Rillor

EXPLAINER

Data labeling vs data annotation: label types for text and documents

How standards and tools use the two words, and the label types for text and documents: classes, entities and spans, relationships, ratings, page regions and tables, with their storage formats.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Data labeling and data annotation name overlapping work, and the label type sets how labels are stored and how their quality is measured.

- The sources reviewed use the two words in overlapping senses. Standards and glossaries treat annotation as attaching information to data and a label as the value a model learns to predict [1, 3, 4], and several vendor guides and tools use the two words for the same work [6, 12].
- The EU AI Act lists annotation and labelling as separate data-preparation operations in Article 10 and defines neither word [2]. No source reviewed fixes one industry-wide difference between data labeling and data annotation.
- Text label types fall into four groups: classes on a whole record, entities and spans inside it, links between its parts (relations, coreference between mentions of the same entity, and events), and ratings or preferences on a response [11, 13, 28].
- Span labels differ in their overlap rules. CoNLL-2003 assumed non-overlapping entities and usually annotated only the top-level one [18], while roughly 17% of entities in the GENIA corpus sit inside another entity [19].
- A relation label set needs an explicit label for no relation. In TACRED, 78.7% of the 119,474 examples carry it [22].
- Relations often cross sentence boundaries. In DocRED, at least 40.7% of relational facts can only be extracted from multiple sentences [23].
- Document labels add position: page regions as bounding boxes [31], form fields linked from question to answer [30] and tables labeled as HTML structure [33].
- The label type also sets the quality measure: exact-match scores for entities [18], an agreement coefficient for span labeling tasks [8], box overlap for page regions [31] and tree edit distance for tables [33].

Background

Two words for one family of tasks

Supervised learning trains on labeled examples. Google's machine learning glossary defines a label as "the 'answer' or 'result' portion of an example", and a labeled example as one that "contains one or more features and a label" [4]. Annotation is the wider word. In corpus linguistics it means adding interpretative information to text, with word-class tags as a common case [5]. On the web, the W3C standardized annotation in 2017 as a body of information related to a target resource [1].

Why the label type matters

A schema has to say what is labeled, in which unit and with which values. A text with one class, a contract with highlighted clauses, two mentions linked as one entity, a ranked set of model responses and a page with boxed regions are all labeling, and their sources give each its own rules, storage format and quality measure [11, 13, 28, 31, 35]. Klie, Eckart de Castilho and Gurevych separate the two documents a labeling task rests on. The annotation scheme "defines the structure, features, and tagsets of the task", and the guidelines "describe how to decide when to apply which label" [8].

This report is for ML engineers, data owners and project leads who write a labeling specification for text or documents, or who buy labeled data and need to read one.

Method

What was examined. How standards, laws, glossaries and tool documentation use "labeling" and "annotation". Then five families of label types for text and documents: classes, entities and spans, relations with coreference and events, ratings and preferences, and document labels (page regions, form fields, tables and question-answer pairs). For each: the defining source, the unit labeled, the values, the storage format and the quality measure its source uses.

Sources. Thirty-eight sources, each opened on 8 November 2026: a W3C Recommendation [1], the EU AI Act on EUR-Lex [2], a page quoting ISO/IEC 22989:2022 [3], glossaries and explainers from three technology companies [4, 6, 7], a corpus linguistics guide [5], a study of annotation quality management [8], tool documentation for SageMaker Ground Truth, Azure Machine Learning, Label Studio, brat, spaCy, Universal Dependencies and TRL [9, 10, 11, 12, 13, 14, 15, 16, 29, 37, 38], and papers or catalogue pages for 19 datasets and methods [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36].

How. Definitions and figures were read in each paper's full text, except DocVQA and PubTabNet, which were read on their abstract pages [33, 34]. Dataset counts come from the papers' own tables. Figures computed for this report are marked "computed here". Tools and datasets are subjects studied, and none is ranked.

Tool status. Two documentation pages carried notices when viewed. SageMaker Ground Truth was "no longer open to new customers" [9], and Azure Machine Learning's text labeling page gave 30 September 2026 as the retirement date of its data labeling [11]. Both are used here only for how they define and store label types.

