

5 Ws Crime News Annotation Guidelines

VERSION 2.1

02-05-2023

Giovanni Bonisoli*, Maria Pia di Buono^,
Laura Po*, Federica Rollo*

*Università degli studi di Modena e Reggio Emilia

^Università di Napoli L'Orientale

1. Goals	3
2. Competency questions	3
3. Annotation	3
3.1. News and events	3
3.2. Span	5
3.3. Synonyms and specific terms	5
4. Multiple annotations	6
4.1 Same label/different spans	6
4.2 Same span/different labels	6
5. Relations	6
6. Stolen object (OBJ)	7
7. Author and group of authors	8
7.1. Author (AUT)	8
7.2. Group of Authors (AUTG)	10
8. Victim, group of victim and injured party	10
8.1. Victim (VIC)	11
8.2. Group of Victims (VICG)	11
8.3. Injured party (PAR)	12
9. Location (LOC)	12
10. Time	13
11. How	13
Appendix A - Doccano	13

1. Goals

These guidelines describe the schema used to annotate texts of crime news articles in order to highlight the elements that characterize the described event.

To the aim of these guidelines we refer to the ITALIAN CRIME NEWS dataset.¹

The main goal is annotating journalistic 5Ws + 1H (What, Where Who, When, Why).

2. Competency questions

In order to identify the crime events described within news and annotate the answers to the 5Ws, we consider four questions and define at least a category to represent the answers to each of them. At the time being, we focus mainly on thefts events, as it follows:

1. Q: What has been stolen?
A: **STOLEN OBJECT**
2. Q: Who committed the theft?
A: **AUTHOR** and/or **GROUP OF AUTHORS**
3. Q: Who was the victim?
A: **VICTIM**, **GROUP OF VICTIMS** and/or **INJURED PARTY**
4. Q: Where did the event happen?
A: **LOCATION**

Answers to these questions should be identified within the news text. Answers can be represented by one or more linguistic elements, that are spans.

We should also consider that sometimes answers cannot be identified in the news, as it does not report all the information suitable to answer the questions.

Annotators should read the whole news article before proceeding with the annotation, in order to identify carefully spans that should be annotated within the text.

3. Annotation

3.1. News and events

News articles can differ in their content with reference to the number of events described within the text. We distinguish two main cases: single and multiple events.

Single event → We define a **single event** as a theft event happening in one location, perpetrated by one or more authors during a specific time.

We annotate just news which report one single event.

Example: (same location, same author, same time, two victims, one theft action)

The thief who entered the apartment stole a computer, a TV, and 200 euros. The owners, Mario Rossi and Anna Verdi, reported the theft.

¹ <https://paperswithcode.com/dataset/italian-crime-news>

Example: (same location, same author, same moment, two different stolen objects and two different victims, one theft action)

The thief who entered the apartment stole a computer belonging to a Spanish student and an iPhone belonging to the homeowner, Luisa Rossi.

Multiple events

We distinguish two types of multiple events, that are unrelated and related events.

- Different unrelated events are described within one news and are characterized by the presence of different authors/victims and happen in different locations/time.

Example:

It was shortly after dinner time when a Ford X car was stolen in the parking lot of the shopping center. For months, the area has been plagued by several car and bike thefts. Two weeks ago, there was the theft of a Fiat Panda, early in the morning... The residents of the area are worried... The patrols in the area are being intensified... A month ago, there were already several complaints and three reports from residents of the neighborhood: the theft of a van, a purse snatching from a lady on a bicycle, and an apartment being ransacked within minutes.

- Different related events describe multiple theft events that share one or more elements, i.e. the same author, the same stolen object, the same location, the same time, the same victim(s) but differ in one of them.

Example:

Yesterday at the shopping center, a man stole a cellphone from a 30-year-old lady, and immediately after that, he took an 80-year-old lady's wallet and bag.

News and events can be also distinguished with reference to the content type.

Side events → In some cases, due to journalistic style, news report information also about other events similar to the main one described in the text (this happens often at the beginning or at the end of the article). Such events are identified as **side events**. If the description of side events is short and quick, such elements are not considered and the main event can be annotated. If this is not the case, we consider a long description of side events as a topic switch and do not annotate the news.

