

ICS 35.240.01

**Semantische Daten-Annotationen zur Unterstützung KI-gestützter
Datenverarbeitung;
Text Englisch**

Semantic Data Annotations to Support AI-enabled Data Processing;
Text in English

Annotations sémantiques des données pour le traitement des données assisté par l'IA;
Texte en anglais

Gesamtumfang 23 Seiten

Dieses Dokument wurde durch die im Vorwort genannten Verfasser erarbeitet und verabschiedet.

Contents

	Page
Foreword	3
Introduction	4
1 Scope	5
2 Normative references	5
3 Terms and definitions	5
4 Semantic data annotations	6
4.1 Classification criteria	6
4.2 List of data annotations	6
5 Annotation syntax and file formats	15
5.1 Annotation syntax	15
5.2 File formats for storing and exchanging annotations	15
6 Name spacing and extensibility	17
6.1 Global annotation namespace	17
6.2 Private annotation namespace	17
6.3 Vendor extensions	17
Annex A (informative) Best practices	18
A.1 Best practices for data annotators	18
A.2 Best practices for implementers	18
A.3 Best practices for extenders	18
Annex B (informative) JSON annotation example	19
Bibliography	23

Tables

Table 1 — Annotations for basic data properties	7
Table 2 — Annotations for basic statistical properties	8
Table 3 — Annotations for advanced statistical properties	8
Table 4 — Annotations for data processing	9
Table 5 — Annotations for process data	11
Table 6 — Annotations for surveys/opinion polling	13
Table 7 — Miscellaneous annotations	14

Foreword

This DIN SPEC has been developed according to the PAS procedure. The development of a DIN SPEC according to the PAS procedure is carried out in DIN SPEC-consortiums and does not require the participation of all stakeholders.

This document has been developed and adopted by the initiator(s) and authors named below:

- Inspirient GmbH
Dr. Guillaume Aimetti, Patrick Mertes, Dr. Georg Wittenburg
- Kantar Public
Johannes Huxoll, Martin Rathje
- Scheer PAS Deutschland GmbH
Dr. Christian Linn

At present, there are no standards covering this topic in the body of German Standards.

DIN SPECs are not part of the body of German Standards.

A draft of this DIN SPEC has not been published.

Despite great efforts to ensure the accuracy, reliability, and precision of technical and non-technical information, the DIN SPEC-consortium cannot give any explicit or implicit assurance or warranty in respect of the accuracy of the document. Users of this document are hereby made aware that the consortium cannot be held liable for any damage or loss. The application of this DIN SPEC does not release users from the responsibility for their own actions and is applied at their own risk.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. DIN shall not be held responsible for identifying any or all such patent rights.

Provision of this document free of charge as a PDF via the Beuth WebShop has been financed in advance.

For current information on this document, please go to the DIN website (www.din.de) and search for the document number in question.

Introduction

Data has been increasingly highlighted as the key resource of the 21st century. The relevance of value creation from data and data streams for both economical and social purposes can hardly be overestimated. Efficiently processing currently available as well as yet-to-be-created data is a key step towards realizing this value. The contribution of this document is to lay the groundwork for fully automated, i.e., autonomous, processing of structured datasets, so that the potential efficiency gains may be realized.

The exact understanding of a dataset is the key prerequisite for its correct processing, evaluation and analysis, e.g., in the context of economic or societal decision making. There are various and in part standardized approaches to document the properties of a dataset and thus to enable the necessary understanding of the data for each new use, including

- document meta-data information about the context of a dataset, including authorship/creator information, dates of creation/last modification, applicable keywords/ontologies, etc. [1];
- syntactic annotations that specify applicable/used data types, data ranges and formatting, e.g., number and currency formats for scientific or commercial use cases [2] [3] [4];
- content annotations and labels document the correct interpretation of individual data items for the purpose of Machine Learning (ML), e.g., annotations of images to train models for automatic classification.

A major gap exists in annotations that enable the correct semantic interpretation and subsequent processing/analysis of a dataset, especially if this is to be done automatically, i.e., without supporting interpretation and contextualization by a human subject matter expert. This aspect of the semantic and structural understanding of a dataset addresses key questions such as:

- Is a field an explicit or implicit ID value, e.g., a database primary key or a postal/ZIP code?
- Does an empty field in a column represent the absence of data ("null") or no countable activity (0)?
- Is it correct to sum the numeric values of a column, which is, e.g., true for price values in a table of sales transactions but not true for the price values in an inventory list?

These questions are currently left to the interpretation of the human subject matter expert. This is a cause for fuzziness in the evaluation of data, which a) can lead to errors in manual evaluation and b) hinder automated evaluation, e.g., by means of Artificial Intelligence methods.

1 Scope

This document comprises two main sections that address the following in-scope aspects of semantic data annotations to enable automated, AI-supported data processing:

- 1) definitions of key data annotations for structured datasets, both of generic nature and for two selected use cases, and including examples with regard to their respective applicability;
- 2) definitions of syntax and file formats to enable storing and processing data annotations.

Annex A also describes best practices when implementing or otherwise working with annotations as defined herein.

