



TECHNOLOGY, TRAINING AND KNOWLEDGE FOR EARLY-WARNING / EARLY-ACTION LED POLICING IN FIGHTING ORGANISED CRIME AND TERRORISM

REPORT ON D4.4 ANNOTATION TOOL FOR THE LAW ENFORCEMENT DOMAIN

Grant Agreement: 786687

Project Acronym: COPKIT

Project Title: Technology, training and knowledge for Early-Warning / Early-Action led policing in fighting Organised Crime and Terrorism

Call (part) identifier: H2020-SEC-2016-2017-2

Document ID: CPK-2005-WP04-004-V1-DV-PU

Revision: V1.0

Date: 31/05/2020

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Dissemination Level		
PU	Public	<input checked="" type="checkbox"/>
CO	Confidential, only for members of the consortium (including the Commission Services)	<input type="checkbox"/>
EU-RES	Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)	<input type="checkbox"/>
EU-CON	Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)	<input type="checkbox"/>
EU-SEC	Classified Information: SECRET UE (Commission Decision 2005/444/EC)	<input type="checkbox"/>

Revision history

Revision	Edition date	Author	Modified Sections / Pages	Comments
0.1	05.05.2020	Sven Schlarb	All	Document draft
0.2	25.05.2020	LIF Team	All	Annex A. Ethical review results and response
0.3	25.05.2020	IGPR	All, extension of section 5	Addressed comments from IGPR document review.
0.4	07.05.2020	UGR	All	Addressed comments from UGR document review.
0.5	29.05.2020	Sven Schlarb	All	Finalizing document
1.0	29.05.2020	Raquel Pastor		Some corrections and formatting for delivery (code, headers, footers...).

Table of Contents

1. Introduction	6
1.1. Background.....	6
1.2. Purpose, Scope and Structure	6
1.3. Document Structure	7
1.4. Applicable and Reference Documents	7
1.5. List of acronyms and abbreviations	7
2. Natural language annotation for machine learning.....	8
2.1. Definition and role in the domain of fight against crime and terrorism.....	10
2.2. Overview of challenges and available technologies.....	10
2.3. Overview of components included in D4.4	11
3. Concepts and architecture overview	12
4. Presentation of the Tool “Recogito”	17
4.1. Technical description	18
4.2. Demonstration of the component	19
4.3. Ethical, Legal and privacy challenges	26
5. Presentation of the Tool “Data Set Repository”	27
5.1. Demonstration of the component	27
5.2. Technical description	32
5.3. Ethical, Legal and privacy challenges	35

6. Conclusion and next steps.....	36
Annex A. Ethical review results and response.....	37
Annex I. Example of web annotations.....	38
Annex II. Example of data set metadata.....	39
Annex III. Example of model evaluation metadata	40
Annex IV. Batch Data Set Submission using the Data Set Repository API	41
Annex V. T4.5 - Expert annotation and integration of extracted information -Virtual workshop – Usability test	44
Bibliography	45

List of Figures

Figure 1 (Pustejovsky and Stubbs 2012 p. 106), Natural Language Annotation for Machine Learning	9
Figure 2 Darknet Advertisement	11
Figure 3 Annotation and model creation steps in the context of the information extraction workflow.....	12
Figure 4 Tools used at the different stages of the data processing pipeline	13
Figure 5 Data set components for working copy and stored entity of the data set.....	16
Figure 6 <i>Recogito</i> annotation user interface	17
Figure 7 Diagram of the annotation example	18
Figure 8 COPKIT Data Set Repository (DSR).....	28
Figure 9 Displaying and changing annotations of weapon named entities	29
Figure 10 Starting the model creation for named entity recognition	30
Figure 11 Creating the named entity recognition model	31
Figure 12 Inspect processing log when the model creation process is finished.....	32

List of Tables

Table 1: List of acronyms and abbreviations.....	8
Table 2 Overview about datasets used in WP4	14
Table 3 Dataset metadata	34
Table 4 Representation metadata.....	34

1. Introduction

1.1. Background

The COPKIT project focuses on the problem of analysing, investigating, mitigating and preventing the use of new information and communication technologies by organised crime and terrorist (OCT) groups. For this purpose, COPKIT proposes an intelligence-led Early Warning (EW) / Early Action (EA) system for both strategic and operational levels. The project duration is 36 months (from 2018 to 2021) and the works are structured in nine work packages (WPs).

Work package WP4 (“Knowledge Extraction Components”) is about collecting data from a variety of data sources and gaining knowledge from unstructured text content. The work package includes five tasks T4.1, T4.2, T4.3, T4.4, and T4.5. The goal of task T4.1 “Collecting and pre-processing of source material” is to make data available in order to support the application of analysis in the other tasks of the work package. Data are harvested from open web and darknet sources. The results of this task have been reported as part of deliverable D4.1 “Compilation of Data Sources”.

Task T4.2 “Event extraction from location-based social media” focuses on the discovery of events based on mining text data. There are existing approaches for discovering events regarding conventional sources, such as news stories, for example. However, the data sources used in WP4, such as darknet forum data and HTML websites from online markets pose specific challenges due to the brevity of the texts and the informal style in which it is written (including orthographic and grammatical errors and use of slang). Furthermore, this task helps enriching the spatiotemporal analytics from Task 6.3 by providing the context of the anomalies.

Task 4.3 “Entity extraction component” is about extracting named entities from unstructured text sources. Many available open-source Named Entity Recognition (NER) tools, such as the Stanford Named Entity Recognizer (NER)¹ or SpaCy’s Named Entity Recognition² are optimized for detecting generic named entity types, such as names of persons, locations, organizations, and products. This task will develop NER tools which also allow extracting other types of entities, such as names of weapons, drugs, or digital fraud items which are relevant regarding the use cases and type of data in COPKIT.

Task 4.4 “Relationship extraction component” takes the results from task T4.4 as a basis to identify relationships between named entities. On the one side, these are relationships between the entity and its properties, such as features of a weapon, for example. On the other side it is the relations between conventional entity types, such as usernames and locations, to the use case specific entity types (e.g. items offered in a web shop).

The report about the prototype results (first release) of tasks T4.2, T4.3, and T4.4 are included in this deliverable. Each of these tasks creates a demonstrator which was part of the M18 demonstration to LEAs (November 2019 in Athens at KEMEA).

Task T4.5 “Expert annotation and integration of extracted information” is about a web-based user interface for visualizing named entities and relationships as a result of the components developed in tasks T4.3 and T4.4. The objective of this task is to implement tools for data validation, annotation, and exploration and support LEA partners in applying them. This deliverable describes the technical component which is used in Task T4.5 for the annotation of data by the experts.

1.2. Purpose, Scope and Structure

This document (“Annotation Tool for the Law Enforcement Domain”) is a report about a software deliverable that is prepared as part of task T4.5 “Expert annotation and integration of extracted

¹ <https://nlp.stanford.edu/software/CRF-NER.html>

² <https://spacy.io/usage/linguistic-features#named-entities>

information". The purpose of this deliverable is to provide an environment where LEAs can evaluate an environment for creating annotations which are required for the supervised learning methods used in Tasks T4.3, and T4.4.

In order to allow using the annotation component, it is embedded in a data management environment which is required to maintain a ground truth data corpus for model training. The software deliverable therefore includes also the environment of the annotation component.

1.3. Document Structure

This document is structured in the following way:

- Section 2 gives background information about natural language annotation for machine learning and the role in the domain of fight against crime and terrorism. Furthermore, this section describes the challenges regarding content harvested from darknet markets and forums. Note that parts of section 2 are repeated from deliverable D4.2 for the reader's convenience to have background information available in this deliverable and for those readers who do not have access to D4.2 which is confidential.
- Section 3 gives an overview about the general concepts and architecture. It is explained how the core software deliverable "Recogito" is integrated with the data set repository component which is also delivered as part of D4.4.
- Section 4 describes the component Recogito which is the web user interface for annotation to be used by LEAs for labelling ground truth data for supervised machine learning algorithms.
- Section 5 describes the data set repository which is used to manage datasets used in the COPKIT project.
- Section 6 describes the conclusions and next steps.

1.4. Applicable and Reference Documents

- Grant Agreement number 786687 - COPKIT - H2020-SEC-2016-2017/H2020-SEC-2016-2017-2.
- D2.1 – Use cases: understanding technological factors driving OCT organisations, COPKIT, CPK-1901-WP02-001-V1.0-DV-EURES
- D2.3 – Technical requirements based on LEAs' needs, COPKIT, CPK-1901-WP02-003-V1.0-DV-EURES
- D2.4 – Compilation of data sources, CPK-1901-WP04-001-V1.0-DV-CO

1.5. List of acronyms and abbreviations

Acronym / abbreviation	Definition
BAYHFOD	Hochschule fur den Offentlichen Dienst in Bayern
COPWIK	Knowledge base of the COPKIT eco-system
EA	Early Action
ELP (Team)	Ethical, Legal and Privacy (Team), the team of partners specialized in Ethical, Legal and Privacy aspects.
EU	European Union
EUROPOL	European Police Office
EW	Early Warning
IE	Information Extraction
LEA	Law Enforcement Agency

Acronym / abbreviation	Definition
NER	Named Entity Recognition
NLP	Natural Language Processing
OCT	Organised Crime and Terrorist
PGP	Pretty Good Privacy
PoS	Part of Speech
VOIP	Voice over IP
WP	Work Package

Table 1: List of acronyms and abbreviations

2. Natural language annotation for machine learning

Natural Language Processing (NLP) covers a wide field of methods and techniques to extract information from text documents. **Information extraction (IE)** is the task of extracting structured information from unstructured or semi-structured text documents. On the one hand, there are **unsupervised algorithms** which do not require the creation of labelled ground-truth data. The language models need to be trained for the type of text given a specific language. Furthermore, they also depend on how the text is presented in the corresponding context (websites, forums, social media, market places, etc.). On the other hand, there are **supervised algorithms** which rely on labelled ground-truth data, i.e. words of interest are manually marked as being of a specific category (drugs, weapons, etc.). The latter is the focus of this deliverable.

This information can then be used to identify these categories in unknown text parts found in dark-net marketplaces, for example. The information gained from unsupervised methods can also be used to support the ground-truth creation of the supervised approaches by suggesting labels which then need to be confirmed or rejected by a human annotator.

Named entity extraction (NER) is a text mining approach which tries to identify specific information entities that are relevant in a given application domain. In the COPKIT project, these entities can be names of objects that are relevant in criminal investigations, such as weapons, drugs, etc., or entities which are related to users or shipment details in the context of an online market, for example. An important means to analyze the text is the so-called **Part-of-Speech (PoS) Tagging** which allows extracting noun phrases (NP) and building grammatical trees to find relationships between the different parts of sentences used in a text.

New NLP techniques trying to go beyond state-of-the-art make use of **Deep Learning** architectures which are also relevant for the classification of named entities. In a supervised approach, a labeled training set is used to create a model for automatically extracting entities. The use of so called “word embeddings” on large text corpora permits gaining knowledge about the general structure of the language by providing comparable vectors for words and expressions. The text corpus which is used as a basis for creating the models depends on the application domain. For this reason, task T4.1 of WP4 consists of collecting data from a variety of web data sources which are specific to the law enforcement domain.

Annotation is a method employed in *corpus linguistics* which aims at providing meta information about language at different levels, i.e. it can be related to morphemes, words, sentences, or documents. It ranges from simply assigning class labels to documents to capturing the grammatical structure of sentences.

Apart from the corpus linguistics domain, the method is used in many other scientific areas where different terms are being used. In computer science and especially in the area of machine learning the assignments are often called *classes*, *targets*, *labels* or *categories*. In social sciences, for example, this kind of categorization is called “*coding*” (Saldaña 2009).

In corpus linguistics, different types of corpora are used with a specific purpose. The following examples are important corpora for natural language processing tasks:

- **Penn Treebank** (Marcus et al. 1993) – The Penn Treebank provided about 7 million words of American English annotated for part-of-speech (POS) information between 1989 and 1996.
- **OntoNotes** (Weischedel et al. 2011) created a corpus encoding information about syntax, predicate-argument structure, word sense, and coreference.
- **FrameNet** (Baker et al. 1998) – The FrameNet corpus provides a collection of annotated examples of how words are used in actual texts, where a frame defines the elements which are required to understand a concept in a specific context.
- **MPQA** (Deng and Wiebe 2015) – The MPQA corpus contains news articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).
- **SQuAD** (Rajpurkar et al. 2016) – The Stanford Question Answering Dataset (SQuAD) is a consists of more than 100000 questions related to specific parts of Wikipedia articles.