Theft attempt → When a news reports an attempt of theft, it is necessary annotating such information at a document-level using Doccano (see Appendix A - Doccano).

Theft related information → Some news report theft related information, such as finding and receiving stolen goods. In such cases, we annotate just the main event of theft described in the news.

Do not annotate:

- News reporting other types of crimes, other than theft;
- Other types of news, such as statistical reports on crimes;
- Multiple events.

When a document cannot be annotated, a comment at the document level is required, as it follows:

- Other type, e.g. assaults (specify)
- Multiple events
- Other

3.2. Span

The term "span" refers to a portion of text that can include one or more words. The annotated span should only refer to the entity and should not include definite/indefinite articles or prepositions, nor any spaces or punctuation marks at the beginning or end.

Examples (underline text indicates the span to be annotated):

- In Modena (...) → In Modena (...)
- The theft was committed in San Cesario sul Panaro. → The theft was committed in San Cesario sul Panaro.
- A person from Modena of Tunisian origins, aged forty (...) → A person from Modena of Tunisian origins, aged forty (...)

Some entities can be formed by more than one word. For instance,

- Ferrari and Giovanardi's auto body shop
- "Re Pipin" pizzeria
- Borgogioioso shopping center
- Playstation 4
- Apple computer
- medicines for animals
- row house

3.3. Synonyms and specific terms

News articles are characterized by the use of a high number of synonyms and terms as well. In case of synonyms, it is advisable annotating the first occurrence, as in the following example:

- The house was ransacked [...]. Rossella reported the theft in her home.

In case there are terms where one is more specific than the other, always annotate the more specific term. For instance,

- The thieves broke into the house [...]. The owners are dismayed because it is already the second theft in the row house.

General term annotation should be avoided, even though these are the only information presented in the text. This is the case of term such as:

- loot, stolen goods, ill-gotten gains, items
- thief, wrongdoer, criminal, assailant
- victim, unfortunate, owner, proprietor, lady/ladies

Some annotation examples are provided:

- "The thief was recognized to be a 50-year-old man."
- "The thief was recognized to be 50 years old."
- "The loot was stolen by two thieves."
- "The loot was stolen by two female thieves."
- "There are 3 victims, who are ladies."
- "There are 3 victims, who are elderly ladies."
- "There are 3 victims, they are elderly ladies."

4. Multiple annotations

Table 2: Crime Entities. * marks the theft-specific entity. Legend: R = Relation; ME = Multi-entity; MS = Multi-span; ML = Multi-label.

W	Entity	Definition	R	ME	MS	ML
Who	AUT	The causer of a crime	y	y	y	n
Who	AUTG	A group of causers	n	n	y	n
Who	VIC	The experiencer of a crime	y	y	y	n
Who	VICG	A group of experiencers	n	n	y	n
Who	PAR	An injured party in a crime	n	n	n	y
Where	LOC	The location of a crime	n	n	y	y
What	OBJ*	Robbed object(s)	y	y	y	n

4.1 Same label/different spans

Different spans may present the same label. This is the case when (i) more than one entity represents the answer to one of the competency questions (e.g., what has been stolen? a wallet and a purse); (ii) different spans refer to the same entity describing its characteristics (e.g., who is the author? a person living in Modena, 50 years-old).

Multiple annotations have different meanings according to the entity they refer to:

- **Stolen object:** multiple labels can refer to different stolen objects, e.g., a wallet and a purse, or to the quantity of stolen objects, e.g., seven laptops (see Section 6);
- **Theft location:** multiple annotations can occur when there are several spans referring to the same location and presenting different levels of informativeness (e.g., drugstore, Grandemilia drugstore. See Section 9));
- **Author/victim:** multiple annotations may refer to single authors/victims, e.g., a person from Modena and a person from Milan, and/or when additional information referring to one entity are identified, e.g., a person from Milan, who is a female and about 50 years old (see Section 7.1).

Some of the multiple annotations can be linked by specifying a relation between entities.

