	CURATORIAL ELEMENT	DESCRIPTION	DOCUMENTATION LEVEL	
			Criteria to meet minimum standard	Criteria to meet standard of excellence
SCOPE				
1	Context, purpose, motivation	This information explains the purpose of dataset creation for the specified domain.	Documentation discusses the problem domain, what problems the new dataset addresses, the relevance of those problems, and the need for a new dataset in comparison to existing datasets.	Documentation explains how the context of the dataset affects possible reuse and includes reflection on the dataset creators' awareness of social, political, and historical context.
2	Requirements	The translation process from a "real-world" problem to a "ML problem" for which the dataset is created [21, 23] consists of numerous decisions, expertise, and worldviews that should be documented in order to understand the context in which the problem situation was framed.	Documentation states how the problem was formulated and how the dataset creation plan was generated.	Documentation includes reflection on how the problem formulation introduces intrinsic biases.
ETHICALITY AND REFLEXIVITY				
3	Ethicality	Ethical considerations are critical to the fair and accountable creation and (re)use of datasets.	Documentation discusses how the benefits of creating the dataset outweigh any harms of creating it (see proportionality principle), and it discusses informed consent if the dataset is about humans.	Documentation goes beyond requirements listed in ethics framings like guidelines/policies/checklists. For example, documentation discusses alternate methods of dataset creation that were not used because of potential ethical harm.
4	Domain knowledge & data practices	Creating a dataset involves, often tacit, expertise about one or more domains as well as data practices. Articulating both types of nuance required in dataset development makes data work more transparent [11, 14, 21, 24, 26].	Documentation states the domain-specific expertise and data skills required in developing the dataset.	Documentation discusses the required expertise needed to understand the intended purpose of the dataset and to reuse it.
5	Context awareness	Context awareness demonstrates an understanding of the subjective, non-neutral nature, and situatedness of data.	Documentation includes a positionality statement.	Documentation adopts a reflexive approach to dataset development. For example, documentation discusses how field epistemologies impact assumptions, methods, or framings.
6	Environmental footprint	This element is for dataset creators to reflect and quantify the footprint of their dataset creation process [1].	Documentation contains a quantitative assessment of environmental footprint and clearly defined scope of what was measured.	Documentation includes a lifecycle assessment and the corresponding environmental footprint, and an assessment of design choices and rationale for the choices.
DATA PIPELINE				
7	Data collection	Disclosing data sources is essential in the data collection process. Further reflection on the process of selecting those sources can reveal important interpretive assumptions [21] and historical and representational biases [14].	If data was collected, documentation states how and why data and metadata were collected from the data source(s). If data was synthesized, documentation discusses: 1) how and why the data was	If data was collected, documentation discusses the process of defining criteria for selecting data source(s), specifies the criteria, explains why those criteria were chosen, and how the selected data sources are evaluated against these criteria. If data was synthesized, documentation includes a reflection on potential intrinsic biases of the synthesis process, how the synthesis process