Findings

1. How the two words are defined

Standards. The W3C Web Annotation Data Model, a Recommendation of 23 February 2017, considers an annotation "a set of connected resources, typically including a body and target", which "conveys that the body is related to the target" [1]. The model lists 13 motivations (count computed here from the list), among them classifying ("to classify an image as a portrait"), identifying, tagging, assessing, commenting and linking, and says there "SHOULD be exactly 1 motivation for each Annotation" [1]. In this model a class label is one kind of annotation.

ISO/IEC 22989:2022, as quoted by itSMF Switzerland, defines data annotation in clause 3.2.1 as the "process of attaching a set of descriptive information to data without any change to that data" [3]. A note adds that the information "can take the form of metadata, labels and anchors" [3]. Clause 3.2.10 defines a label as the "target variable assigned to a sample" [3]. Read together, the two clauses make a label one form of annotation, the value a model is trained to predict [3, 4].

Corpus linguistics. Geoffrey Leech's chapter in a 2004 guide to developing linguistic corpora defines corpus annotation as "the practice of adding interpretative linguistic information to a corpus" [5]. It gives as one common type "the addition of tags, or labels, indicating the word class to which words in a text belong" [5]. The chapter lists other levels of annotation: phonetic, prosodic, syntactic, semantic, pragmatic, discourse, stylistic and lexical, with discourse annotation covering anaphoric links [5].

Glossaries and explainers. Google's glossary, last updated on 6 November 2026, has no entry for annotation [4]. The related word annotator appears in two entries: inter-rater agreement is "also sometimes called inter-annotator agreement", and "'Annotator' is another name for rater" [4]. IBM's explainer opens with "Data labeling, or data annotation, is part of the preprocessing stage when developing a machine learning (ML) model" and treats the two as one [6]. AWS's explainer defines data labeling as identifying raw data and "adding one or more meaningful and informative labels" [7]. In its passage on consensus, it calls the separate responses for one item "annotations" and consolidates them "into a single label" [7].

Tools. Label Studio's export schema describes its annotations array as "Array containing the labeling results for the task" and annotations.result as the results "of the labeling or annotation task" [12]. SageMaker Ground Truth's output for a named entity recognition labeling job holds its labels and entities inside an "annotations" object [10]. In both tools the two words name the same records.

Law. In the EU AI Act (Regulation (EU) 2024/1689, published 12 July 2024), Article 10 says the data governance practices for training, validation and testing data concern "relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation" [2]. Annex IV asks the technical documentation to describe "labelling procedures (e.g. for supervised learning)" [2]. Article 3 defines neither word [2]. The law also uses labelling in a second sense, for marking generated content as "artificially created or manipulated" [2].

Reading. Where a source defines both words, annotation is the wider term and a label is one form of it, the value a model learns [3, 4, 5]. Where a tool or guide uses both, they often name the same work [6, 10, 12]. AWS's split, with annotations as individual responses and the label as their consolidated result, is a third use [7]. No source reviewed sets an industry-wide distinction, so a project's schema has to define its own terms.

2. Classes on the whole record

A class label assigns a whole item to a category from a fixed set. Google's glossary says a classification model "predicts a class", and its multi-class example chooses among poodle, beagle and pug [4].

Tool documentation fixes three shapes for text. Azure Machine Learning offers multi-class "to apply only a single label to each piece of text from a set of labels", multi-label "to apply one or more labels to each piece of text", and named entity recognition for labels on "individual text words or to multiple text words in each entry" [11]. Each .txt file "represents one item to be labeled", and in a .csv or .tsv file each row is one item [11]. SageMaker Ground Truth lists the same three text tasks: single-label classification, multi-label classification and named entity recognition [37].

Multi-label sets. GoEmotions labels about 58,000 English comments for "one or more of 27 emotion(s) or Neutral", and its classification experiments use "a sigmoid cross entropy loss function to support multi-label classification" [26]. Neutral is a value of its own beside the 27 emotions [26].