4.2 Same span/different labels

A span may present more than one label. This is the case where the same entity represents more than one concepts, e.g. *Marechiaro pizzeria* can be LOCATION and INJURED PART.

5. Relations

Relations are used to connect entities when there are additional information which specify their characteristics (see Appendix A - Doccano).

Not all the entities may be connected using relations. Entities that can be linked by a relation are:

- **author/victim:** annotations which refer to the same entity due to the fact that they specify further information about authors/victims are linked by a relation.
- **stolen object:** annotations referring to the amount of stolen objects can be linked by a relation.

Warning: relations are directed, this means that there is an entity from which the relation starts. It is important to carefully consider the starting entity for each of the categories.

6. Stolen object (OBJ)

	Allowed	Use
Multiple annotations	yes	they refer to different stolen objects
Relations	yes	between stolen object and their amount

The stolen object is the entity which has been stolen during the theft event. In case, there are more terms referring to the same entity, it is advisable annotating the most specific term (e.g., *diamonds* should be preferred to *jewellery*). If the most specific term cannot be identified, as they have the same level of descriptive granularity, e.g., Lagotto and truffle dog, the first term occurring within the text should be preferred.

Object spans do not include object characteristics, as the value, the material, the colour (for instance, in *golden rings*, we annotate just *rings*). Further specializations of the stolen object, as type or brand, are considered specification of the more general categories, so that they are annotated together with the general term, e.g. *Apple laptop*, *pet meds*, when they occur next to the other, otherwise only the most specific term has to be annotated.

For instance,

- Yesterday, a *vehicle* was stolen and then found in the ditch. It was a *Volvo VX60*.

Multiple annotations are allowed to annotate:

- 1) Different stolen objects.

For instance,

- A *bag* containing a *wallet* and a *cellphone* was stolen.

- 2) Numbers and precise amounts referring to the stolen objects. In such cases, **relations** have to be used to link multiple annotations

For instance,

- *Five Apple computers*
- Some *Apple computers* (*some* is not annotated because it represents a generic amount)

Warning: relations are directed, this means that there is an entity from which the relation starts. It is important to carefully consider the starting entity for each of the categories.

7. Author and group of authors

The entity committing the theft can be described as one or more living beings and/or a group. The annotation reflects this distinction by identifying:

- **AUT** (living beings): e.g. *woman, 38 years old, person from Modena*, “first name/family name” *etc.*,
- **AUTG** (a group of living beings): e.g. *Three men from Modena*.

7.1. Author (AUT)

	Allowed	Use
Multiple annotations	yes	They refer to different authors or to authors' characteristics
Relations	yes	Between the author and their characteristics

AUT represent the entity perpetuating the theft. A theft event can be perpetuating by one or more authors.

Single spans are annotated with reference to different authors.

For instance,

- A person from Milan and a person from Naples have been reported to the police for aggravated theft

Within the news article, **further information about authors'** may be retrieved. In such cases, where it is required the annotation of multiple spans referring to the same author, the annotation should be performed by a three-step procedure:

1) First step:

For each author mentioned within the article, annotators have to identify the main information, such as first/family name, the initials of the first name and last name, and socio-demographic reference information, such as age, race, ethnicity, residence, inhabitant/native, gender, occupation, legal status (e.g., clean record, convicted, with previous convictions). No other characteristics or conditions or roles should be annotated (e.g., perpetrator, criminal, blonde, young, elderly, under arrest, on probation, escaped, husband, wife...).

Examples:

- A 45-year-old person from Milan and a 45-year-old person from Naples have been reported to the police for theft. The Milanese individual, E. E. (...)

Altri esempi (a destra della freccia sono sottolineati gli span da annotare):

- A young person from Carpi, who is a minor → A young person from Carpi, who is a minor (Carpi can indicate that the minor either lives in Carpi or is native to Carpi, so it is annotated)
- A minor residing in Carpi → A minor residing in Carpi
- A minor living in Carpi → A minor living in Carpi
- A minor originally from Naples, living in Carpi → A minor originally from Naples, living in Carpi (annotations such as "works in Carpi" or "studies in Carpi" should not be included)

Exception:

- A young person from Carpi → A young person from Carpi (in this case, "young person" is annotated because it is the only noun that identifies the subject and allows the connection to the socio-demographic information "Carpi").