			synthesized and 2) whether the data was synthesized to match labels, if used.	shaped the features of the data, the limitations of the synthesis process, and how the synthesized data relates to the real-world distribution of the data it represents.
8	Data processing	Data processing involves cleaning, transforming, and wrangling data. Data processing decisions have impacts on the ultimate “cleaned” data that is used [18, 21]. Detailed documentation of this process enables outcomes of the model to be traced back to processing decisions.	Documentation discusses the process of cleaning, transforming, or wrangling data.	Documentation goes beyond what is done to discuss how the decisions about data processing were made and why, and potential impacts of the processing decisions.
9	Data annotation	Data annotation or labelling, regardless of the guidelines provided to reduce worker bias, can lead to disagreements on how data should be annotated (either between annotators or between dataset creators and annotators).The inclusion of this documentation highlights what is considered the “ground truth” [4, 21, 22] by the dataset creators which impacts how annotation is performed [15].	Documentation discusses the process of annotation. If any labels are used, the documentation includes the following: If labels are derived from the data: documentation discusses how data was interpreted to generate labels. If the labels were created first and the data was derived from the labels: documentation discusses how the relationship of the data to the labels was verified. If labels are obtained from elsewhere: documentation discusses where they were obtained from, how they were reused, and how the collected annotations and labels are combined with existing ones.	Documentation discusses the process of annotation with depth and reflexivity by including a reflection on how annotations (including labels, if used) represent differing worldviews and social backgrounds. Additionally, if labels are derived from the data: documentation discusses how the labels are robust, i.e., not sensitive to variability and how disagreements on annotation were reconciled.
DATA QUALITY				
10	Suitability	Suitability is a measure of a dataset’s quality with regards to the purpose defined.	Documentation discusses how the dataset is appropriate for the defined purpose.	Documentation discusses how dimensions such as accuracy, completeness, timeliness, and consistency contribute to the quality of the dataset in being used for the defined purpose. For example, timeliness (i.e., age) of data should be appropriate for the defined purpose.
11	Representativeness	Representativeness is a measure of how well a sample set of data represents the entire population. Sampling procedures and decisions about data sources can introduce extrinsic bias [21]. For example, choosing Reddit or Twitter as a data source can perpetuate dominant social biases rather than being a representative sample of the target population [1].	Documentation defines the population and discusses the extent to which the sampling procedure is representative of the population.	Documentation includes reflection on how the dataset creation process overall, and the sampling procedures specifically, affect extrinsic bias.

12	Authenticity	Authenticity of a dataset is about whether the dataset "is what it purports to be" [5, 7, 8, 12, 25], which is a responsibility of dataset creators [17]. Authenticity can be established by assessing the identity and the integrity of the record [5, 6, 10, 13, 16, 19]. Integrity of a dataset is about whether "the material is complete and unaltered" [2, 3, 9, 12, 20].	Documentation discusses how authenticity has been established and maintained, i.e.,
			• Has the identity and origin of all data been verified?
			  • For data that is obtained, it is clear how the dataset creators have verified the identity of the dataset they reuse.
			  • For data that is generated, it is clear how they have been created and by whom.
			• Has the integrity of all data been verified?
			  • For data that is processed in any way, it is clear how processing steps may have impacted integrity.
				Documentation states how others can establish the authenticity of this dataset, i.e.,
				• Documentation provides a persistent identifier and provenance information for the dataset in order for reusers to establish identity.
				• Documentation provides mechanisms for reusers to verify the integrity of their dataset.

13	Reliability	Reliability is about how well the dataset is "capable of standing for the facts to which it attests" [5], i.e., how certain we can be that its data points reflect what they represent.	Documentation discusses how the reliability of the dataset has been established and maintained, including the verification steps taken to ensure reliability, where necessary, i.e., • It is clear for each data element what synthetic or real-world phenomenon it represents.	Documentation states how others can establish the reliability of the dataset, i.e., • Documentation provides mechanisms to enable verification of what synthetic or real-world phenomenon each data element represents.
14	Structured documentation	Context documents in standardized structures provide information on the content of the dataset which is critical in establishing its usage in a well defined format.	Documentation includes a standardized context document. Acceptable formats include context documents that follow an established structure such as datasheets, data statements, and nutrition labels.	The context document addresses all mandatory items.
		DATA MANAGEMENT		
15	Findability	Ensuring findability is about enabling the dataset to be discovered for reuse after its development [27].	Documentation discusses how the dataset is findable by providing a globally unique and persistent identifier (URLs are not persistent).	Documentation includes metadata and both the metadata and data are stored in a searchable repository.
16	Accessibility	Accessibility is about enabling the dataset to be obtained after its development [27].	Documentation states all information and tools required to access the content of the data, and the identifier navigates to the metadata and data.	Documentation includes a communications protocol, an authentication and authorization procedure, and provides metadata that will be available even if data access is removed.

17	Interoperability	Interoperability ensures that the dataset can be integrated with other applications and workflows [27].	Documentation discusses how the dataset integrates with other data, workflows, applications, etc. (i.e., that both the metadata and data are readable by humans and machines).	Documentation has metadata and data that both use controlled vocabularies and link to other resources using qualified references.
18	Reusability	Ensuring reusability requires providing information such as relevant provenance and usage [27].	For both metadata and data, provenance information includes at least all of the following: 1) where the data came from, 2) who collected it, and 3) when it was collected.	Documentation has metadata and data that are both described using domain-relevant standards, state license and usage information, and provide additional provenance documentation as described by FAIR best practices.