Classes on pairs. SNLI labels 570,152 sentence pairs as entailment, contradiction or neutral [25]. Its validation phase "collected four additional judgments for each label for 56,941 of the examples" [25]. In 98% of them a label reached three of five votes, and 58% were unanimous [25]. Where no label had three votes, in about 2% of cases, the pair received the placeholder label "-" [25]. The item here is a pair of texts, and the schema includes a value for no agreed label.

Classes on tokens. Universal Dependencies defines 17 universal part-of-speech tags, from ADJ to X (count computed here from the list) [16]. Each token gets a tag from that set, which makes the task a class label at the token level.

Tagset size. Klie and colleagues note that percent agreement "is influenced by the size of the tagset. Therefore, it is difficult to compare across annotation schemes" [8]. Agreement on a two-class schema and on GoEmotions' 28 values (27 emotions and Neutral) does not compare directly [8, 26].

3. Entities and spans inside the record

Entities. SageMaker's guide describes named entity recognition as locating "noun phrases, called named entities", and categorizing each with a label such as person or organization. It adds that the task can be broadened "to label longer spans of text" with predefined labels [9]. brat defines a text-bound annotation as one that "identifies a specific span of text and assigns it a type" [13].

CoNLL-2003 used four types: "persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC)" [18]. Its English training set holds 946 articles, 14,987 sentences and 203,621 tokens, with 7,140 location, 3,438 miscellaneous, 6,321 organization and 6,600 person entities [18]. OntoNotes, as used in the CoNLL-2012 task, has "a set of 18 well-defined proper named entity types" [21].

Overlap and nesting. CoNLL-2003 states: "We assume that named entities are non-recursive and non-overlapping. When a named entity is embedded in another named entity, usually only the top level entity has been annotated" [18]. Finkel and Manning write that "Most corpus designers have chosen to skirt the issue entirely, and have annotated only the topmost entities", and name CoNLL, MUC-6 and MUC-7 as flatly annotated [19]. In the GENIA corpus, labeled with types such as protein and DNA, roughly 17% of entities are embedded within another entity, and in the AnCora corpus of Spanish and Catalan news text nearly half are [19]. A flat schema leaves the inner entity unlabeled, so the nesting policy changes what the data holds [18, 19].

spaCy draws the same line in its components. Its entity recognizer "identifies non-overlapping labelled spans of tokens", and training data with overlapping spans raises an error [15]. Its span categorizer is for "labeling potentially overlapping spans of text": a suggester proposes candidate spans and a labeler "predicts zero or more labels for each candidate" [14].

Unitizing. Klie and colleagues describe span tasks as unitized: "annotators first divide the instances into smaller units and only then assign labels", as when "marking spans that contain entities" [8]. The span boundary is itself part of the label.

Long spans. CUAD, a contract review dataset, contains "510 contracts and 13101 labeled clauses" in "41 label categories" [35]. Labeled clauses make up about 10% of each contract on average, so only about 0.25% of each contract is highlighted for each label [35]. Clause spans are long and sparse.

Attributes and normalization. brat adds two layers on top of spans. Attributes are "binary or multi-valued 'flags' that specify further aspects of other annotations", with Negation and Confidence as examples [13]. Normalization links an annotation to an entry in an external resource, written as a resource ID and entry ID pair [13].

4. Relations, coreference and events between parts

Relations. In brat, binary relations "are defined by their type (e.g. Origin, Part-of) and their arguments", and unlike events they do not identify a trigger, the text expressing the relation [13]. TACRED is "a large (119,474 examples) supervised relation extraction dataset" with 42 relation labels, counting no relation among them, and 41 slot types [22]. Its authors "fully annotate all negative instances", and 78.7% of its examples are negative [22].

Document-level relations. DocRED's annotated set, as counted in its Table 1, has 5,053 documents, 40,276 sentences, 132,375 entities, 96 relation types, 63,427 relation instances and 56,354 relational facts [23]. At least 40.7% of its relational facts "can only be extracted from multiple sentences" [23]. Its annotation begins

with "Named Entity and Coreference Annotation" before relations are labeled [23]. The mentions of each entity are linked before relations across sentences are labeled [23].