Out-of-scope topics of this document thus include

- data annotations used for training Machine Learning models;
- use case specific annotations beyond the two use cases included herein.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

DIN and DKE maintain terminology databases for use in standardization at the following addresses:

- DIN-TERMinologieportal: available at <https://www.din.de/go/din-term/>
- DKE-IEV: available at <https://www.dke.de/DKE-IEV>

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <https://www.electropedia.org/>

3.1

dataset

identifiable collection of data

[SOURCE: ISO 19115-1:2014, 4.3, modified — Note 1 to entry deleted]

3.2

structured dataset

identifiable collection of data with machine-adressable sub-elements

3.3

metadata

data about data or data elements, possibly including their data descriptions, and data about data ownership, access paths, access rights and data volatility

[SOURCE: ISO/IEC 2382:2015, 2121505, modified — Notes to entry deleted]

3.4

annotation

metadata associated with an item at a given instance in time

Note 1 to entry: The term “annotation” as used in this document is meant to be understood in the literal sense of the word, as opposed to data annotations as used in the field of Machine Learning (ML): The primary purpose of data annotations in this document is *not* to train ML-based classifiers, but rather to enable fully automated, intelligent processing of structured datasets.

[SOURCE: DIN EN IEC 62541-11:2021-08, 3.1.1, modified — Note 1 to entry deleted and replaced by a new Note 1 to entry]

4 Semantic data annotations

4.1 Classification criteria

There are four types of data annotations:

- **Filter** — Annotations that perform various filters on the input data
- **Transformation** — Annotations used to carry out a transformation on the input values
- **Semantic** — Annotations to explicitly communicate column meaning
- **Analysis** — Annotations that affect the analysis

Depending on the type, annotations can either denote a context-free property of the data (e.g., the fact that a dimension contains natural-language text), a context-dependent property of the data (e.g., being an independent variable in the context of the specific use case), or use-case-specific intended use of a dimension (e.g., a necessary filter on one of the dimensions that then enables the dataset to be processed correctly).

4.2 List of data annotations

The full list of supported annotations and their effects are listed in Table 1 to Table 7.

Effects of annotations are additive if multiple annotation are applied to the same column of a dataset. In case the effects of multiple annotations are mutually exclusive, the annotation that is applied later overrules any preceding annotations.

Some annotations also have a general negated version that causes the stated effect of the annotation *not* to be applied or considered when evaluating the dataset. For the negated version, the prefix **NOT_** (including the trailing underscore character) is prepended to the regular annotation. For example, the annotation **SUMMABLE** becomes **NOT_SUMMABLE** in its negated form. Annotations that support this form of general negation are marked with “(N)” in their type fields below.

Table 1 — Annotations for basic data properties

Annotation	Type	Description	Examples
DEFAULT_VALUE	Transformation	Specifies a default value for this column that is used whenever a value is not given, i.e., it is absent or given as “null”.	— DEFAULT_VALUE (“No answer given”) applied to a column of opinion polling answers inserts the phrase “No answer given” into all empty cells, to be used in any output of results
IGNORE_VALUE	Transformation	Ignore specified value(s), i.e., treat as absent or null	— IGNORE_VALUE (“John Doe”) applied to a column of names causes all cells that contain the string “John Doe” to be treated as empty
MORE_IS_BETTER	Semantic	Specifies that higher values in a numeric columns are “better” in the context of this analysis, i.e., for any optimization sets the aim to achieve higher values for this dimension (as opposed to lower values). See also LESS_IS_BETTER .	— MORE_IS_BETTER applied to a column of prices in the context of <i>selling</i> goods ensures that higher prices are preferred — MORE_IS_BETTER applied to Likert-scale survey responses ensures that the higher Likert-scale values are aggregated into the Top 2/Top 3 bucket (and the lower values into the Bottom 2/Bottom 3 bucket). This is also the default behavior
LESS_IS_BETTER	Semantic	Specifies that lower values in a numeric column are “better” in the context of this analysis, i.e., for any optimization sets the aim to achieve lower values for this dimension (as opposed to higher values). See also MORE_IS_BETTER .	— LESS_IS_BETTER applied to a column of prices in the context of <i>buying</i> goods ensures that lower prices are preferred — LESS_IS_BETTER applied to Likert-scale survey responses ensures that the lower Likert-scale values are aggregated into the Top 2/Top 3 bucket (and the higher values into the Bottom 2/Bottom 3 bucket)
HAS_SUBTOTALS	Semantic	Specifies that a numeric column contains subtotals, which implies that certain analyses are not permissible for this column, e.g., calculating sums and averages over all values (or even selections of values that still include the subtotals).	— HAS_SUBTOTALS applied to a revenue column on a business report that also aggregates revenues by business unit or geographies ensures that no sums/averages are incorrectly calculated that include these subtotals
ID	Semantic	Specifies that numeric values in a column represent ID values that can be used for aggregating other numeric values or joining data from another table but should not be summed or averaged.	— ID applied to a column that contains five-digit postcodes (as used in Germany) ensures that the analysis results may contain aggregations by postcode (as opposed to averaged or summed postcodes)
NATURAL_LANGUAGE_TEXT	Semantic	Text values in this column should be treated as natural language text for semantic text analysis.	— NATURAL_LANGUAGE_TEXT applied to a column of open answers in a survey table will ensure that key phrases are extracted from each text segment for further statistical analysis