Figure 1 shows the typical flow of the annotation activity for corpus building (Pustejovsky and Stubbs 2012 p. 106). According to this workflow, an annotation project provides a specification and guidelines about how the annotation should be done by the annotators. If there is a low inter-annotator agreement, a revision of the guidelines is done. Generally, the higher the inter-annotator agreement, the better is the basis for machine learning tasks later because, basically, one cannot expect a machine learning algorithm to distinguish classes where even humans do not agree on how to distinguish them. Based on a set of annotations created in several rounds, the adjudication step decides on which of the annotations should be used in the gold standard used for machine learning.

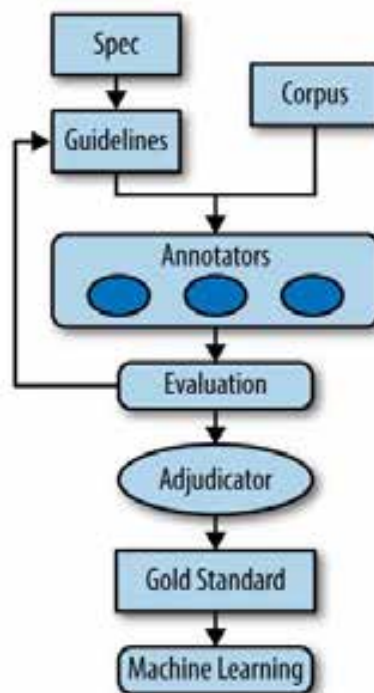


Figure 1 (Pustejovsky and Stubbs 2012 p. 106), Natural Language Annotation for Machine Learning

Generally, it depends on the type of task, how much variance in the inter-annotator agreement can occur. In COPKIT, the variance in annotating weapon entities in texts is probably much lower than the task to highlight dangerous events in texts. In this sense, the task of annotating weapon entities has a lower degree of subjectivity compared to the judgement of whether an aspect in a text should be considered dangerous or not.

For this reason, the redundancy required for the annotation of text documents needs to be decided specifically for a given use case and task.

In WP4, the annotation task T4.5 is focused on annotating named entities and relationships between these entities, which are tasks T4.3 and T4.4 respectively. The specification are the guidelines which LEA partners receive with respect to a specific use case. For D4.4, this use case is the annotation of weapon advertisements in darknet markets. The goal of the annotation activity in T4.5 is not only to create a gold standard required for machine learning, but also to learn about what kind of entities and relationships LEAs are interested in.

2.1. Definition and role in the domain of fight against crime and terrorism

During the last decade, the darknet has provided unprecedented opportunities for trafficking illegal goods, such as weapons, drugs, or instructions and tools for digital fraud. Along with the possibilities of concealing financial transactions with the help of crypto currencies, the darknet offers sellers the possibility to operate in covert. Often, the illegal marketplaces are uncovered and closed by the police, as was the case with some of the more well-known marketplaces, such as *Silk Road*, *AlphaBay*, *Dream Market* and most recently the *Wall Street Market*. One of the biggest challenges is that data needs to be acquired and analysed quickly because operators of illegal data markets might close them on the slightest suspicion of police surveillance. In addition, the amount of illegal trading objects is unmanageably large, so that investigators need automated tools that help getting insights into the data.

Many information sources that WP4 of the COPKIT project is dealing with is web data retrieved from open or darknet forums and marketplaces. The information available on shop websites is usually available in semi-structured form, i.e. the HTML format of the web pages allows deriving some information entities, such as usernames, price or shipment information, from tabular or layer definitions of the websites. After transforming the collected web data into a suitable format for data analysis, methods for archiving, pre-processing and data transformation are required so that statistical approaches and the use of machine learning can be used to identify patterns in the data and thus gain information faster and more efficiently.

The general approach of WP4 is therefore to provide components which allows gaining knowledge in form of use case specific and meaningful information entities and the relations between them from unstructured web data as it is available in online shops of darknet markets and forums.

2.2. Overview of challenges and available technologies

Annotation is used in WP4 to provide a human validated basis of information to adapt generic language models for the law enforcement domain for the tasks of Named Entity Recognition (NER) and Relationship Detection.

Regarding Named Entity Recognition (NER) the goal is to identify selected information elements, so called Named Entities (NE), a term which was originally coined at the 6th Message Understanding Conference (MUC) to denote names for people, organizations, localizations, and numerical expressions.³ In the Automatic Content Extraction (ACE) Program lead by the US National Institute of Standards and Technology (NIST)⁴ additional entity types, such as organization, geo-political, facility, vehicle, weapon, were introduced.

Regarding the classical NER element types, the task usually achieves high success ratios over 95% in terms of precision and recall on task specific evaluation data sets.⁵ While the task is very successful on

³ R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in 16th Conference on Computational linguistics, 1996, pp. 466-471.

⁴ <https://www.nist.gov/>

⁵ Marrero, M.; Urbano, J.; Sánchez-Cuadrado, S.; Morato, J. & Berbís, J. M. G. Named Entity Recognition: Fallacies, challenges and opportunities Computer Standards & Interfaces (CSI), 2013, 482-489, p. 1.

typical entity types, it remains challenging to optimize NER classifiers to perform well on new types of domain-specific content.

One of the main challenges in the COPKIT project is therefore to optimize NER to extract the entity types of interest, such as the weapon, drugs, or digital fraud, as well as common entity types, such as locations and organizations, for example.

Another challenge arises because of the specific characteristics of the texts published in online markets. The most outstanding difference compared to standard corpora is that the data is not always structured in sentences and paragraphs. Figure 2 Darknet Advertisement shows an example of a drug advertisement with special characters, emoji and ASCII art which are used to structure the text and to make the advertising more appealing.



Figure 2 Darknet Advertisement

Furthermore, grammar and syntax are not always strictly followed. In many cases the offers consist of bullet lists and enumerations rather than full sentences. The style of the remaining text can be compared to the type of language used in typical advertisements. Additionally, offers occur in multiple languages and the dataset contains PGP⁶ keys and lists of keywords for search engine optimization that need to be filtered. Furthermore, the true nature of products is often obfuscated by the use of codewords or vague language.

2.3. Overview of components included in D4.4

Deliverable D4.4 includes two main components:

1. **Data Set Repository (DSR):** creation and management of datasets.
2. **RecogitoJs:** UI for annotation which belongs to a family of open source software components for online document annotation.

The **Data Set Repository** is used for the creation and management of datasets. The component is used as a repository for storing data crawled from darknet markets and forums. It provides the means to transfer data together with metadata describing the dataset, and to manage the different types of data that are derived from harvested data during its lifecycle. The data set repository represents the basis of the annotation environment by providing the actual text data to be annotated as ground truth for creating models required by the named entity recognition and relationship extraction components.

⁶ Pretty Good Privacy, see https://en.wikipedia.org/wiki/Pretty_Good_Privacy

Recogito is a family of open source software components for online document annotation, developed under the leadership of AIT. In addition to providing a comprehensive, standalone platform for the collaborative annotation of texts and images,⁷ the *Recogito* project also develops a set of separate software libraries, which can be seamlessly embedded into other environments.⁸ For this deliverable, the *RecogitoJS* component was used, which provides embeddable user interface elements for text highlighting and tagging.

3. Concepts and architecture overview

Figure 3 illustrates the model creation process which is separated into the *Harvesting*, *Scraping*, *Labelling*, *Model Creation*, and *Information Extraction* steps.

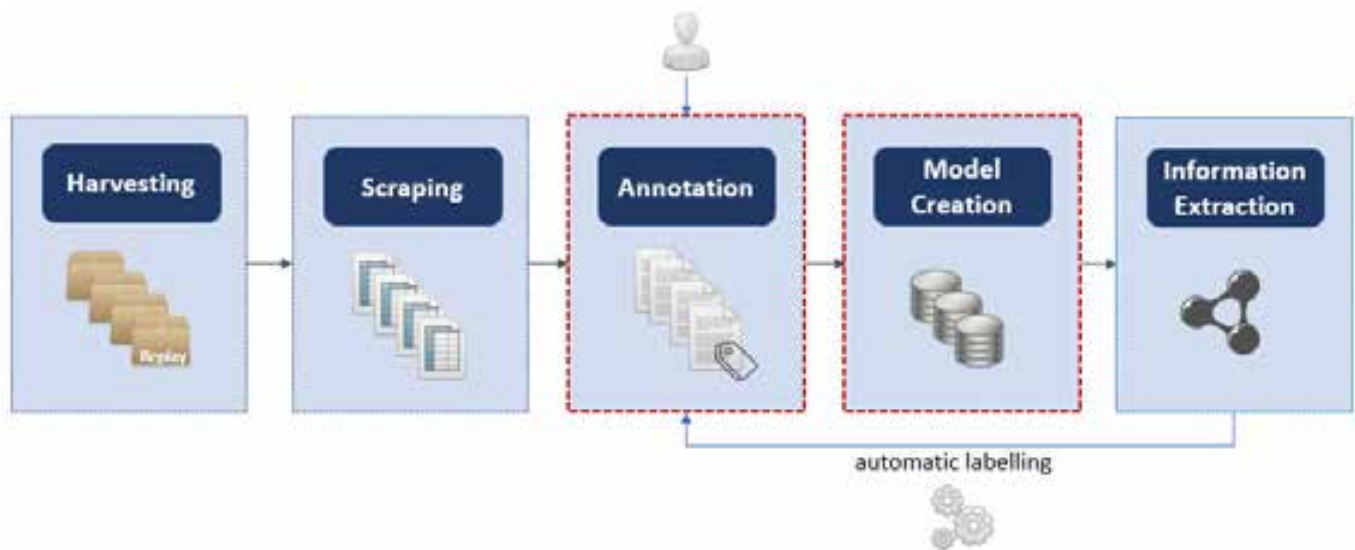


Figure 3 Annotation and model creation steps in the context of the information extraction workflow

The process starts with the **Harvesting** step by collecting web data from online or darknet markets or forums.⁹ This process preserves the evidence of collected web data in its original form. The result is packaged web data in form of the data produced by Mitmproxy¹⁰ or alternatively in the HAR format¹¹ which are currently the agreed upon web data packaging standards of the COPKIT project.

After the Harvesting step, the **Scraping** step performs the rough transformation of unstructured web data into structured information tables. These tables can already contain specific information entities. For example, in a typical online market, information entities, such as the vendor, price or shipment details appear on specific locations of the web pages, and it is therefore possible to directly extract them. Apart from these structured information entities, the scraping also extracts unstructured information in form of descriptive texts which are available for further analysis. In the context of this deliverable, the tabular data produced by the scraping step represent the basis for separating a data set into training and test subsets

⁷ <http://github.com/pelagios/recogito2>

⁸ <http://github.com/recogito>

⁹ The data used for building the initial model for named entity recognition does not rely on harvested data. The models are built based on a subset of the Darknet Market Archives called Grams, which contains crawls of thirteen darknet markets in the period from 09.06.2014 to 12.07.2015, containing approximately 10 million product offers. Filtering for unique product descriptions resulted in 226 661 datapoints. See Gwern Branwen et al. Dark Net Market archives, 2011-2015. Accessed: 2019-01-23. July 2015. url: <https://www.gwern.net/DNM-archives>.

¹⁰ <https://docs.mitmproxy.org>

¹¹ <https://w3c.github.io/web-performance/specs/HAR/Overview.html>

for creating machine learning models. The data resulting from the scraping step are pseudonymized before processing and using it in the subsequent steps.

The focus of this deliverable is on the **Annotation** and **Model Creation** steps highlighted in red Figure 3. The goal of these steps is to create human revised ground truth required to create models for the named entity recognition and relationship extraction tasks. Figure 3 shows that the Annotation step requires a human to correct annotations suggested by the system or create new annotations. The result of this step is used for the model creation which then allows the system to automatically suggest labels. The labelling and model creation steps are tightly coupled because the two steps are processed in an iterative manner.

Figure 4 gives an overview about the architecture of the D4.4 components. On the left side there are two groups of components, first the components for **Annotation & Model Creation**, and second the components for **Information Extraction**. The components are connected to the Data Set Repository which they use to load data or store processing results.

The right side of Figure 4 represents the **Data Set Repository** which provides a user interface and REST API to upload, access, and manage data sets used by the information extraction components. Apart from the use in WP4, this API is also used to integrate with WP4 components for enrichment and knowledge discovery.

