

Category	Criteria	Sub-criteria	Guidance and Examples	
1) Dataset Properties	a) Following international standards and norms		For example, use ISO-3 codes for countries or ISO-8601 for timing data.	
	b) Semantic and logical consistency across entries		As 'heart attack' and 'cardiac arrest' are synonymous terms, only one should be used in a medical dataset. In domain-specific cases like this, labels should adhere to internationally recognised vocabularies such as ICD-10, SNOMED CT, or similar standards.	
	c) Identifiable class and source imbalance		CommonCorpus , an aggregate text dataset for AI training, clearly presents the source of each entry.	
	d) De-identification and Anonymisation where necessary		Transactions are anonymised with Principal Component Analysis in the Credit Card Fraud Detection dataset to ensure confidentiality without compromising quality.	
	e) Appropriate file format		The comma-separated value or .csv (particularly .csv on the Web, or CSVW) format is commonly preferred for structured datasets. Formats like Apache Parquet, or .rdf for graph-based data, also offer advantages. Selection should reflect the intended AI application and interoperability needs.	
2) Metadata	a) Machine-readable metadata format		Using the machine-readable Croissant metadata standard provides discoverability and interoperability for a dataset.	
	b) The dataset served to users with attached metadata		API queries for a dataset should return its metadata alongside it, as seen with the WorldPop API .	
	c) Basic technical specifications	i) Modalities		A dataset should clearly describe the data types (such as text, image, video, time series) within it.
		ii) Dimensionality		A user should know a dataset's number of rows and columns and any nested data or layering.

		iii) Update frequency	A dataset should clearly specify their update schedule, data refresh cycles, and change notification procedures.	
		iv) Semantics	Defining how columns and rows should be interpreted ensures users clearly understand the data, thereby using it better and more responsibly.	
		v) Bias	The Croissant Responsible AI extension allows authors to describe potential biases in their dataset.	
		vi) Basic summary statistics	Users appreciate descriptions of dataset column distribution, including calculations of averages, ranges and variances.	
		vii) Synthetic data	Any synthetic data-generation methods or machine annotation of data points should be demarcated.	
		d) Supply chain information	i) Collection	Ontologies like Prov-O or Croissant-RAI enable clear descriptions of datasets' provenance, including their collection, preprocessing (such as anonymisation or normalisation) and labelling.
			ii) Preprocessing	
iii) Annotation				
e) Legal and sociotechnical information	i) Context and original purpose	Clear documentation on why data was initially collected and the conditions under which it was gathered, in alignment with the EU AI Act, Article 10 . Dataset usage benefits when clear, machine-readable, statements regarding its socio-technical information, including the name of its licence, usage permissions, retention permissions and a URL link, are included in its attached machine-readable metadata. Statements on intended or permitted users, and notices about data protection, ensure AI practitioners have complete confidence in using a dataset.		
	ii) Licence name(s), permissions and URL(s)			
	iii) Intended access controls			
	iv) Data protection declaration(s)			

3) Surrounding Infrastructure	a) Accessibility via a user-centric data portal	Datasets should be sourced from a user-centric data portal like the data.europa.eu .
	b) Accessibility via API	RESTful API architectures are industry standard, mainly when exact dataset use cases vary or remain undefined.
	c) Version control and monitoring infrastructure	Dataset Version Control enables the tracking of a dataset's entire lifecycle, including post-publication. Integrated quality control processes enable monitoring for quality degradation, data drift, and emerging issues over the course of the dataset's lifecycle.
4) Governance	a) Governance policy-as-code	Key data governance policies—including data access, compliance checks, audit logging, and consent management—should be codified into machine-readable formats, such as ODRL , executable across governance tools and platforms.
	b) Documented roles and responsibilities	The dataset's documentation must clearly define and assign key governance roles with explicit accountability, such as a data owner or data steward.
	c) Publicly identifiable points of contact	Metadata should include contact information for the designated data steward, providing a feedback loop for users and a channel for reporting data quality or governance issues.
	d) Clear data access processes	The processes required for data users to access data must be clearly articulated, with the requirements, criteria and timelines described in full.