Warning: Uncertain information should not be annotated. (e.g., "they seemed Maghrebi" → do not annotate; "Eastern European citizens, probably Romanian" → do not annotate because "Eastern European citizens" is not socio-demographic information, and "Romanian" is not certain information).

2) Second step:

Among the annotated spans referring to the same author, it is necessary to choose one unique span called **identifier** of an author within the article.

The identifier should be the author's name, if present, or the first occurrence denoting the author.

If a span has been selected to represent an author it cannot be used as identifier for other authors.

Example:

The theft was committed by a 45-year-old man, a Neapolitan resident of Carpi, who acted together with an accomplice, another man from Milan resident of Modena.

"Il furto è stato commesso da un uomo, napoletano 45enne residente a Carpi che ha agito assieme a un complice, un altro uomo milanese residente a Modena"

In the previous example there are two authors.

The **first author** is described by four spans, that are: *man, Neapolitan, 45-year-old, resident of Carpi*.

According to the aforementioned rule, the first occurrence referring to the first author is *man*, so we can select this as an identifier.

The **second author** is described by three spans that are *man, from Milan, resident in Modena*. In this case, the identifier is *from Milan*.

3) Third step:

The identifier should be linked to the other spans which describe an author.

Warning: relations are directed, this means that the relations should be created from the identifier to the other spans.

If there is information regarding unidentifiable authors for whom neither names nor socio-demographic information are provided, such authors should not be annotated, as *man* in the following example:

- The Carabinieri have not yet identified the third man involved in the robbery.

7.2. Group of Authors (AUTG)

	Allowed	Use
Multiple annotations	yes	They refer to the group of authors
Relations	no	--

The group of authors label is used to annotate multiple authors when they are described as collective entity, e.g., A gang of 5 Moroccan thieves (...).

Just like for individual authors, the spans describing the group should be annotated, which means the socio-demographic reference information such as age, race, ethnicity, residence, inhabitant/native, gender, occupation, legal status (e.g., clean record, convicted) that will be attributed to the group that committed the theft and not to individual authors or subgroups.

Warning: The use of relations for the AUTG class is not allowed.

Any descriptions of subgroups (e.g., "*two other accomplices* have not been identified") should not be labeled as AUTG.

General terms such as "criminals," "gang," "thieves" should not be annotated, but only the number of members and/or socio-demographic characteristics.

When a news article contains information about both the group of authors and the individual authors, all the information should be annotated using respectively AUTG and AUT labels.

8. Victim, group of victim and injured party

The entity that experiences the theft can be a living being or an organization. The annotation reflects this difference by identifying:

- VICTIM (living being): e.g., 38 years old, from Modena, "name and surname", etc.
- INJURED PARTY (organization): e.g., I Gelsi cooperative, hospital, Pediatrics department, jewelry store, "business name", etc.

8.1. Victmin (VIC)

	Allowed	Use
Multiple annotations	yes	They refer to different victims
Relations	yes	Between victim span and their characteristics

Victims are annotated following the annotation rules for authors.

If the news indicates one or more victims of the theft, as already mentioned for the author, the information referring to the same subject is connected through relationships.

We follow the annotation process in 3 steps presented in Section 7.1.

If a theft has been committed against a store, it is not correct to annotate the owner as the victim. Instead, the store should be annotated as the injured party (see Section 8.3).

If there are multiple victims in a theft, it is necessary to determine whether it constitutes a single event or a multi-event. If the event occurs in a single location by the same author(s) at a single moment in time, it is a single theft event. However, if the location varies, or the authors of the theft vary, or the moments in time are subsequent, it is considered a multi-event (see examples in Section 3.1).

Warning: Relations between different labels are not allowed. If in a news article the stolen items are attributable to individual victims, victims and stolen items should be annotated separately, and no relations should be inserted between the stolen items and the victims.