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '21*), March 01, 2021. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[2] June M. Besek and Philippa S. Loengard. 2007. Maintaining the Integrity of Digital Archives. *Columbia J. Law Arts* 31, (2007), 267.

[3] Li Cai and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* 14, (May 2015), 2–2. https://doi.org/10.5334/dsj-2015-002

[4] Catherine D'Ignazio and Lauren F. Klein. 2023. *Data Feminism*. MIT Press.

[5] Luciana Duranti. 1995. Reliability and Authenticity: The Concepts and Their Implications. *Archivaria* (May 1995), 5–10.

[6] Luciana Duranti. 1998. *Diplomatics: New Uses for an Old Science*. Scarecrow Press.

[7] Luciana Duranti. 2005. The long-term preservation of accurate and authentic digital data: the INTERPARES project. *Data Sci. J.* 4, (2005), 106–118. https://doi.org/10.2481/dsj.4.106

[8] Luciana Duranti. 2007. The InterPARES 2 Project (2002-2007): An Overview. *Archivaria* (2007), 113–121.

[9] Luciana Duranti and Heather MacNeil. 1996. The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project. *Archivaria* (October 1996), 46–67.

[10] Luciana Duranti and Randy Preston. 2009. International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records. *Rec. Manag. J.* 19, 1 (January 2009). https://doi.org/10.1108/rmj.2009.28119aae.003

[11] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (2022), 1–29. https://doi.org/10.1145/3555760

[12] Sarah Higgins. 2009. DCC DIFFUSE Standards Frameworks: A Standards Path through the Curation Lifecycle. *Int. J. Digit. Curation* 4, 2 (October 2009), 60–67. https://doi.org/10.2218/ijdc.v4i2.93

[13] Asen O Ivanov. 2019. The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation: Curation Culture and Evaluative Practice. University of Toronto.

[14] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 27, 2020. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

[15] Julian Posada. 2023. *Platform Authority and Data Quality*. Retrieved from https://www.berggruen.org/ideas/articles/decoding-digital-authoritarianism/

[16] Brent Lee. 2005. Authenticity, Accuracy and Reliability: Reconciling Arts-related and Archival Literature. (2005).

[17] Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, Wim Hugo, Reyna Jenkyns, Varsha Khodiyar, Maryann E. Martone, Mustapha Mokrane, Vivek Navale, Jonathan Petters, Barbara Sierman, Dina V. Sokolova, Martina Stockhause, and John Westbrook. 2020. The TRUST Principles for digital repositories. *Sci. Data* 7, 1 (May 2020), 144. https://doi.org/10.1038/s41597-020-0486-7

[18] Lydia R. Lucchesi, Petra M. Kuhnert, Jenny L. Davis, and Lexing Xie. 2022. Smallset Timelines: A Visual Representation of Data Preprocessing Decisions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. ACM, Seoul Republic of Korea, 1136–1153. https://doi.org/10.1145/3531146.3533175

[19] H. MacNeil. 2013. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. Springer Science & Business Media.

[20] Reagan Moore. 2008. Towards a Theory of Digital Preservation. *Int. J. Digit. Curation* 3, 1 (August 2008), 63–75. https://doi.org/10.2218/ijdc.v3i1.42

[21] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*, 2022. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644

[22] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445402

[23] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19*), 2019. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[24] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. 2021. Advances in Neural Information Processing Systems.

[25] Alex H. Poole. 2015. How has your science data grown? Digital curation and the human factor: a critical literature review. *Arch. Sci.* 15, 2 (June 2015), 101–139. https://doi.org/10.1007/s10502-014-9236-y

[26] Andrea K. Thomer, Dharma Akmon, Jeremy J. York, Allison R. B. Tyler, Faye Polasek, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (2022), 414:1-414:29. https://doi.org/10.1145/3555139

[27] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18