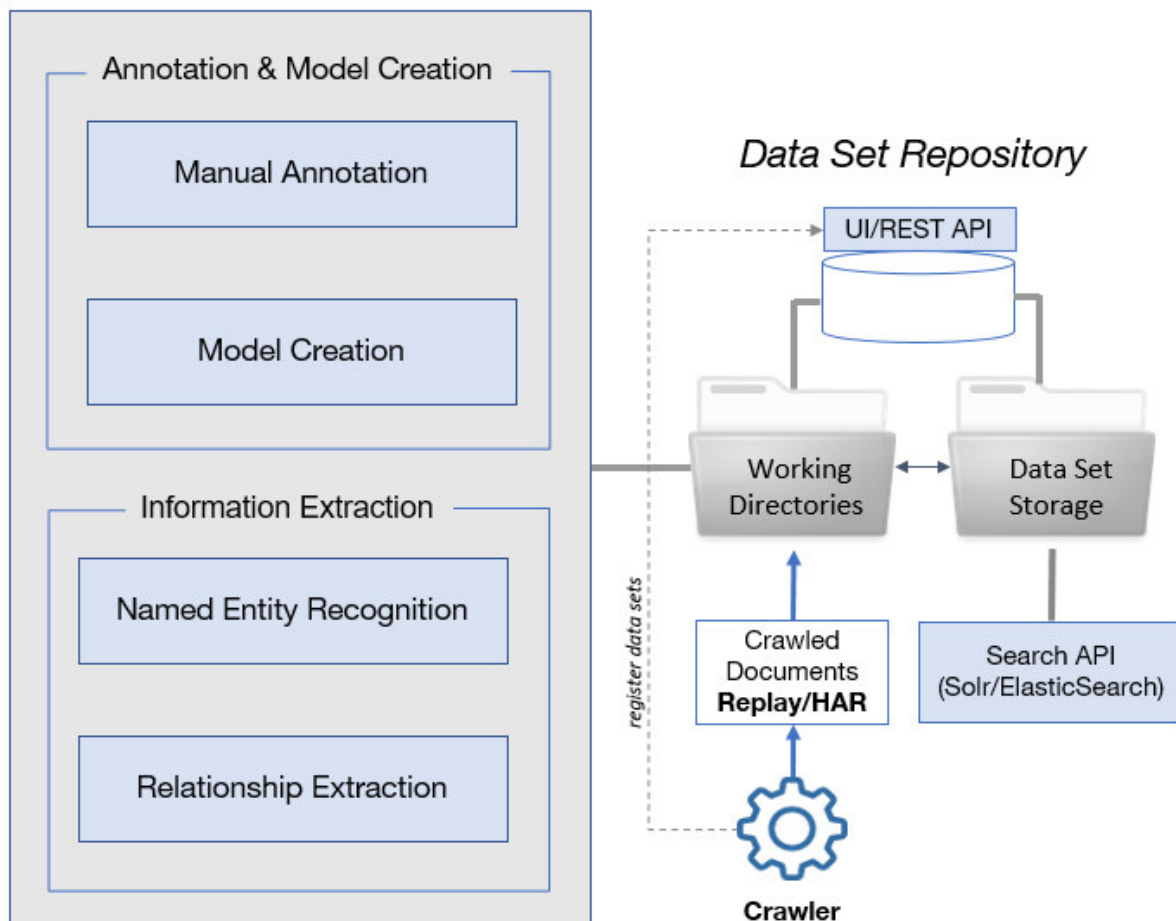


Figure 4 Tools used at the different stages of the data processing pipeline

The Data Set Repository allows creating and managing data sets with different types of **representations** of the data set where data set representation means either the data stored in some format or any transformed or enriched form of the original data.

Only pseudonymized data sets are deployed to the data set repository, restricted to those that are used to demonstrate functionality of the tools.

The demonstration scenarios make use of custom models and training/test sets for the detection of weapons. The model is trained on a subset of the Grams crawl in Gwern's archive (Gwern 2015) that contains strings from a list of weapon related terms.

Table 2 gives an overview about the datasets that were used during the implementation of the components included in this deliverable.

ID	Title	Format	Description	Date of Availability	Related tasks
1	ngram darknet ads	Table	Open data: Ngram dark net archive. Collection of advertisement from various darknet markets. Darknet advertisement texts, Structured metadata User alias	February 2019, see (Gwern 2015)	T4.3
1.1	classified ngrams ads	Table. Filtered Classified	Classified ngrams advertisements. A derivation of dataset 1, where firearms ads were semi automatically filtered and classified with respect to the type of weapon. Darknet advertisement texts, Structured metadata User	April 2019	T4.3
1.1.1	classified ngrams ads. NoUser	Table. Filtered Classified	classified ngrams ads. A derivation of dataset 1.1, Darknet advertisement texts, Structured metadata. Column with user login has been removed, but info may remain in ad text.	April 2019	T4.3
3	GN darknet ads 2018S2	Raw	A collection of darknet advertisements collected by GN before and at the beginning of the Project. Darknet advertisement texts, Structured metadata User alias Not filtered or classified, replay format	September 2018	T4.3, T4.4
4	GN darknet ads 2019Q3	Web crawl	A collection of darknet advertisements collected by GN in 2019Q3. Specifically focused on firearms and DaaS Not filtered or classified, replay and har format	September 2019	T4.3, T4.4
5	darknet forums	Web Crawl (HTML files)	A collection of darknet market forums archive. Dataset includes usernames which are Pseudonymized during extraction of text from HTML files	February 2019	T4.2

Table 2 Overview about datasets used in WP4

Following the steps shown in Figure 3, data is harvested by the crawler and registered in the repository as a new dataset. At this stage, the data set contains the data as it was collected from defined data sources.¹² In the case of data harvested from web or darknet data, this is the base representation of the data that is used to extract data from web documents using web scraping techniques. The result of this process is tabular data which serves as a basis for the Annotation & Model Creation components. The subsequent process of creating ground truth data requires human intervention in a supervised learning scenario.¹³ First, filtering and data cleaning is done to prepare data for the model creation.

Figure 5 shows a general concept of storing datasets in the data set repository. When creating a new dataset, a “Working Directory” (left side of Figure 5) allows collecting data, grouping it into sub-components and describing the data set with metadata, such as Title, Description, Category, Creator, etc. Once this process is finished, the actual Data Set is stored (right side of Figure 5).

The working copy and the stored data set have a set of components with the actual data. This can be harvested data (“crawl”), data extracted from the harvested data (“tabular”), pseudonymized data (“pseudonymized”). Distinguishing different types of data components allows defining a policy which regulates access to data sets on the level of these components, called data set “representations”. For example, for most development tasks, it will be enough to use “pseudonymized” data only, and it would not be required to use the original crawl which contains personal information. Or a policy can prescribe that specific dataset components have to be deleted after a certain time period.

At the same time, the identification of datasets and their derivatives allows keeping track of users that access the data or apply a processing pipeline, and also preserve the relationship of the machine learning models and the data sources they depend on.

¹² It is important to note that measures filtering of objectionable, inappropriate, or illegal content is being performed during harvesting. The data stored in the repository can therefore not be considered raw data.

¹³ At this step, the human intervention is required to annotate text data which has been pseudonymized for this purpose.

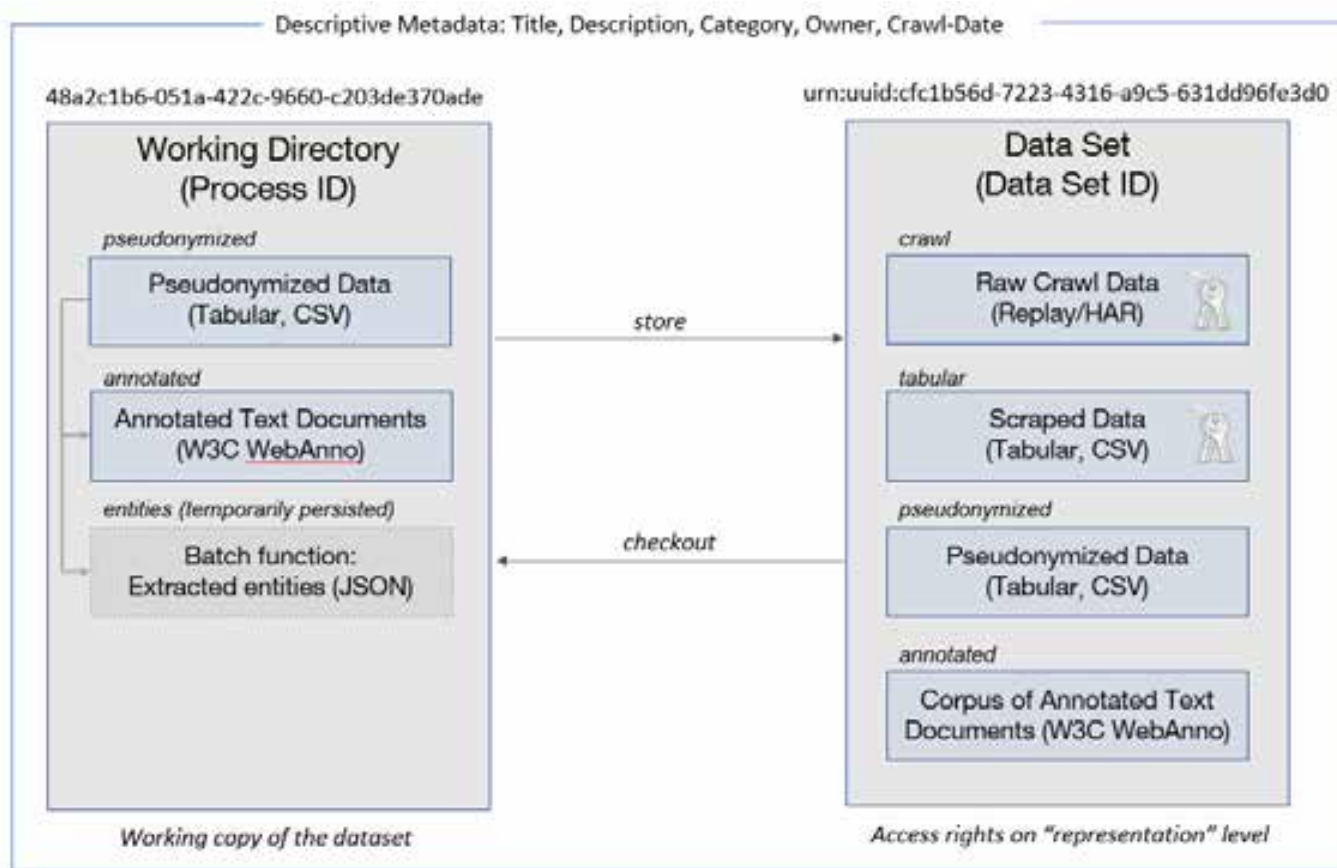


Figure 5 Data set components for working copy and stored entity of the data set

4. Presentation of the Tool “Recogito”

Recogito is a family of open source software components for online document annotation, developed under the leadership of AIT. In addition to providing a comprehensive, standalone platform for the collaborative annotation of texts and images,¹⁴ the *Recogito* project also develops a set of separate software libraries, which can be seamlessly embedded into other environments.¹⁵ All *Recogito* components are licensed under the permissive BSD license. For this deliverable, the *RecogitoJS* component was used, which provides embeddable user interface elements for text highlighting and tagging. As part of this deliverable, we integrated *RecogitoJS* into the Data set repository, and furthermore developed an extension that better serves the specific requirements of the COPKIT stakeholders. The extension replaces the standard *Recogito* annotation editor with an alternative user interface element that allows quicker browsing, validation, and correction of the tagging results produced by the machine learning process than the original *RecogitoJS* would.

Figure 6 shows a screenshot of *Recogito*'s user interface. In this example, two named entities have been annotated. One “weapon” entity and one “price” entity. A directed relation named “hasPrice” was defined to start from the “weapon” entity “Beretta 92G Elite II” to the “price” entity “\$750”.

The *RecogitoJS* component is designed to be easily integrated into other web user interface environments and provides JavaScript functions to retrieve the result of the annotation process for further processing.