8.2. Group of Victims (VICG)

	Allowed	Use
Multiple annotations	yes	They refer to the group of victims
Relations	no	--

If the theft is suffered by a group of victims, the text may contain references describing that group (e.g., 3 elderly ladies → 3 elderly ladies).

Just like for individual victims, for a group as well, the spans describing it should be annotated. This includes socio-demographic reference information such as age, race, ethnicity, residence, inhabitant/native, gender, occupation, legal status. These annotations will be attributed to the group that committed the theft and not to individual authors or subgroups.

Warning: Relations are not allowed.

Any descriptions of subgroups (e.g., the owner, Luisa Rossi, and two other tenants were robbed) should not be annotated with the VICG label.

General terms like "victims," "people," "unfortunate individuals," "owners" should not be annotated. Instead, the number of members and socio-demographic information should be annotated.

8.3. Injured party (PAR)

	Allowed	Use
Multiple annotations	no	--
Relations	no	--

The term *injured party* refers to organizations or companies that are victims of theft. The rules mentioned earlier for other subjects also apply to victims and the injured party.

If the news indicates the injured party of the theft, the victim should not be annotated.

If an event involves the injured party instead of the victim, it is likely that it coincides with the location where the theft occurs. In such cases, the corresponding span will have a dual annotation as both a location and an injured party.

If the specialized term is in proximity to the general term, they should be annotated together, for example, "Re Pipin pizzeria."

If the spans are discontinuous and there is no proximity, the more specific term should be annotated. For example, "A pizzeria was robbed. The owner of Re Pipin declares that..."

9. Location (LOC)

	Allowed	Use
Multiple annotations	yes	All the span refer to the same location where the theft happened
Relations	no	--

The location where a theft occurs is annotated with multiple spans denoting the generic place (e.g., apartment, residence, logistics company, supermarket, store name - e.g., "coop Grandemilia," "Borgogioioso shopping center") and/or the geographic area (street, city, etc.).

The generic place allows identifying the type of location where the theft took place, while the geographic area allows geolocation.

If the geographic area is indicated with different levels of specificity, such as "via della Chimica, 11," "historic center," "Carpi," choose the most specific information that allows geolocating the place of the theft, such as "via della Chimica, 11" and "Carpi."

If there is no specific place mentioned, but there are general pieces of information, annotate the general areas like "historic center," "artisanal zone," "Madonnina neighborhood."

Warning: In some cases, the news text begins with the indication of a geographic area/city name (that is the news dateline), usually in uppercase, which is the city of the correspondent. This location should not be confused with the place where the theft occurs, which is mentioned in the news text. The city of the correspondent should never be annotated.

Some locations consist of multiple-word spans.

If the specialized term is in proximity to the general term, they should be annotated together, for example, "Borgogioioso shopping center."

If the spans are discontinuous and there is no proximity, each span should be annotated separately.

Warning: be careful to identify the location where the theft happens and not the location where the stolen object(s) are retrieved. For instance, in the following sentence *Via Rainusso* refers to the location where the *Fiat Panda* has been retrieved, so we do not annotate it

- *Yesterday, a police patrol managed to track down a stolen Fiat Panda in Via Rainusso, Modena, which had been stolen about ten days ago (...).*

10. Time


Time references are not annotated.

