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) Semantics	Defining how columns and rows should be interpreted ensures users clearly understand the data, thereby using it better and more responsibly.
		iv) Bias	The Croissant Responsible AI extension allows authors to describe potential

			biases in their dataset.
		v) Basic summary statistics	Users appreciate descriptions of dataset column distribution, including calculations of averages, ranges and variances.
		vi) 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 and processing.
		ii) Preprocessing	
	e) Legal and sociotechnical information	i) Licence name(s) and URL(s)	Dataset usage benefits when clear statements regarding its socio-technical information, including the name of its licence and a URL link, are included in its attached metadata. Statements on intended or permitted users, and notices about data protection, ensure AI practitioners have complete confidence in using a dataset.
		ii) Intended access controls	
		iii) 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 European Data Portal.
	b) Accessibility via API		RESTful API architectures are industry standard, mainly when exact dataset use cases vary or remain undefined.
	c) Version control infrastructure		Dataset Version Control enables tracking a dataset's entire lifecycle, including post-publication.