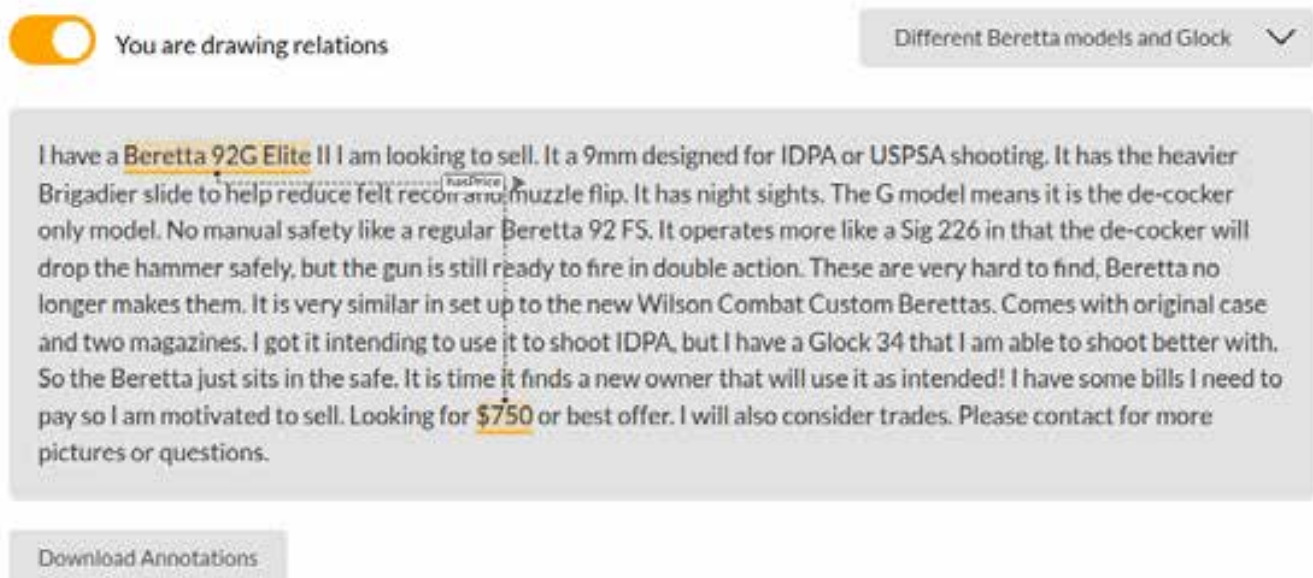


Figure 6 *Recogito* annotation user interface

¹⁴ <http://github.com/pelagios/recogito2>

¹⁵ <http://github.com/recogito>

4.1. Technical description

4.1.1. Technical framework

The user interface of *RecogitoJS* is a client-side implementation in JavaScript based on the React JavaScript library.¹⁶

4.1.2. Serializing annotations

Annotations are serialized in JSON-LD, a JSON-based format to serialize Linked Data, following the W3C recommendation Web Annotation Data Model.¹⁷

Figure 7 shows a diagram of the example represented in Figure 6. For the reader interested in the code representation of this diagram, the corresponding example in JSON-LD format is added as an appendix in Listing 1 on page 39.

In this example the string “Beretta 92G Elite I” is tagged as a “weapon” and the amount of money “\$750” is tagged as a “price” entity respectively. These entities are then related to each other via the relationship “hasPrice” using the IDs of the weapon and price entity.



Figure 7 Diagram of the annotation example

¹⁶ <https://reactjs.org>

¹⁷ <https://www.w3.org/TR/annotation-model>

4.2. Demonstration of the component

RecogitoJs is a client-side HTML/JavaScript application and can be opened in a Firefox or Chrome web browser. Note that *RecogitoJs* is the component which allows creating new or change existing annotations. This interface is integrated with the Data Set Repository to allow getting named entity suggestions displayed in the *RecogitoJs* user interface. After adapting the label suggestions, a new version of the named entity recognition model is created and evaluated against the test data set.

4.2.1. Component installation

No installation of the component is required because *RecogitoJS* is a client-side application.

Pre-requisite: Firefox or Chrome web browser (JavaScript must be enabled)

Installation:

- Obtain the distribution package file:

```
copkit_recogito.zip
```

Unpackage the ZIP file and open the HTML page `index.html` in Firefox or Chrome web browser.

4.2.2. Component usage

This section describes the usage of the *Recogito* annotation UI only, i.e. creating annotations for texts that do not contain any annotations. The usage of the integrated version of *RecogitoJs* is described in section 5.

The component is used by opening the HTML page `index.html` which is delivered as part of the component packaged as a ZIP file or via URL provided for the demonstration.

Generally, we recommend using the style guidelines for naming and labeling ontologies in the multilingual web (Montiel-Ponsoda et al. 2011).

For the demonstration scenario, the following recommendations are given:

- Named entities can be all lower case for single-token entities, such as “weapon” or “price”, for example, and should be camel case starting with a capital letter for composited entities, such as “AssaultRifle” or “MachineGun”, for example.
- Relations start in lower case and use camel case for composited names, such as “hasPrice” or “isSellingTo”, for example.

4.2.2.1. Overview

When opening the *Recogito* web page, the following view will be available in the web browser:



You are annotating entities

Different Beretta models and Glock



I have a Beretta 92G Elite II I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

By default, the application starts in the “annotation mode” which is shown on the top left by the notice “You are annotating entities”. This switch can be toggled to change between “relationship mode” and “annotation mode”. On the top right, there is a pulldown menu which allows choosing from different examples which were provided for the demonstration. In the middle, you can find the area for the text to be annotated. And, at the bottom, there is a download button which allows downloading the results of the annotation process.

4.2.2.2. Creating annotations

In order to create annotations, the user must be in “named entity annotation mode”, which means that the switch at the top left has a blue background and shows the notice “You are annotating entities”.

An entity is marked up by holding down the mouse button to mark the beginning of the named entity and release the button when the end of the named entity string is reached. Single words can also be selected by double clicking them. Selection will open a popup element below the named entity which points to it and contains a field with the “Add tag ...” placeholder.



You are annotating entities

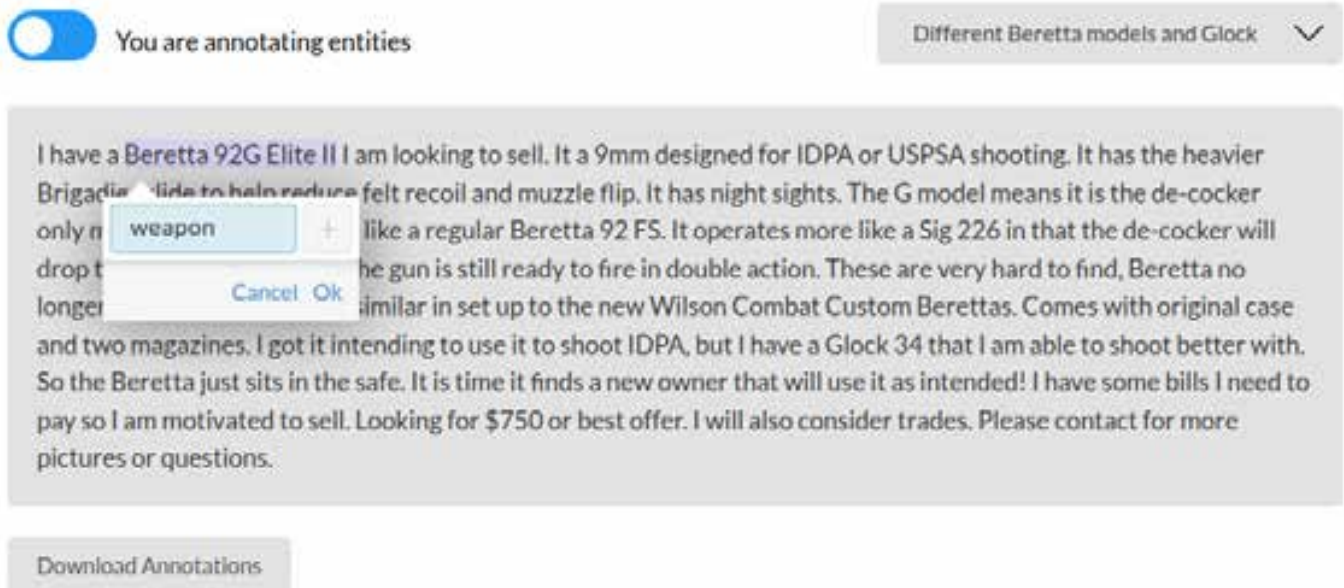
Different Beretta models and Glock



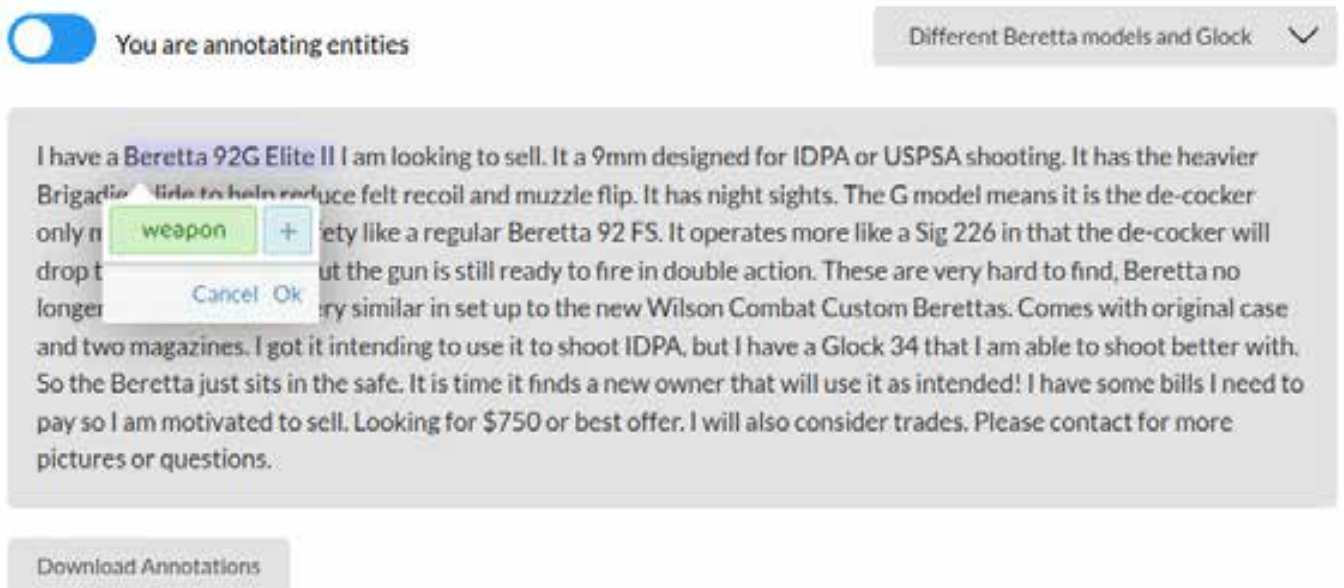
I have a Beretta 92G Elite II I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

The named entity type is typed into this field as shown in the example below, where the named entity type “weapon” is indicated for the string “Beretta 92G Elite II”.



The keyboard can be used to add the entity. Once a named entity type is added, it is highlighted in green. Pressing the Enter key once will add the tag and allow entering more named entity types by pressing the “plus” symbol on the right side of the tag field.



Pressing either the “Ok” button in the bottom-right corner of the tag entry dialog using the mouse, or pressing Ctrl-Enter (the two keyboard keys “Ctrl” and “Enter” at the same time) will add the annotation and close the tag entry dialog to finalize adding entities for the given string. Once this is done, the marked-up string is underlined and highlighted in orange as shown below.

☒ You are annotating entities

Different Beretta models and Glock

I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

By hovering over the marked-up text parts, the tag or tags which were annotated for the string are shown in a pop up as shown in the following screenshot:

☒ You are annotating entities

Different Beretta models and Glock

weapon

I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

The process of entering tags is repeated for other named entities in the text, as, for example, the amount of money which is going to be tagged in the example below.

☒ You are annotating entities

Different Beretta models and Glock 

I have a **Beretta 92G Elite** ^{weapon} looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

Add tag_ +
Cancel Ok

Download Annotations

The “\$750” string will be tagged as a “price” named entity:

☒ You are annotating entities

Different Beretta models and Glock 

I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

price +
Cancel Ok

Download Annotations

As a result, two named entities have been tagged. It is now possible to switch to the “relationship annotation mode” by toggling the switch in the top left corner, so that it changes the color to orange and displays “You are drawing relations”.



You are drawing relations

Different Beretta models and Glock



I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for **\$750** or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

A relationship between two entities is drawn by, either: pressing the mouse button on the first entity, then holding it down while moving the mouse to the second entity and then releasing the mouse button; or by clicking on the first entity once, then moving the mouse over the second entity and clicking again. A dotted line is drawn between the entities and a text input field opens where the type of relationship can be entered.



You are drawing relations

Different Beretta models and Glock



I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for **\$750** or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

In this example, a relation named “hasPrice” is defined between the “weapon” entity “Beretta 92G Elite II” and the “price” entity “\$750”.