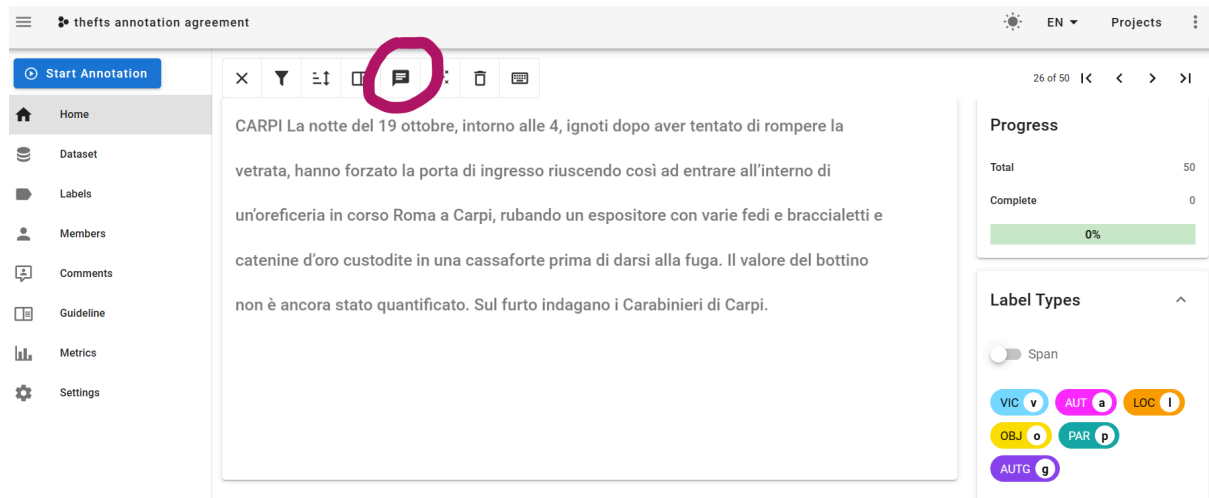
11. How

We do not annotate the information about how the theft happened.

Appendix A - Doccano

Comments

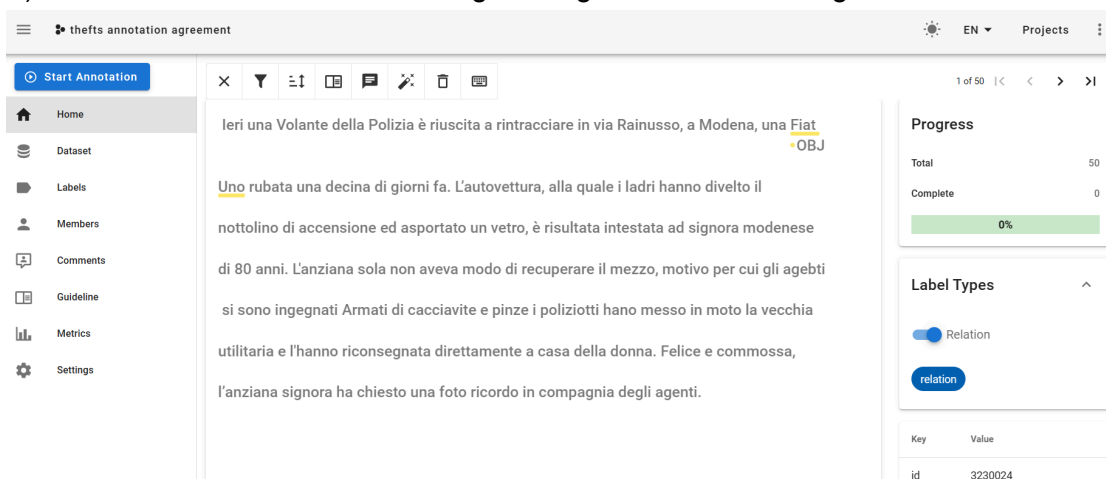
Doccano allows you to insert comments at the document level. Click on the icon  , type the comment and then click “Send” and “Close”.



The screenshot shows the Doccano interface for a document titled "thefts annotation agreement". The left sidebar contains navigation links: Home, Dataset, Labels, Members, Comments, Guideline, Metrics, and Settings. The main text area displays a paragraph of Italian text. A comment icon (speech bubble) in the top toolbar is circled in red. The right sidebar shows a "Progress" section with a 0% completion bar and a "Label Types" section with various labels like VIC, AUT, LOC, OBJ, PAR, and AUTG.