Coreference. The CoNLL-2012 task was "to automatically identify mentions of entities and events in text and to link the coreferring mentions together to form entity/event chains" [21]. Its coreference layer "does not contain any singleton mentions", so a mention with no partner carries no coreference label [21]. In brat, equivalence relations "define sets of annotations to be equivalent in some sense (e.g. referring to the same real-world entity)" [13].

Events. brat defines an event "by type (e.g. MERGE-ORG), event trigger (the text stating the event) and arguments" [13]. An event label therefore holds a span for the trigger and links to its arguments. ACE 2005 contains "approximately 1,800 files of mixed genre text in English, Arabic, and Chinese annotated for entities, relations, and events" [24].

Layers on one text. OntoNotes Release 5.0, released on 16 October 2013, annotates text in English, Chinese and Arabic with syntax, predicate argument structure, word sense linked to an ontology and coreference, in "2.9 million words" [20]. One corpus carries several label types on the same text [20].

5. Ratings and preferences on a response

Ratings. The W3C model lists assessing as a motivation, "for example to write a review or assessment of a book, assess the quality of a dataset" [1]. SummEval rated summaries "on a Likert scale from 1 to 5 (higher better)" for coherence, consistency, fluency and relevance [27]. It rated summaries for 100 articles from a news summarization test set, "amounting to 12800 summary-level annotations" [27]. A rating schema names each dimension and the scale.

Rankings and preferences. In the instruction-following study by Ouyang and colleagues, responses were ranked in sets of "anywhere between K = 4 and K = 9", which "produces (K choose 2) comparisons for each prompt" [28]. A set of 4 yields 6 comparisons and a set of 9 yields 36 (computed here). Ties were dropped [28]. For evaluation, the overall quality of each response was judged "on a 1-7 Likert scale", with binary labels such as "Fails to follow the correct instruction / task" and "Hallucination" [28]. The study's data hold three label types on responses: ranks, ratings and binary flags [28].

Storage. Hugging Face's TRL documentation defines a preference dataset as one that "includes a 'prompt', a 'chosen' completion, and a 'rejected' completion", where the model "is trained to select the 'chosen' response over the 'rejected' response" [29]. Its example is {"prompt": "The sky is", "chosen": " blue.", "rejected": " green."} [29]. An unpaired preference stores one completion with a label of true or false [29]. The documentation recommends "using explicit prompts whenever possible" [29]. A ranking of K responses becomes pairs to fit the paired format, as the study's K choose 2 comparisons show [28, 29].

6. Document label types

Page regions. DocLayNet contains 80,863 annotated pages with "labelled bounding-boxes with a choice of 11 distinct classes": Caption, Footnote, Formula, List-item, Page-footer, Page-header, Picture, Section-header, Table, Text and Title [31]. Its authors compare that with 5 labels in PubLayNet and 13 in DocBank [31]. Of its pages, 7,059 carry two annotations and 1,591 carry three [31]. Scanned documents were "excluded wherever possible because they can be rotated or skewed", which would not allow annotation with rectangular bounding boxes [31]. Google's glossary describes intersection over union (IoU) as the measure of a predicted bounding box's accuracy "with respect to the ground-truth bounding box" [4].

Labels from source markup. PubLayNet was built "by automatically matching the XML representations and the content of over 1 million PDF articles that are publicly available on PubMed Central" [32]. It contains "over 360 thousand document images, where typical document layout elements are annotated", such as text, title, list, figure and table [32]. Where a document's source markup exists, labels can be derived from it [32, 33].

Form fields. FUNSD holds "199 real, fully annotated, scanned forms" [30]. Each semantic entity has "a unique identifier, a label (i.e., question, answer, header or other), a bounding box, a list of links with other entities, and a list of words" [30]. Links "are directed and formatted as [id_from, id_to]" [30]. The training set holds 149 forms with 22,512 words, 7,411 entities and 4,236 relations, and the test set 50 forms with 8,973 words, 2,332 entities and 1,076 relations [30]. Its training entities split into 441 header, 3,266 question, 2,802 answer and 902 other [30]. One form label combines a box, a class and a link.