☒ You are drawing relations

Different Beretta models and Glock

I have a **Beretta 92G Elite II** I am looking to sell. It is a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for **\$750** or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

Note that a relation is directed which means that it matters where the relation starts and where it ends. An arrow pointer right after the text box indicates the direction of a relation. Bi-directional relationships do not need to be annotated separately but can apply by convention. For example, by defining the relation “isBuyingFrom” between a “buyer” and a “seller” named entity, the convention can define (as part of an ontology) that the inverse relationship “isSellingTo” also applies.

☒ You are drawing relations


Different Beretta models and Glock

I have a **Beretta 92G Elite II** I am looking to sell. It is a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for **\$750** or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

In the top right corner, there is a select box which provides several examples which were prepared for the demonstration of the *RecogitoJS* component. It is possible to change from one example to the other.

Important: The annotations are stored in the browser’s local storage. Annotations can be downloaded and stored as a file.

 You are drawing relations


I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for **\$750** or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

Different Beretta models and Glock

- Different Beretta models and Glock
- Beretta, contact and price
- Smith & Wesson with price (type gun)
- Winchester, multiple labels**
- Ammunition
- Customized listing ruger
- Conversion of civilian to military weapon
- Drug offer

After the annotations have been added, please make sure to store the results. In the bottom left corner, there is the “Download Annotations” button which allows storing a JSON file which contains the annotations added to the example documents. The Download Annotations button will export all annotations, from all sample texts together. Therefore, it is not necessary to make a separate download for each sample text.

 You are drawing relations

Different Beretta models and Glock

I have a **Beretta 92G Elite II** I am looking to sell. It a 9mm designed for IDPA or USPSA shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two magazines. I got it intending to use it to shoot IDPA, but I have a Glock 34 that I am able to shoot better with. So the Beretta just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for **\$750** or best offer. I will also consider trades. Please contact for more pictures or questions.

Download Annotations

4.3. Ethical, Legal and privacy challenges

There are no ethical, legal, and privacy challenges to be considered for the *RecogitoJS* component itself, because the component relies on data to be provided by the environment in which it is embedded. For this reason, potential issues will be explained in the corresponding section 5.3 because *RecogitoJS* is integrated in the dataset repository which exposes the data for annotation.

5. Presentation of the Tool “Data Set Repository”

This section contains information about the component “Data Set Repository” (DSR).

5.1. Demonstration of the component

5.1.1. Component installation

Pre-requisite: The OS of the host machine (or VM) should be Ubuntu 18.04. Docker should be installed (see **¡Error! No se encuentra el origen de la referencia.**).

Installation:

- Obtain the source code for DSR:

```
dsrc_source_v0.1.tar.gz
```

The configuration file `settings/settings.cfg.docker` is used for the docker deployment.

Note that this file is copied to the docker container as the settings file:

```
dsrc1:/dsrc/settings/settings.cfg
```

5.1.1.1. Build

Build the docker containers from source:

```
docker-compose build
```

5.1.1.2. Run

Start the docker containers:

```
docker-compose up
```

Open the web page (where xxx.xxx.xxx.xxx is the IP address of the server on which the component is deployed) and open the web application running on port 5801:

```
http://xxx.xxx.xxx.xxx:5801
```

A task monitoring service is available on port 5805:

```
http://xxx.xxx.xxx.xxx:5805
```

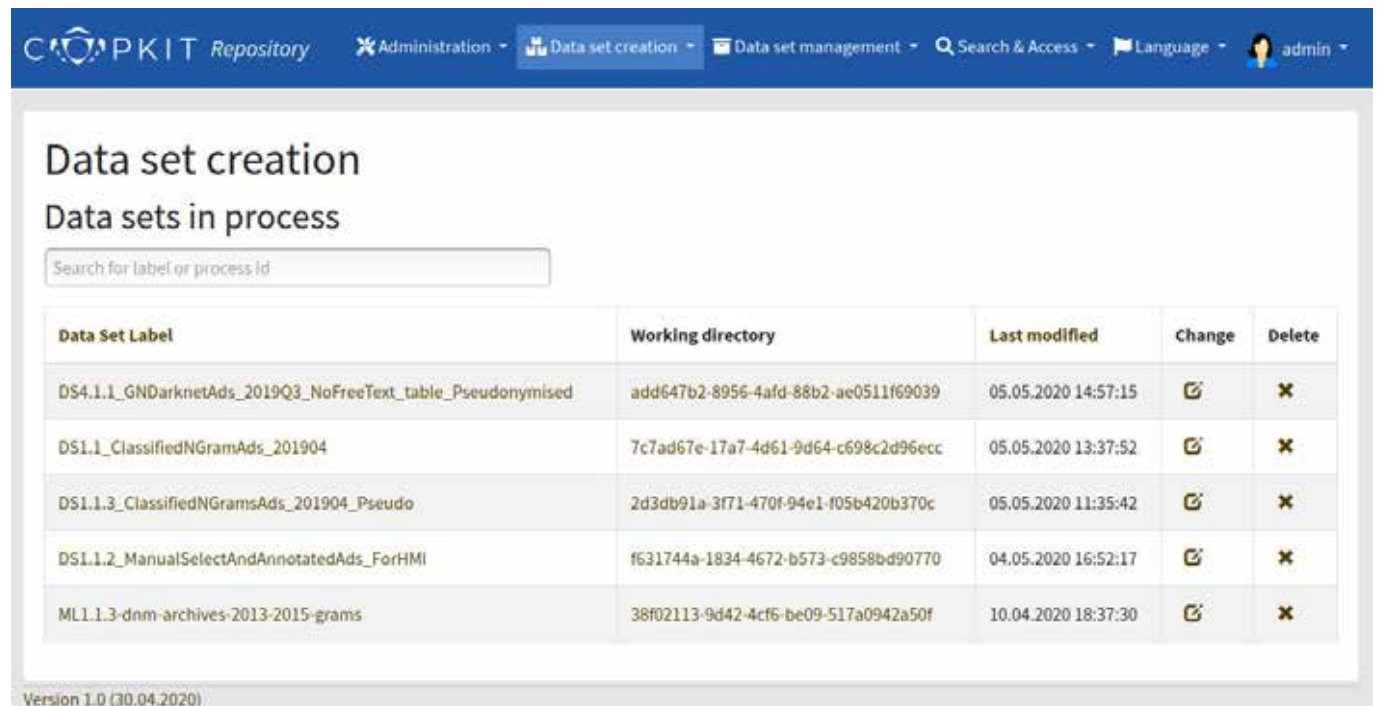
5.1.2. Component usage

The Data Set Repository (DSR) is used

1. to manage data sets created as part of the activities in Task T4.1,
2. to demonstrate the integration of the annotation tool Recogito with the possibility to improve an existing training/test data set,
3. and to demonstrate the model creation process together with a supervised learning approach.

Regarding 1), the screenshot in Figure 8 shows a list of data sets collected as part of T4.1. Dataset “ML1.1.3-dnm-archives-2013-2015-grams” was used to demonstrate the model creation for recognizing

weapon named entities. This is a publicly available dataset which was pseudonymized by the technical team to allow using it for demonstration purposes.



The screenshot shows the 'Data set creation' page of the COPKIT Repository. It features a search bar for labels or process IDs and a table listing data sets in process. The table has columns for Data Set Label, Working directory, Last modified, Change, and Delete. The data sets listed are DS4.1.1_GNDarknetAds_2019Q3_NoFreeText_table_Pseudonymised, DS1.1_ClassifiedNGramAds_201904, DS1.1.3_ClassifiedNGramsAds_201904_Pseudo, DS1.1.2_ManualSelectAndAnnotatedAds_ForHMI, and ML1.1.3-dnm-archives-2013-2015-gramis.

Data Set Label	Working directory	Last modified	Change	Delete
DS4.1.1_GNDarknetAds_2019Q3_NoFreeText_table_Pseudonymised	add647b2-8956-4afd-88b2-ae0511f69039	05.05.2020 14:57:15		
DS1.1_ClassifiedNGramAds_201904	7c7ad67e-17a7-4d61-9d64-c698c2d96ecc	05.05.2020 13:37:52		
DS1.1.3_ClassifiedNGramsAds_201904_Pseudo	2d3db91a-3f71-470f-94e1-f05b420b370c	05.05.2020 11:35:42		
DS1.1.2_ManualSelectAndAnnotatedAds_ForHMI	f631744a-1834-4672-b573-c9858bd90770	04.05.2020 16:52:17		
ML1.1.3-dnm-archives-2013-2015-gramis	38f02113-9d42-4cf6-be09-517a0942a50f	10.04.2020 18:37:30		

Version 1.0 (30.04.2020)

Figure 8 COPKIT Data Set Repository (DSR)

The Data Set Repository (DSR) serves as a shared data storage to store data sets and processing outcomes.

Regarding 2), the screenshot in Figure 9 shows the integration of the *RecogitoJs* component into the DSR. As shown in this screenshot, the *RecogitoJs* component displays existing annotations (in this example: “calibre”: “9mm” and “weapon-type”: “pistol”) from the ground truth data set. Existing annotations can be changed, or new annotations can be added (in this example “HK SP5K is annotated as “weapon”). The result can be stored to create an improved version of the training/test data set.

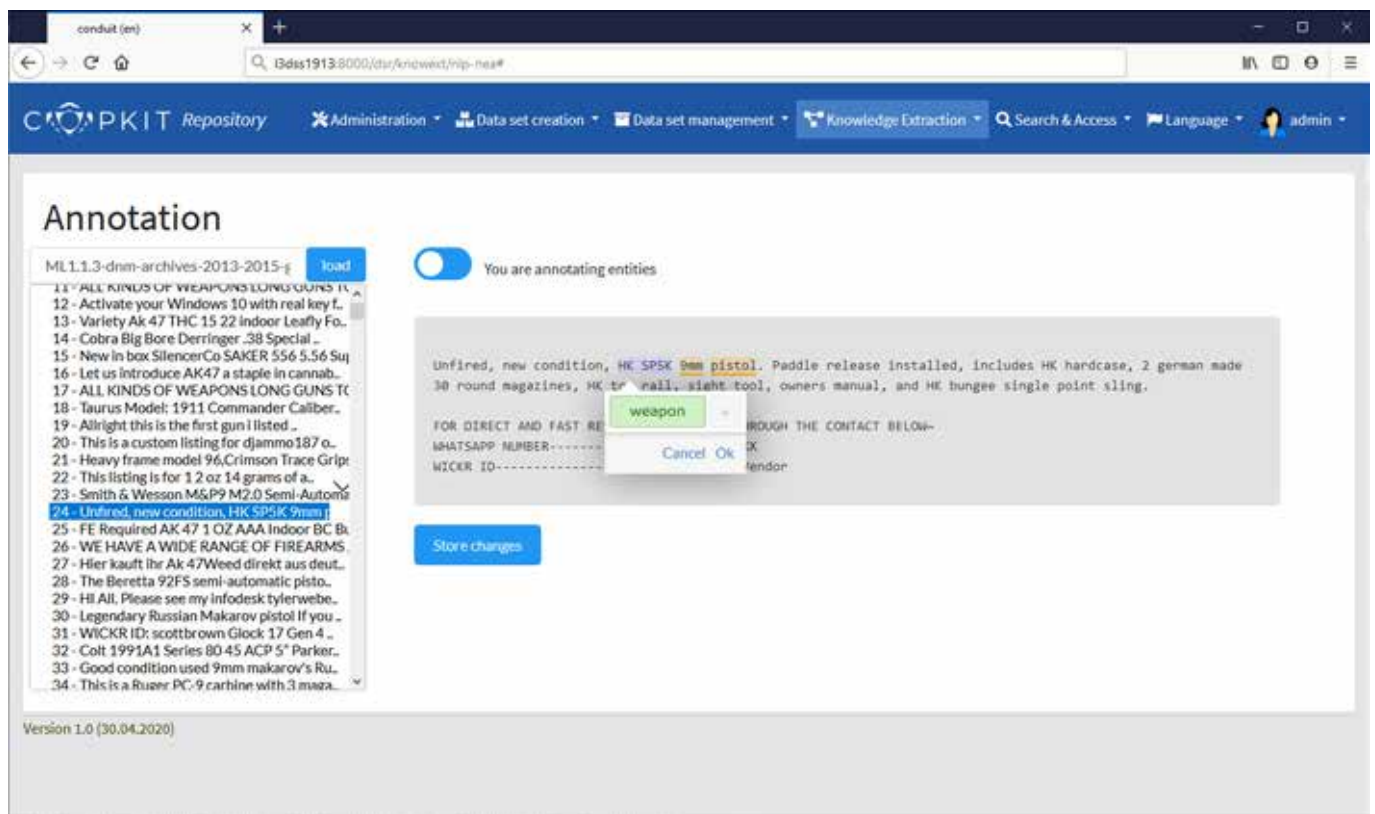


Figure 9 Displaying and changing annotations of weapon named entities

Regarding 3) the screenshot in Figure 10 shows how the input for the model creation is defined by selecting the data source from the DSR. By convention, the data source must contain a data set representation with the label "mldata" to use it for building a model (in this case: a named entity recognition model for weapon named entities).