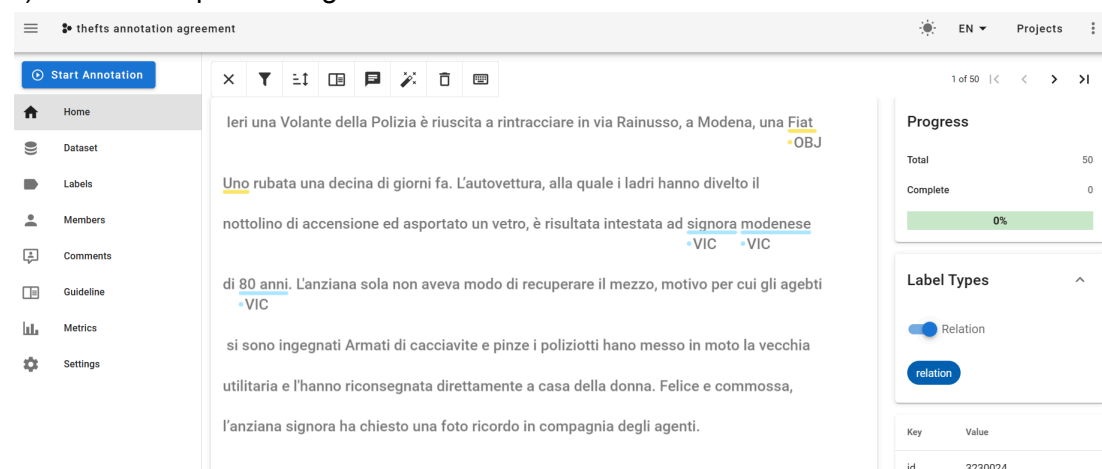
Relations

1) Relations can be added selecting the flag “Relation” in the right menu



The screenshot shows the Doccano interface with a document titled "thefts annotation agreement". The left sidebar is the same as the previous screenshot. The main text area shows a paragraph of Italian text. The right sidebar shows the "Progress" section and the "Label Types" section. The "Relation" flag is selected in the "Label Types" section, and a "relation" button is visible below it. The "Key" and "Value" table shows "id" with the value "3230024".

2) Annotate spans using the available labels



The screenshot shows the Doccano interface with a document titled "thefts annotation agreement". The left sidebar is the same as the previous screenshot. The main text area shows a paragraph of Italian text with several spans annotated using labels: "Uno" (OBJ), "rubata" (VIC), "signora modenese" (VIC), "di 80 anni" (VIC), and "si sono ingegnati" (VIC). The right sidebar shows the "Progress" section and the "Label Types" section. The "Relation" flag is selected in the "Label Types" section, and a "relation" button is visible below it. The "Key" and "Value" table shows "id" with the value "3230024".

- 3) Click on the label of the first entity to be connected and then click on the label of the second entity. A dropdown menu shows the available relations, select the desired one.

The screenshot displays the 'thefts annotation agreement' web application. The interface includes a left sidebar with navigation options: Home, Dataset, Labels, Members, Comments, Guideline, Metrics, and Settings. The main workspace shows a text document with several sentences. Entities are highlighted with colored labels: 'Fiat' is labeled 'OBJ' (Object), 'signora modenese' is labeled 'VIC' (Victim), and '80 anni' is labeled 'VIC'. A 'relation' arrow connects the 'VIC' label for 'signora modenese' to the 'VIC' label for '80 anni'. The right sidebar contains a 'Progress' section showing 'Total' as 50 and 'Complete' as 0, with a 0% progress bar. Below this is the 'Label Types' section, which has a toggle for 'Relation' and a 'relation' button. At the bottom right, a table shows the 'id' as '3230024'.

thefts annotation agreement

Start Annotation

Home

Dataset

Labels

Members

Comments

Guideline

Metrics

Settings

1 of 50

leri una Volante della Polizia è riuscita a rintracciare in via Rainusso, a Modena, una Fiat
•OBJ

Uno rubata una decina di giorni fa. L'autovettura, alla quale i ladri hanno divelto il

nottolino di accensione ed asportato un vetro, è risultata intestata ad signora modenese
•VIC •VIC

di 80 anni. L'anziana sola non aveva modo di recuperare il mezzo, motivo per cui gli agebti
•VIC

si sono ingegnati Armati di cacciavite e pinze i poliziotti hano messo in moto la vecchia

utilitaria e l'hanno riconsegnata direttamente a casa della donna. Felice e commossa,

l'anziana signora ha chiesto una foto ricordo in compagnia degli agenti.

relation

Progress

Total 50

Complete 0

0%

Label Types

Relation

relation

Key	Value
id	3230024