Tables. PubTabNet pairs "568k table images with corresponding structured HTML representation", generated by matching the XML and PDF versions of articles in the PubMed Central Open Access Subset [33]. Its authors propose "a new Tree-Edit-Distance-based Similarity (TEDS) metric for table recognition" [33]. A table label here is the table's HTML structure.

Question-answer pairs. DocVQA has "50,000 questions defined on 12,000+ document images", with human performance of "94.36% accuracy" [34]. The label here is a question and its answer on a page image.

7. Storage formats and offsets

Standoff. In brat, "annotations are stored separately from the annotated document text, which is never modified by the tool", one .ann file per .txt file [13]. Spans are tied to text "through character offsets", all "indexed from 0", including the character at the start offset and excluding the character at the end offset [13]. Each annotation has an ID whose prefix gives its kind: T for a text-bound annotation, R relation, E event, A attribute, M modification, N normalization and # note [13]. The W3C Text Position Selector uses the same counting: "The first character in the full text is character position 0", and the end character "is not included within the segment" [1]. Its Text Quote Selector describes a segment by quoting it, plus passages before or after it, in its exact, prefix and suffix properties [1].

Offsets in JSON. SageMaker Ground Truth gives each entity "a label value that matches one in the labels list, an integer startOffset value for labeled span's starting Unicode offset, and an integer endOffset value for the ending Unicode offset" [10]. Its example entity reads {"label": "Date", "startOffset": 75, "endOffset": 88} [10].

Token columns. CoNLL-2003 data has one word per line in four fields: "the word, its part-of-speech tag, its chunk tag and its named entity tag" [18]. Words outside entities are tagged O, words inside an entity of type XXX are tagged I-XXX, and B-XXX marks the first word of a second adjacent entity of the same type [18]. The scheme is "the IOB scheme originally put forward by Ramshaw and Marcus (1995)" [18]. Ramshaw and Marcus used "the chunk tag set {I, O, B}" for base noun phrases [17]. CoNLL-U, the Universal Dependencies format, has word lines "in 10 fields separated by single tab characters" [38]. spaCy exposes the same encoding on tokens as `ent_iob_`, "The IOB part of the named entity tag" [15]. CoNLL-2003 pairs its single entity column with the assumption that entities do not overlap [18].

Tool exports. Label Studio's export formats include COCO, CoNLL2003, CSV, JSON, Pascal VOC XML, TSV and YOLO, among others. For spaCy it has no direct export: a CoNLL2003 export is converted with spacy convert [12]. "When a prediction is used to create an annotation, the result IDs stay the same in the annotation field" [12], so a model's proposed region and its reviewed version share an ID.

Format	Where labels sit	Span position	Also holds
brat standoff	A separate .ann file per .txt file	Character offsets from 0, end excluded	Relations, events, attributes, normalizations, notes [13]
W3C Web Annotation	An annotation linked to its target	Character positions from 0, end excluded, or a quote with prefix and suffix	Usually one motivation per annotation (SHOULD), from 13 [1]
SageMaker NER output	A JSON "annotations" object	Unicode startOffset and endOffset	The job's label list [10]
CoNLL-2003	One line per word in the data file	Tokens, with IOB tags	Part-of-speech and chunk tags [18]
CoNLL-U	One line per word	Tokens	10 tab-separated fields per word [38]
TRL preference	One JSON record per prompt	None, whole completions	Chosen and rejected completions [29]

Table 1. Storage patterns in the sources reviewed.

A span delivered as character offsets and the same span delivered as IOB tags describe one label in two units, characters and tokens [13, 18]. The schema records which one a delivery uses.

8. How the label type changes the quality measure

Entities. CoNLL-2003 counts "a named entity is correct only if it is an exact match of the corresponding entity in the data file", and scores systems with the F measure at beta = 1 over precision and recall [18]. spaCy's documentation notes that its entity recognizer's loss "optimizes for whole entity accuracy", and that low inter-annotator agreement on boundary tokens makes the component likely to perform poorly [15]. The boundaries are part of what has to agree.