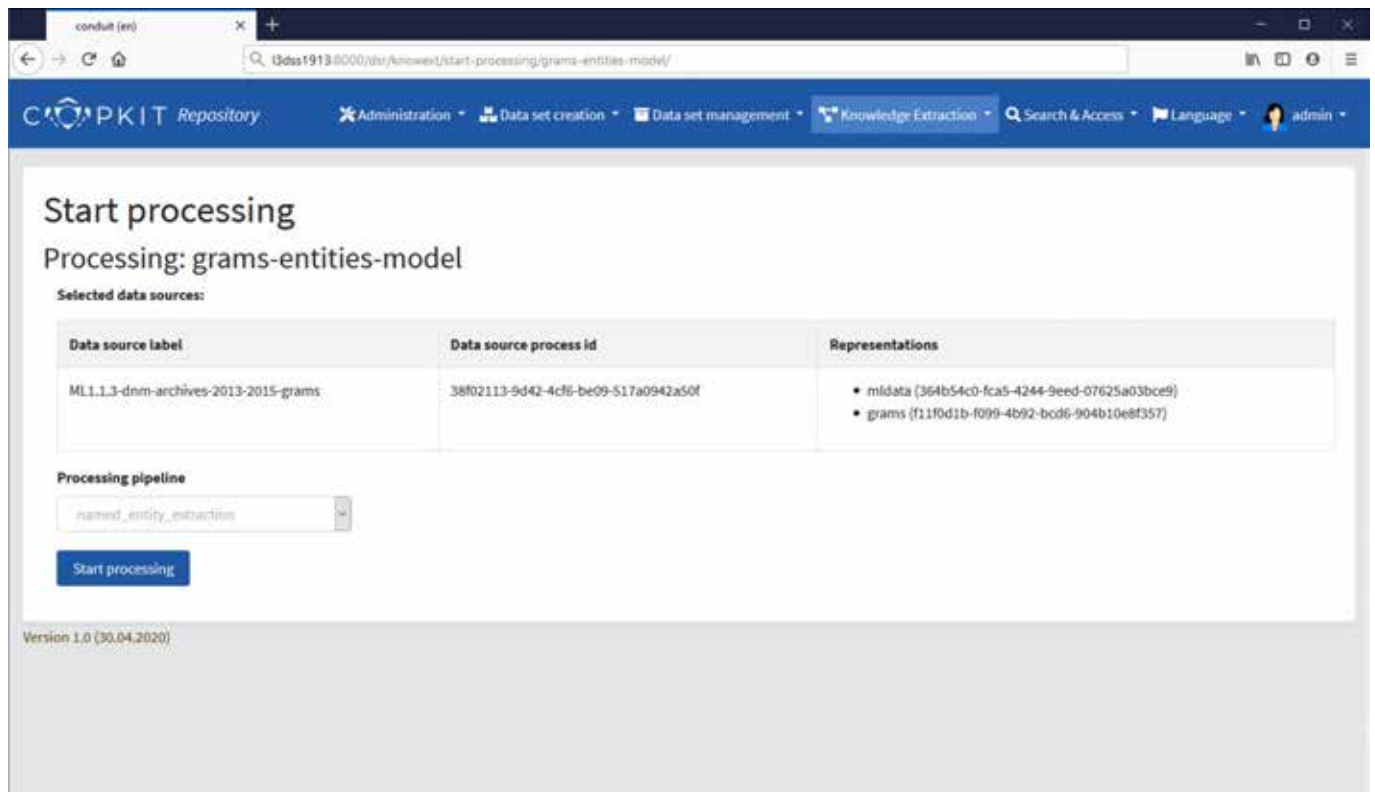


Figure 10 Starting the model creation for named entity recognition

And the screenshot in Figure 11 shows how the named entity recognition model is created using the input data source. It is possible to follow the log of the model creation process which at the end indicates the results of the performance evaluation using the test data set included in the “mldata” representation of the data set used.

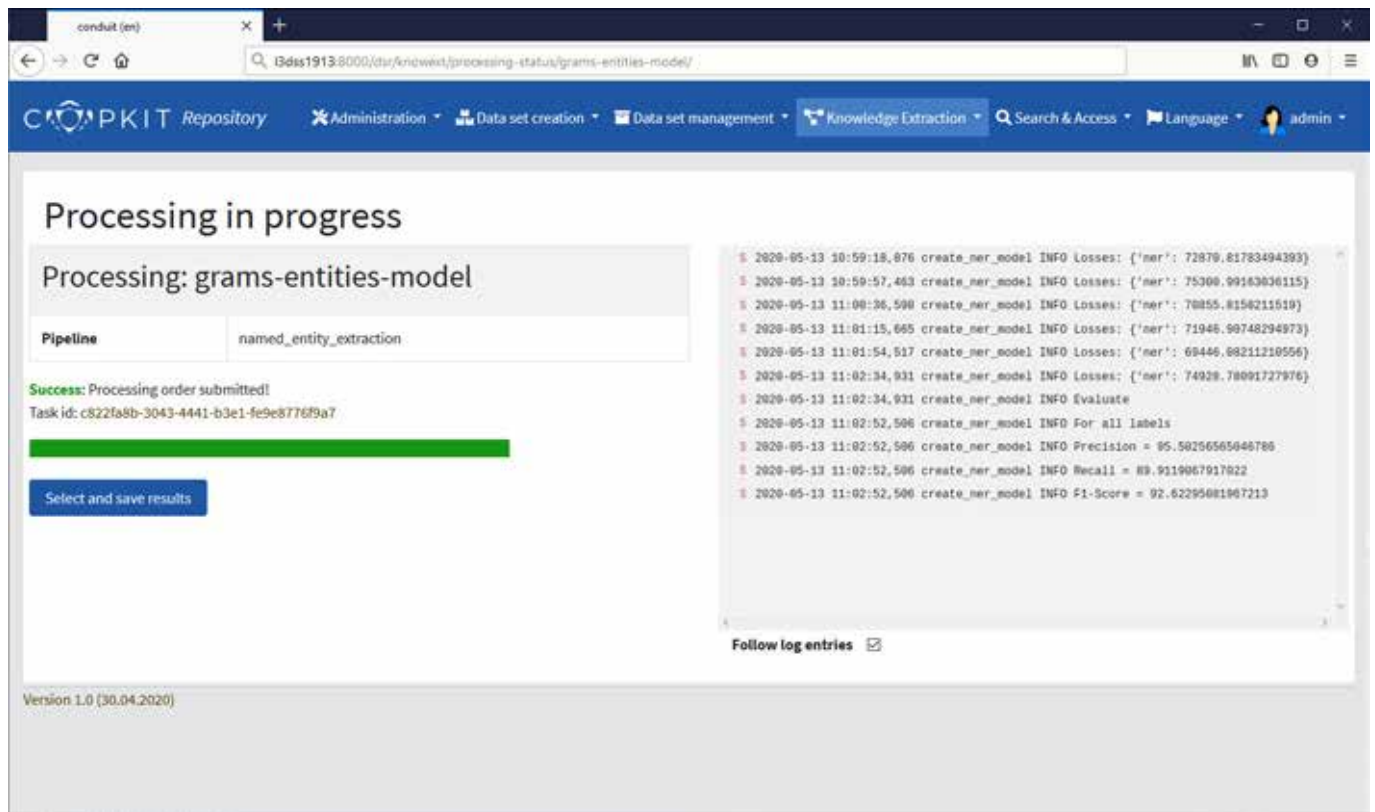


Figure 11 Creating the named entity recognition model

It is possible to inspect the log when the model creation process is finished, as shown in Figure 12 or to look into the model creation metadata ("processing_metadata.json" – an example is shown in "Annex III. Example of model evaluation metadata" on page 40).

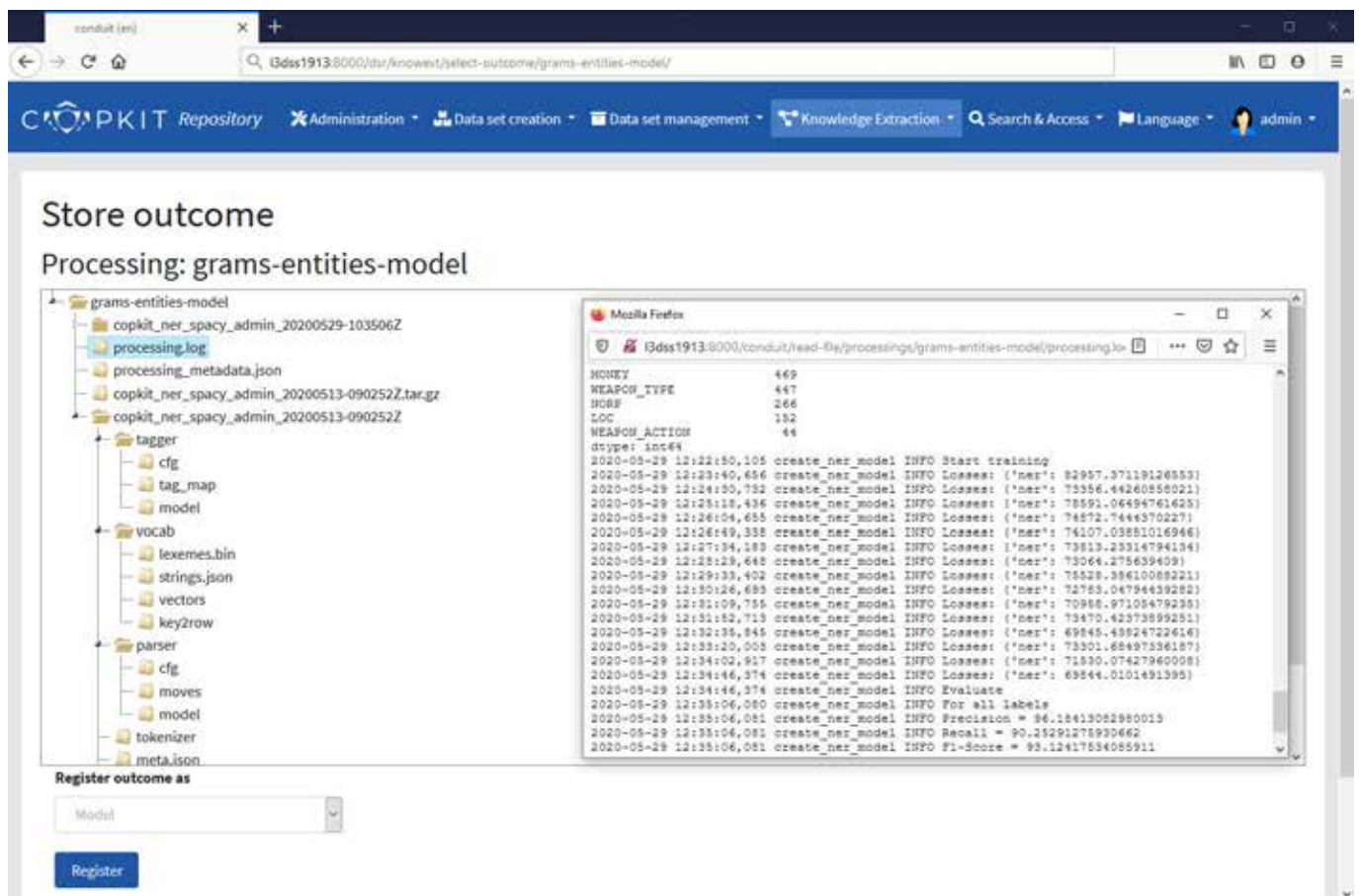


Figure 12 Inspect processing log when the model creation process is finished

The relevant indicators for the performance of the model are in this example “Precision = 96.18%”, “Recall = 90.25%”, and “F1-Score=93.12%”.

5.2. Technical description

The Data Set Repository is an application which allows managing datasets and serves as an integration environment for WP4 components. It provides REST interfaces to allow programmatically integrating with other components, for example, to exchange data between WP4 and WP6.

Furthermore, the Data Set Repository is the system environment in which the RecogitoJs component described in section 4 is integrated. This allows rendering named entity labels suggested by existing models and create new versions of the model after improving the annotations.

The Data Set Repository provides a frontend web application together with a task execution system based on Celery¹⁸ which allows synchronous and asynchronous processing of data sets. It is a Python/Django-based web application which uses a MySQL database for storing information about data sets and a Celery/RabbitMQ/Redis backend for asynchronous task processing.

