for power data exchange	Service for purchase: documentati	Wikipedia documentation

the International Electrotechnical Commission	modeling of electrical power systems		**B:** High, established protocols, backed up by many international standards and based on technical consensus, **C:** Limited, restricted to purchasers (incl. access to metadata), **D:** High, linked through its implementation, **E:** High, XML and RDF Schema available	on and software for knowledge documentation in the field	
Hotmaps	Building stock, building energy demand, space heating and cooling, domestic hot water usage data for a mapping and planning toolbox	Own standard	**A:** High, both in terms of types and coverage. **B:** High. Aggregates and combines data from various databases. The use of estimations for missing data is a limitation. **C:** High. Accessible through standard web query, downloadable without any additional authorization or permissions. **D:** Medium. Metadata is attached, however not rich and fully explained in all cases **E:** High. Using Creative Commons License. JSON, GeoTiff, .shp, and .csv formats are used (datapackage.json from Fristionlessdata.io).	Online platforms where the datasets and the planning tools are available. Also, project documentation and web page.	Project website, Gitlab, Dataset documentation
Emodnet	Data from the marine environment: Bathymetry, Geology, Seabed habitats, Chemistry, Biology, Physics, Human activities	SeaDataNet CDI metadata standard. Common Data Index (CDI) metadata format is based upon the ISO19115	**A:** High, for all data portals **B:** High, international working groups; formal process to extend the catalog. **C:** High, information, data curation, documentation, standardizations are defined and available. **D:** High, data are interlinked and available from different portals to the same map **E:** High, metadata and file formats comply with standards and community agreements.	Many services in each data portal and the central platform, from data ingestion to data analysis, mapping, and training.	www.emodnet.eu www.emodnet-ingestion.eu
Taxonomy supporting the	Wind energy	Own standard based on	**A:** High in the field of application. **B:** Medium, developed in the	Ontology service	Wind energy taxonomy,

ShareWind		literature review and expert elicitation	context of EU projects **C:** High, available in, e.g., CEDAR workbench. **D:** Medium to high, taxonomy only. **E:** High, converted to RDF by SKOS lexicalization.		
PV-GIS	Geographic Information System for Photovoltaics	Own standard	**A:** High for the field of application. **B:** High, approved by EU (PV-GIS is located on *the European Commission's science and knowledge service)* **C:** High, can be operated by a human through webpage tools (PV Performance, Solar radiation, TMY). **D:** Low. **E:** Medium to High, uses JSON-LD without semantic support.	PV-GIS contact points	PV GIS 5.2 documentation
NOMAD	The web platform for FAIR sharing and analyzing materials science data	Own standard: NOMAD Meta Info	**A:** High, 168 public metadata according to the NoMAD Meta Info. **B:** Medium to high, developed in EU projects (EU partners only) opened worldwide. **C:** High, can be operated by a human through webpage tools (e.g., Periodic element table) or Python tools. **D:** High, linked through its implementation; possibility to explore the database. (Encyclopedia & Repository) or to analyze data. **E:** High, uses JSON and MessagePack file formats.	Community discussion forum, Technical support, Online tutorials	Nomad Meta Info Home
SEMANCO - Semantic tools for carbon reduction in urban planning	Energy efficiency in buildings	Own standard, translation of Suggested Upper Merged Ontology (SUMO)	**A:** High, several standard tables (25) build with objects and attributes. **B:** Medium, developed in EU projects (EU partners only); dedicated to EU policy development for CO2 reduction at an urban level. **C:** Medium, operated to dedicated software (Clock-on, Map-on, Search-on).	Tools web page: http://www.semanco-tools.eu/search-on Wiki publications	SEMANCO Ontology

			D: High, ontology-based, linked RDF. **E:** Low, does not support free code editing in terms of XML code.		
Materials Experiment and Analysis Database (MEAD)	Material Science toward solar fuels	Own standard	**A:** Low, metadata is related to specific identifiers. **B:** Medium, developed in a US initiative, opened worldwide. **C:** Low, navigation for humans complicated, not dedicated for machine search. **D:** Medium, metadata are linked through experimentation plans. **E:** Low, Java implementation.	Publications in Scientific Journals	MEAD, https://solarfuelshub.org/materials-experiment-and-analysis-database
Other practices used in the domain but not necessarily domain-specific					
General repositories and institutions providing metadata guidelines	Administrative information	DDI, DC, CSMD; Specific for research: Datacite, Dataverse, Dryad, FigShare; (RAMON); OECD Glossary (link), EPA (link); W3C Standard Provenance	Not individually assessed. Assessment comment: Overall low-medium level of implementation, depending on the requirements from the repositories or institutions used to publish. E.g., Zenodo requests very detailed administrative metadata.	Supports linking across domains, a relatively high degree of standardization which supports machine-actionability	See Farnel & Shiri 2014 for assessments of the available repositories.

References

Bruce, T. R., Hillmann, D., I. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In Metadata in Practice, Hillmann, and Westbrooks, eds. ALA Editions (2004).

Day, M. Metadata in a nutshell. Information Europe 6, p. 11 (2001).

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

Gilliland, A.J. In M. Baca, Introduction to Metadata, Getty Publishing, 2016

Gregg, W. J, Erdmann, C., Paglione L.A.D., Schneider, J., Dean, C., A literature review of scholarly communications metadata, Research Ideas and Outcomes 5: e38698 (2019)

Haynes, D. Metadata for Information management and Retrieval. Facet publishing (2018). London.

Hill, T. In M. Baca, Introduction to Metadata, Getty Publishing, 2016

Kahn and Wilensky (1995) http://www.cnri.reston.va.us/k-w.html. (accessed 01.12.2020)

Kemp, J., Dean, C., and Chodacki, J. Can Richer Metadata rescue research? The Serials Librarian 74, 207-211 (2018).

Király, P., A Metadata Quality Assurance Framework, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), (2015), Göttingen, Germany

Mitchell, D., Counsell, F. Metadata 2020: A collaborative effort to improve metadata quality in scholarly communications. Septentrio Conference Series, (2018). 10.7557/5.4471.

NISO 2007, National Information Standards Organization (NISO), A framework of guidelines for building good digital collections (3rd edition). https://www.niso.org/sites/default/files/2017-08/framework3.pdf

Ochoa, X., Duval, E. Automatic Evaluation of Metadata Quality in Digital Repositories. International Journal on Digital Libraries 10 (2-3), 67-91 (2009) https://lirias.kuleuven.be/bitstream/123456789/255807/2/xavuxavier-pre.pdf

Pomerantz, Jeffrey. Metadata, MIT press, 2015

Schneider J. Metadata 2020: Updates and Plans, Research Data Alliance, Berlin 2018. Available at: https://rd-alliance.org/system/files/documents/RDA_Metadata2020.pdf (accessed: 28.12.2020)

Smiraglia, Richard P. Metadata: A A Cataloger's Primer, Psychology Press, 2005

Stiller, J., Király, P., Multilinguality of Metadata Measuring the Multilingual Degree of Europeana's Metadata, In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th—15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 164—176.

Wilkinson, M.D., Dumontier, M., Sansone, SA. et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci Data 6, 174 (2019). https://doi.org/10.1038/s41597-019-0184-5

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18