Spans. Klie and colleagues write that "span labeling tasks like named entity recognition or relation extraction can be evaluated using a coefficient of the Krippendorff's unitized α (αu) family" [8]. Hripcsak and Rothschild note that studies that involve "searching the Internet or marking phrases usually lack a well-defined number of negative cases", which "prevents the use of traditional interrater reliability metrics like the κ statistic" [36]. They show that the average pairwise F-measure equals the average positive specific agreement, and that κ approaches these measures as the number of negative cases grows large [36].

Classes. Percent agreement depends on tagset size and is hard to compare across schemes [8]. SNLI's gold labels come from a three-vote majority, with 58% of validated pairs unanimous [25].

Relations. A model that predicts no relation for every TACRED example matches 78.7% of them (computed here from [22]). The negative share is the baseline a relation accuracy figure starts from.

Page regions. DocLayNet computes agreement on its triple-annotated pages as "the `mAP@0.5-0.95` metric between pairwise annotations" [31]. Agreement is 82 to 83 overall and 60 to 72 for Title, and the models trained on it "fall approximately 10% behind the inter-annotator agreement" [31]. The agreement figure varies by class.

Tables. PubTabNet scores table recognition with TEDS, a similarity based on tree edit distance [33]. The score compares table structures as trees.

Label type	Unit	Example source	Measure named in the source
Class	Whole record or pair of texts	SNLI	Majority label, percent agreement by tagset size [8, 25]
Entity	Span, non-overlapping	CoNLL-2003	Exact-match F1 [18]
Span	Span, unitized	Klie and colleagues, Hripcsak and Rothschild	Unitized alpha family, pairwise F-measure [8, 36]

Label type	Unit	Example source	Measure named in the source
Relation	Pair of entities	TACRED	Micro-averaged precision, recall and F1, against a 78.7% no-relation share [22]
Page region	Bounding box	DocLayNet	`mAP@0.5-0.95` between annotations [31]
Bounding box	Box against a ground-truth box	Google's glossary	Intersection over union [4]
Table	Table structure	PubTabNet	TEDS [33]

Table 2. Label types, units and the quality measures their sources name.

The coefficients for each annotation design, with worked figures, are set out in measuring label quality with inter-annotator agreement (https://rillor.com/insights/measuring-label-quality-inter-annotator-agreement).

Implications

For teams writing a labeling specification. Define the terms first, since the sources give "annotation" and "label" several senses [1, 3, 6, 7]. Then name the label type, the unit and the value set. The scheme defines the labels and the guidelines define when each applies [8]. For spans, state whether they may overlap or nest [14, 15, 18, 19]. For relations, include a no-relation label and say whether negatives are fully labeled, as TACRED does [22]. For preferences, say how ties are handled and whether rankings become pairs [28, 29]. For document regions, state the box type and the classes, and note that DocLayNet excluded scanned documents wherever possible, because they can be rotated or skewed and rectangular boxes would not fit them [31].

For buyers of labeled data. Ask for the label type and unit of every field, the offset convention (characters or tokens, the start index, end included or excluded) [1, 13, 18], the export format [12], the share of negative and placeholder labels [22, 25] and the quality measure used for each label type [8, 18, 31, 33]. A single agreement figure across several label types hides differences by type and by class [8, 31].

For document projects. A document label often combines text, position and links, as FUNSD's question-to-answer links show [30]. Where the source markup of a document exists, labels can be derived from it [32, 33]. Tables need a structure label and a structure score [33].

For anyone reading a regulation or glossary. The EU AI Act names annotation and labelling as separate operations and asks for labelling procedures in technical documentation [2]. Since the law defines neither word [2], a project's own documentation is where the terms get their meaning.

Limits of this analysis

- ISO/IEC 22989:2022 wording is taken from a page that quotes it [3]. The ISO catalogue page answered with a bot challenge and was not read.
- DocVQA and PubTabNet were read on their abstract pages only [33, 34]. Nothing beyond their abstracts is claimed.
- Tool documentation can change after the viewing date. Two of the tools carried service notices when viewed on 8 November 2026 [9, 11].
- The papers reviewed date from 1995 to 2024 [8, 17]. They are used for the structure of their labels, never as current accuracy levels.
- This report covers text and documents. Image, audio and video label types appear only where documents need them.
- No source reviewed gives the cost or time of labeling by label type, so none is stated.