The Data Set Repository is prepared for container-based deployment based on Docker¹⁹ in order to support simple and modular installation of the software in cloud environments. Docker is an open-source

¹⁸ <http://www.celeryproject.org>

¹⁹ <https://www.docker.com>

engine that automates the deployment of any application as a lightweight and portable container that will run on any platform where the Docker engine is supported.²⁰

The software modules for Docker deployment are the following:

- MySQL²¹
- Solr²²
- RabbitMQ²³
- Redis²⁴
- DSR
- Celery²⁵
- Flower²⁶

5.2.1. Data Set Metadata

A technical team with participants from WP4 and WP6 defined a core set of metadata for storing data sets which allows searching in metadata or in the full-text of files stored as part of the data sets.

Table 3 shows the metadata which was defined for datasets in the COPKIT project.

Identifier	Description	Example	Cardinality
identifier	Identifier of the dataset (URN-UUID).	"urn:uuid:42658bbd-a76f-46f5-85da-f0ad2bed94dc"	1
version	Version number of the dataset.	1	1
label	The label is a short URI suitable form of the title without white spaces using alphanumeric characters plus hyphen ('-'), underscore ('_'), dot ('.'), and tilde ('~') ²⁷	"example_crawl_09-2019"	1
title	Title of the dataset.	"Example Crawl 09-2019"	1
description	Description of the dataset.	"A collection of advertisements for weapons and drugs"	1
tags	List of keywords or tags assigned to the dataset.	"weapons", "drugs"	0...N
language	Main language of the dataset.	"en"	0..1
contactPoint	Person or organization that can be contacted regarding questions related to the dataset.	"COPKIT project"	1
contactEmail	Email address of the person or organization that can be contacted regarding questions related to the dataset.	"copkit@copkit.eu"	1
representation	A data container of the dataset. This can be the data in a specific format, or a derivative of the original data, for example.	See Table 4.	1...N

²⁰ <https://docs.docker.com/engine/installation>

²¹ <http://www.mysql.com>

²² <https://lucene.apache.org/solr>

²³ <http://www.rabbitmq.com>

²⁴ <http://redis.io>

²⁵ <http://www.celeryproject.org>

²⁶ <https://github.com/mher/flower>

²⁷ Unreserved characters in RFC3986, see <https://tools.ietf.org/html/rfc3986#section-2>

Identifier	Description	Example	Cardinality
maintainer	Person or organization maintaining the dataset.	"COPKIT project"	1
maintainerEmail	Email address of the person or organization maintaining the dataset.	"copkit@copkit.eu"	1
sourceDate	Date when the original dataset was created or published.	"22.05.2019"	1
created	Date when the dataset was created in the repository.	"2020-04-10T16:37:30Z"	1
lastChange	Date when metadata or data of the dataset was changed the last time.	"2020-04-10T16:37:30Z"	1

Table 3 Dataset metadata

Identifier	Description	Example	Cardinality
identifier	Identifier of the representation (UUID)	"f11f0d1b-f099-4b92-bcd6-904b10e8f357"	1
label	The label is a short URI suitable form of the tile without white spaces using alphanumeric characters plus hyphen ('\'), underscore ('\'), dot ('\'), and tilde ('\~\') ²⁸	"crawl"	1
description	Description of the dataset.	"Harvested data"	1
maintainer	Person or organization maintaining the representation.	"COPKIT project"	1
maintainerEmail	Email address of the person or organization maintaining the representation.	"copkit@copkit.eu"	1
created	Date when the representation was created.	"2020-04-10T16:37:30Z"	1

Table 4 Representation metadata

Listing 2 shows an example of a metadata file in JSON format.

5.2.2. REST API

The Data Set Repository provides a REST API using the Swagger²⁹ framework for API provisioning. The API is generated automatically using the Django REST Swagger framework³⁰ which allows implementation and documentation of the API by using annotations and comments of API view functions.³¹ **Error! No se encuentra el origen de la referencia.** The swagger API UI is actionable, i.e. it is possible to test API calls using the web user interface and following the API documentation available in the DSR menu at "Administration > REST API" or at:

`http://{server-name}:{server-port}/dsr/api`

Web request tools, such as cURL³² can be used to test the API functions.

²⁸ Unreserved characters in RFC3986, see <https://tools.ietf.org/html/rfc3986#section-2>

²⁹ <https://swagger.io>

³⁰ <https://django-rest-swagger.readthedocs.io/en/latest>

³¹ <https://gitlab.com/datamarket/conduit/blob/master/api/views.py>

³² <https://en.wikipedia.org/wiki/CURL>

5.3. Ethical, Legal and privacy challenges

The COPKIT Data Set Repository is the most critical component regarding ethical, legal, and privacy challenges. It serves as a data hub and integration point for technical work packages and contains data from several crawls which might also contain personal information about sellers and buyers in darknet markets.

The technical team is aware that careful planning of security measures is required when this component is being set in place.

Therefore, a set of principles apply:

- The component is only installed in the protected demonstration environment where only a limited set of project partners can access via VPN.
- External access or user accounts outside of the project consortium are not foreseen as part of the WP4 tasks.
- Datasets will be maintained by a limited group (about 4 persons) from the technical team.
- As an additional security measure, data set representations that contain originally harvested data will be encrypted using GPG³³, so that only developers who have the required key can use these files. For demonstration purposes, only the pseudonymized tabular data will be used.
- The purpose of the data repository is to manage different versions of the datasets with their metadata and derivatives that are created (e.g. machine learning models).
- Data access is planned via defined interfaces which allow exchanging information between WP4 and WP5. This is a planned information exchange in line with the DoW: "AIT and UGR collaborate on using and improving the COPWIK knowledge base. On the one side, this is to make sure that COPWIK vocabulary can be used for annotation and, on the other side, that outcomes of annotation and machine learning can feed back into the knowledge base."

As already mentioned in section 3, the data harvested from darknet sources contains personal information, such as email addresses, phone numbers, bank account numbers, as well as contact information for various social media communication platforms, such as Telegram, WhatsApp, Skype, etc. However, it must be pointed out that the data may still contain hints and traces to which darknet marketplace accounts were involved in the publishing of advertisements. The technical team will take all measures available to remove traces to the largest possible extent.

In section 3 it was explained, that distinguishing different types of data set components allows defining a policy which regulates access to data sets on the level of data set components, called data set "representations". Based on these concepts, a policy can define the periods for which it is allowed to keep a working copy, the components which are allowed to be stored as the data set storage entity for the long term, which dataset components are allowed for a specific purpose, etc.

At the same time, the identification of datasets and their derivatives allows keeping track of users that access the data or apply a processing pipeline, and also preserve the relationship of the machine learning models and the data sources they depend on. Standard logging procedures will be applied to build a protocol of activities concerning the use of data sets.

³³ <https://gnupg.org/>

6. Conclusion and next steps

Task 4.5 establishes the common ground for task T4.3 where named entity recognition was established by providing custom NER models that were trained on darknet market datasets related to drugs and weapons and task T4.4, the relationship extraction which combines rule-based and approaches to extract features.

The next steps are to:

- Improve integration of RecogitoJs with the Data Set Repository (DSR).
 - Stabilize the annotation, ground truth, and model creation process.
- Providing pre-defined vocabulary for RecogitoJs annotation based on vocabulary received from the COPWIK knowledge base using a REST API provided by WP5.

Annex A. Ethical review results and response

This deliverable was reviewed by the ELP team of the COPKIT project, more precisely by our parent num. 8. Law and Internet Foundation from Bulgaria.

From ethical, legal and privacy perspective, the Annotation Tool for the Law Enforcement Domain need to further provides opportunities for the potential end-users to precise the scope of data collection, processing, as well to enable them to assign tailor policies in view of storage and access to the data – both the Harvested data and the Working Directories. Ideally, the end users are to be able to set criteria specifying the time limits within which data are to be stored with regards to their national legislation and Art. 5 of Directive 2016/680.

Furthermore, it is also recommended to explicitly indicate at which point the data pseudonymization is be carried out. In order to prevent any possibility of misuse of personal data, it should be carried out at the stage of data collection.

Additional safeguards should be implemented to prevent misuse of data. The COPKIT Data Set Repository should integrate different level of access not only for the developers but also for the users within the LEAs which will ensure that only authorized personnel will have access to the available data. The embedding of appropriate logging procedures is a welcomed step in this direction.

In view of the ethical aspects of the employment of such a tool, it should be very clearly noted that no automated decisions are previewed, and human intervention will be sought in order to define any relationship, and potentially start the investigation process. The tool is to be only developed in a way that underlines this important relationship.

From legal point of view, aspects related to fair trial and the presumption of innocence should be also considered. The tool might underline existing relationships, however the results yielded from its application and to be always regarded as part of evidence collection. In order to strengthen the procedural rights of the persons suspect or accused of crime, the tool and a description of its modus operandi should be made available to experts appointed by the Court the evaluate and present an expert statement.

Overall, the presented tool is in line with the relevel ethical, legal and privacy requirements – there are suitable safeguards in place such as pseudonymization, access rights and logging procedures that that will ensure compliance with data protection by design and by default and with other fundamental principles of the EU Data Protection Framework, including Directive 2016/680. This is really reassuring in order to accomplish transparency and accountability of the data processing activities that are exercised not only under the COPKIT project but also in the operational environment once the tool is adopted by LEAs.

The authors made their best to address these comments along the document although some comments would need to be discussed further after the submission of this deliverable.

Annex I. Example of web annotations

```
[{
  "text": "Different Beretta models and Glock",
  "annotations": [{
    "type": "Annotation",
    "body": [{
      "type": "TextualBody",
      "value": "weapon",
      "purpose": "tagging",
      "confirmed": true
    }],
    "target": {
      "selector": [{
        "type": "TextQuoteSelector",
        "exact": "Beretta 92G Elite I"
      }, {
        "type": "TextPositionSelector", "start": 9, "end": 28
      }]
    },
    "@context": "http://www.w3.org/ns/anno.jsonld",
    "id": "#54887600-89fa-11ea-9a32-0bd2b41ab69b"
  }, {
    "type": "Annotation",
    "body": [{
      "type": "TextualBody",
      "value": "price",
      "purpose": "tagging",
      "confirmed": true
    }],
    "target": {
      "selector": [{
        "type": "TextQuoteSelector",
        "exact": "$750"
      }, {
        "type": "TextPositionSelector", "start": 879, "end": 883
      }]
    },
    "@context": "http://www.w3.org/ns/anno.jsonld",
    "id": "#5869a7d0-89fa-11ea-9a32-0bd2b41ab69b"
  }, {
    "@context": "http://www.w3.org/ns/anno.jsonld",
    "type": "Annotation",
    "id": "#5a238cd0-89fa-11ea-9a32-0bd2b41ab69b",
    "body": [{
      "type": "TextualBody",
      "value": "hasPrice",
      "purpose": "tagging"
    }],
    "target": [{
      "id": "#54887600-89fa-11ea-9a32-0bd2b41ab69b"
    }, {
      "id": "#5869a7d0-89fa-11ea-9a32-0bd2b41ab69b"
    }],
    "motivation": "linking"
  }]
}]
```

Listing 1 Annotations stored in W3C web annotations in JSON-LD

Annex II. Example of data set metadata

```
{
  "identifier": " urn:uuid:42658bbd-a76f-46f5-85da-f0ad2bed94dc ",
  "version": 1,
  "label": "grams-crawl-2015",
  "title": "Grams Crawl 2015",
  "description": " A collection of advertisements for weapons and drugs",
  "provenance": "Collected by Gwern; machine learning data by COPKIT",
  "tags": [
    "weapons ",
    "drugs",
  ],
  "contactPoint": "Person to contact",
  "contactEmail": "copkit@copkit.eu",
  "maintainer": "Creator to contact",
  "maintainerEmail": "copkit@copkit.eu",
  "language": "en",
  "sourceDate": "22.05.2019",
  "lastChange": "2020-04-10T16:37:30Z",
  "created": "2020-04-10T16:37:30Z",
  "representation": {
    "364b54c0-fca5-4244-9eed-07625a03bce9": {
      "label": "mldata",
      "description": "Training and test data.",
      "maintainer": "Creator to contact",
      "maintainerEmail": "copkit@copkit.eu",
      "created": "2020-04-10T16:37:30Z",
      "access": "limited"
    },
    "f11f0d1b-f099-4b92-bcd6-904b10e8f357": {
      "label": "crawl",
      "description": "Harvested data",
      "maintainer": "Creator to contact",
      "maintainerEmail": "copkit@copkit.eu",
      "created": "2020-04-10T16:37:30Z",
      "access": "limited"
    }
  }
}
```