Sources

1. Web Annotation Data Model, W3C Recommendation. Robert Sanderson, Paolo Ciccarese and Benjamin Young (editors), World Wide Web Consortium, 23 February 2017, viewed 8 November 2026. https://www.w3.org/TR/annotation-model/

2. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). European Parliament and Council of the European Union, Official Journal of the European Union, 12 July 2024, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

3. Understanding AI terms and definitions according to the standard ISO 22989. itSMF Switzerland, 19 February 2025, viewed 8 November 2026. https://www.itsmf.ch/blogs/post/understanding-ai-terms-and-definitions-according-to-the-standard-iso-22989

4. Machine Learning Glossary. Google for Developers, last updated 6 November 2026, viewed 8 November 2026. https://developers.google.com/machine-learning/glossary

5. Adding Linguistic Annotation. Geoffrey Leech, in Developing Linguistic Corpora: a Guide to Good Practice, ed. Martin Wynne, Arts and Humanities Data Service, 2004, viewed 8 November 2026. https://users.ox.ac.uk/~martinw/dlc/chapter2.htm

6. What is data labeling? IBM Think, IBM, 28 September 2021, updated 23 January 2026, viewed 8 November 2026. https://www.ibm.com/think/topics/data-labeling

7. What is Data Labeling? Amazon Web Services, undated, viewed 8 November 2026. https://aws.amazon.com/what-is/data-labeling/

8. Analyzing Dataset Annotation Quality Management in the Wild. Jan-Christoph Klie, Richard Eckart de Castilho and Iryna Gurevych, Computational Linguistics 50(3), pp. 817-866, MIT Press, September 2024, viewed 8 November 2026. https://aclanthology.org/2024.cl-3.1/

9. Extract text information using named entity recognition. Amazon SageMaker Developer Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/sagemaker/latest/dg/sms-named-entity-recg.html

10. Labeling job output data. Amazon SageMaker Developer Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/sagemaker/latest/dg/sms-data-output.html

11. Set up a text labeling project and export labels. Azure Machine Learning documentation, Microsoft Learn, last updated 31 December 2025, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-create-text-labeling-projects?view=azureml-api-2

12. Export annotations and data from Label Studio. Label Studio documentation, HumanSignal, undated, viewed 8 November 2026. https://labelstud.io/guide/export

13. brat standoff format. brat rapid annotation tool, undated, viewed 8 November 2026. https://brat.nlplab.org/standoff.html

14. SpanCategorizer. spaCy API documentation, Explosion, undated, viewed 8 November 2026.

https://spacy.io/api/spancategorizer

15. EntityRecognizer. spaCy API documentation, Explosion, undated, viewed 8 November 2026. https://spacy.io/api/entityrecognizer

16. Universal POS tags. Universal Dependencies contributors, undated, viewed 8 November 2026. https://universaldependencies.org/u/pos/index.html

17. Text Chunking using Transformation-Based Learning. Lance Ramshaw and Mitch Marcus, Third Workshop on Very Large Corpora, ACL Anthology W95-0107, 1995, viewed 8 November 2026. https://aclanthology.org/W95-0107/

18. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Erik F. Tjong Kim Sang and Fien De Meulder, Proceedings of CoNLL at HLT-NAACL 2003, pp. 142-147, 2003, viewed 8 November 2026. https://aclanthology.org/W03-0419/

19. Nested Named Entity Recognition. Jenny Rose Finkel and Christopher D. Manning, Proceedings of EMNLP 2009, pp. 141-150, August 2009, viewed 8 November 2026. https://aclanthology.org/D09-1015/

20. OntoNotes Release 5.0, LDC2013T19. Linguistic Data Consortium, 16 October 2013, viewed 8 November 2026. https://catalog.ldc.upenn.edu/LDC2013T19

21. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina and Yuchen Zhang, Joint Conference on EMNLP and CoNLL, Shared Task, pp. 1-40, July 2012, viewed 8 November 2026. https://aclanthology.org/W12-4501/

22. Position-aware Attention and Supervised Data Improve Slot Filling. Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli and Christopher D. Manning, Proceedings of EMNLP 2017, pp. 35-45, September 2017, viewed 8 November 2026. https://aclanthology.org/D17-1004/

23. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. Yuan Yao and others, Proceedings of ACL 2019, pp. 764-777, July 2019, viewed 8 November 2026. https://aclanthology.org/P19-1074/

24. ACE 2005 Multilingual Training Corpus, LDC2006T06. Christopher Walker and others, Linguistic Data Consortium, 15 February 2006, viewed 8 November 2026. https://catalog.ldc.upenn.edu/LDC2006T06

25. A large annotated corpus for learning natural language inference. Samuel R. Bowman, Gabor Angeli, Christopher Potts and Christopher D. Manning, Proceedings of EMNLP 2015, pp. 632-642, September 2015, viewed 8 November 2026. https://aclanthology.org/D15-1075/

26. GoEmotions: A Dataset of Fine-Grained Emotions. Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade and Sujith Ravi, Proceedings of ACL 2020, pp. 4040-4054, July 2020, viewed 8 November 2026. https://aclanthology.org/2020.acl-main.372/

27. SummEval: Re-evaluating Summarization Evaluation. Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher and Dragomir Radev, Transactions of the Association for Computational Linguistics 9, pp. 391-409, 2021, viewed 8 November 2026. https://aclanthology.org/2021.tacl-1.24/

28. Training language models to follow instructions with human feedback. Long Ouyang and others, arXiv 2203.02155, version 1, 4 March 2022, viewed 8 November 2026. https://arxiv.org/abs/2203.02155

29. Dataset formats and types. TRL documentation, Hugging Face, undated, viewed 8 November 2026. https://huggingface.co/docs/trl/dataset_formats

30. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. Guillaume Jaume, Hazim Kemal Ekenel and Jean-Philippe Thiran, arXiv 1905.13538, version 2, 29 October 2019, viewed 8 November 2026. https://arxiv.org/abs/1905.13538

31. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar and Peter W J Staar, KDD 2022, arXiv 2206.01062, version 1, 2 June 2022, viewed 8 November 2026. https://arxiv.org/abs/2206.01062

32. PubLayNet: largest dataset ever for document layout analysis. Xu Zhong, Jianbin Tang and Antonio Jimeno Yepes, arXiv 1908.07836, version 1, 16 August 2019, viewed 8 November 2026. https://arxiv.org/abs/1908.07836

33. Image-based table recognition: data, model, and evaluation. Xu Zhong, Elaheh ShafieiBavani and Antonio Jimeno Yepes, arXiv 1911.10683, version 5, 4 March 2020, viewed 8 November 2026. https://arxiv.org/abs/1911.10683

34. DocVQA: A Dataset for VQA on Document Images. Minesh Mathew, Dimosthenis Karatzas and C.V. Jawahar, WACV 2021, arXiv 2007.00398, version 3, 5 January 2021, viewed 8 November 2026. https://arxiv.org/abs/2007.00398

35. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. Dan Hendrycks, Collin Burns, Anya Chen and Spencer Ball, NeurIPS 2021 Datasets and Benchmarks, arXiv 2103.06268, version 2, 8 November 2021, viewed 8 November 2026. https://arxiv.org/abs/2103.06268

36. Agreement, the F-Measure, and Reliability in Information Retrieval. George Hripcsak and Adam S. Rothschild, Journal of the American Medical Informatics Association 12(3), pp. 296-298, May-June 2005, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC1090460/

37. Built-in Task Types. Amazon SageMaker Developer Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/sagemaker/latest/dg/sms-task-types.html

38. CoNLL-U Format. Universal Dependencies contributors, undated, viewed 8 November 2026. https://universaldependencies.org/format.html

About this report

Published by Rillor on 11 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Data labeling vs data annotation: label types for text and documents. 11 November 2026. https://rillor.com/insights/data-labeling-vs-data-annotation

Work with Rillor

Send a sample of the records, the labels your schema sets and the accuracy you need. Contact us at rillor.com/contact?type=labeling or contact@rillor.com. More on labeling and annotation: rillor.com/data#labeling