Listing 2 Metadata of a data set in JSON format

Annex III. Example of model evaluation metadata

```
{
  "out": "/var/data/repo/processing/admin/grams-entities-model",
  "createdBy": "admin",
  "performance": {
    "f": {
      "WEAPON_TYPE": 99.68152866242038,
      "DRUG": 95.3416149068323,
      "DATE": 88.45528455284553,
      "LOC": 87.00564971751412,
      "GPE": 93.58752166377815,
      "WEAPON_CALIBRE": 96.72977624784853,
      "PRODUCT": 77.99315849486888,
      "WEAPON": 98.13953488372094,
      "WEAPON_MANUFACTURER": 98.60031104199068,
      "MONEY": 97.79005524861878,
      "NORP": 93.43629343629344,
      "QUANTITY": 85.66493955094991,
      "WEAPON_ACTION": 97.95918367346938,
      "COMBINED": 92.57042767479199
    },
    "p": {
      "WEAPON_TYPE": 100.0,
      "DRUG": 98.71382636655949,
      "DATE": 92.83276450511946,
      "LOC": 86.51685393258427,
      "GPE": 93.42560553633218,
      "WEAPON_CALIBRE": 98.59649122807016,
      "PRODUCT": 86.36363636363636,
      "WEAPON": 98.5981308411215,
      "WEAPON_MANUFACTURER": 100.0,
      "MONEY": 100.0,
      "NORP": 97.58064516129032,
      "QUANTITY": 86.7132867132867,
      "WEAPON_ACTION": 100.0,
      "COMBINED": 95.16806722689076
    },
    "r": {
      "WEAPON_TYPE": 99.36507936507937,
      "DRUG": 92.1921921921922,
      "DATE": 84.472049689441,
      "LOC": 87.5,
      "GPE": 93.75,
      "WEAPON_CALIBRE": 94.93243243243244,
      "PRODUCT": 71.10187110187111,
      "WEAPON": 97.68518518518519,
      "WEAPON_MANUFACTURER": 97.23926380368098,
      "MONEY": 95.67567567567568,
      "NORP": 89.62962962962962,
      "QUANTITY": 84.64163822525597,
      "WEAPON_ACTION": 96.0,
      "COMBINED": 90.11082693947145
    }
  },
  "modelCreationDate": "2020-04-24T19:02:20Z",
  "modelPath": "/var/data/repo/processing/admin/grams-entities-model/copkit_ner_spacy_admin_20200424-190218Z",
  "modelFramework": {
```



```
{
  "name": "spacy",
  "version": "2.1.0"
},
{
  "baseModel": "en_core_web_lg-2.1.0",
  "modelName": "copkit_ner_spacy_admin_20200424-190218z",
  "modelTags": [
    "nlp",
    "ner",
    "weapons"
  ]
}
```

Listing 3 Model creation metadata in JSON format

Annex IV. Batch Data Set Submission using the Data Set Repository API

The API of the DSR allows creating data sets, uploading data, and start processing the data.

Each request must be executed with a user token header which is omitted in the query examples below for the sake of readability:

```
-H 'Authorization: Token $usertoken'
```

The procedure is as follows:

1. Initialise a new data package:

```
curl -X POST -d 'package_name=genesis.harvest.scraped.20191024'
http://$server:$port/dsr/api/datasets/
```

In case of success the HTTP response code is "201 CREATED" and a new process ID (`process_id`) is returned which is required for uploading data in a subsequent step.

```
{
  "process_id": "73483984-debd-4d04-a14c-5acb11167719",
  "work_dir": "/var/data/repo/work/73483984-debd-4d04-a14c-5acb11167719",
  "package_name": "genesis.harvest.scraped.20191024",
  "version": 0,
  "last_change": "2020-03-20T15:38:23.026106+01:00"
}
```

2. Upload data file (here a CSV file `/home/$user/datafile.csv` (note that the process ID (`process_id`) returned by the previous request is used in this request to identify the target data set where the files are going to be uploaded).

For the first file of a representation only the process ID (`process_id`, here: 73483984-debd-4d04-a14c-5acb11167719) needs to be provided and the representation ID can be omitted:

```
curl -F "file=@/home/$USER/datafile.csv"
http://$server:$port/dsr/api/datasets/73483984-debd-4d04-a14c-5acb11167719/data/upload/
```

This will generate a random UUID identifier (`representationId`) for the representation which is returned as part of the response message in case of success:

```
{
  "message": "File upload successful",
  "sha256":
"7c10a5a8e79989b608d5e63ed58c031676f43ee4cc01a00d013400941cf7f2d1",
  "processId": "73483984-debd-4d04-a14c-5acb11167719",
  "representationId": "5ed2c8b6-4f4b-46f7-a1f3-192472a76a41"
}
```

Additionally, the `sha256` hash sum allows verifying if the file was uploaded correctly.

If a file needs to be added to a representation, the representation ID (`representationId`, here: `5ed2c8b6-4f4b-46f7-a1f3-192472a76a41`) is given as a parameter after the process ID.

```
curl -F "file=@/home/$USER/datafile.csv"
http://$server:$port/dsr/api/datasets/73483984-debd-4d04-a14c-
5acb11167719/5ed2c8b6-4f4b-46f7-a1f3-192472a76a41/data/upload/
```

To upload metadata, a JSON metadata (`metadata.json`) file can be created:

```
{
  "title": "Data set title",
  "description": "Data set description",
  "contactPoint": "Contact",
  "contactEmail": "contact@email.com",
  "publisher": "Publisher",
  "publisherEmail": "contact@email.com",
  "language": "English",
  "representations": {
    "5ed2c8b6-4f4b-46f7-a1f3-192472a76a41": {
      "distribution_label": "csv",
      "distribution_description": "CSV table",
      "access_rights": "limited"
    }
  }
}
```

Note that if metadata for representations should be added, the representation ID in the metadata file must match the ID of the corresponding representation, e.g. as in this example, the one created previously: `5ed2c8b6-4f4b-46f7-a1f3-192472a76a41`.

An example for a metadata upload request is the following:

```
curl -F "file=@/home/$USER/metadata.json"
http://localhost:8000/dsr/api/datasets/cb755987-9e83-4e71-b000-
dea9324e5dea/metadata/upload/
```

which returns a similar response as the data file upload:

```
{
  "message": "File upload successful",
  "sha256":
"7c10a5a8e79989b608d5e63ed58c031676f43ee4cc01a00d013400941cf7f2d1",
```

```
"processId": "73483984-debd-4d04-a14c-5acb11167719"
}
```

3. Update archived dataset without working copy:

If an archived dataset does not have a working copy, it must be checked out first:

```
curl -X POST http://localhost:8000/dsr/api/datasets/urn:uuid:42658bbd-a76f-46f5-85da-f0ad2bed94dc/checkout-working-copy/
```

Which in case of success returns the new process ID (`process_id`) of the working copy and the corresponding job ID which allows monitoring the process.

```
{
  "message": "Checkout request submitted successfully.",
  "job_id": "59a7da2e-7496-4b70-b2c5-1fc1c7a41a02",
  "process_id": "650abd36-1203-4d6b-aa10-d8305271db9b"
}
```

4. Store data:

```
curl http://$server:$port/dsr/api/datasets/cc3e95de-71d9-4e9e-8de7-128a1c92774f/startingest
```

Once the data package is stored, it is indexed and gets an identifier of the form `"urn:uuid:f90668b9-112b-4723-8344-07449e7b657e"`.

The data package can be changed afterwards which increases the version number and the index for the data package is updated.

File upload does not have to be done via this API. It is possible to just create the data folders and transfer the data using other means of file transfer (scp, rsync, etc.).

An example bash script can be found in:

```
dsr/util/scripts/import_script.sh
```

Annex V. T4.5 - Expert annotation and integration of extracted information - Virtual workshop – Usability test



TECHNOLOGY, TRAINING AND KNOWLEDGE FOR EARLY-WARNING / EARLY-ACTION LED POLICING IN FIGHTING ORGANISED CRIME AND TERRORISM

Expert annotation and integration of extracted information (T4.5) virtual workshop – usability test

Grant Agreement: 786687

Project Acronym: COPKIT

Project Title: Technology, training and knowledge for Early-Warning / Early-Action led policing in fighting Organised Crime and Terrorism

Call (part) identifier: H2020-SEC-2016-2017-2

Document ID: CPK-WP04-T4.5-RecogitoJS-UsabilityTest-DokumentationAndResults

Revision: V1.3

Date: 29.05.2020

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Dissemination Level		
PU	Public	<input type="checkbox"/>
CO	Confidential, only for members of the consortium (including the Commission Services)	<input checked="" type="checkbox"/>
EU-RES	Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)	<input type="checkbox"/>
EU-CON	Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)	<input type="checkbox"/>
EU-SEC	Classified Information: SECRET UE (Commission Decision 2005/444/EC)	<input type="checkbox"/>

Revision history

Revision	Edition date	Author	Modified Sections / Pages	Comments
0.1	27/03/2020	BayHfoeD	All	Initial draft for T3.5 virtual workshop
1.0	27/03/2020	BayHfoeD	----	for decision to project coordination team
1.1	06/04/2020	BayHfoeD	No. 2	Modifications after review R.P.P.
1.2	26/05/2020	BayHfoeD	All	Implementation of the rest and evaluation results.
1.3	29.05.2020	AIT	All	Review

Table of Contents

1. Introduction	5
1.1. Background.....	5
1.2. Purpose and Scope.....	5
1.3. Document Structure	5
1.4. Glossary.....	6
2. General objectives of T4.5.....	6
3. RecogitoJS Usability Test	7
3.1. Roadmap	7
3.2. Virtual Workshop.....	7
3.2.1. Execution of the workshop.....	7
3.2.2. Participants.....	7
3.2.3. Platform used	8
3.2.4. Technical Issues.....	8
3.3. Manual LEA Participants.....	8
3.3.1. Manual for Installation RecogitoJS.....	8
3.3.2. Manual for Using RecogitoJS	8
3.4. Usability Test	8
3.4.1. Access to RecogitoJS.....	8
3.4.2. Submission of results.....	8
4. Results of Usability Test Evaluation.....	8
4.1. Structure of the Questionnaire	8
4.1.1. General information about the participants	9
4.1.2. Evaluation of the virtual workshop	9
4.1.3. Evaluation of the RecogitoJS manual	9
4.1.4. Short implemented SUS part	10
4.1.5. Evaluation Usability of RecogitoJS in Detail.....	10
4.1.6. Additional Questions about RecogitoJS.....	11
4.1.7. Open Space for Additions.....	11
5. Ethical issues	13
5.1. Access to the component.....	13
5.2. Evaluation	13

6. Risks for RecogitoJS Usability Test and Mitigation.....	13
7. References.....	13

Tables and Figures

Table 1. Glossary	6
Table 2. Roadmap RecogitoJS Usability Test	7
Table 3. Evaluation of the virtual workshop	9
Table 4. Evaluation of the manual	9
Table 5. Evaluation of the usability in detail	11
Table 6. Additional Questions.....	11

1. Introduction

1.1. Background

The COPKIT project focuses on the problem of analysing, investigating, mitigating and preventing the use of new information and communication technologies by organised crime and terrorist groups. For this purpose, COPKIT proposes an intelligence-led Early Warning (EW) / Early Action (EA) system for both strategic and operational levels. The project duration is 36 months (from 2018 to 2021).

Main objectives of COPKIT project:

- Apply new EW/EA-led policing technologies for improved situational awareness.
- Develop a toolkit for knowledge production and exploitation in investigative and strategic analysis work to support the EW/EA paradigm.
- Ensure that the new tools, detection capabilities and knowledge-sharing respect EU data protection regulation and ethical principles.
- Develop innovative curricula for educating and training LEAs for the methodological and workflow aspects of EW/EA-led policing.

1.2. Purpose and Scope

This document describes the planned roadmap for the next the progress in the T 4.5, focused on planning a virtual workshop for testing the usability of the developed AIT tools.

1.3. Document Structure

The document has the following structure

- Section 2: description of the general objectives of the task 4.5.
- Section 3: description of the RecogitoJS usability test
- Section 4: results of the usability test evaluation
- Section 5: possible ethical issues in context of the usability test
- Section 6: possible risks for RecogitoJS usability test and mitigation

Attached documents:

- CPK-WP04-T4.5-RecogitoJS-Manual-v1.0
- CPK-WP04-T4.5-RecogitoJS-EvaluationForm-v1.0
- CPK-WP04-T4.5-RecogitoJS-OnlineWorkshop-Slides
- CPK-WP04-T4.5-RecogitoJS-UsabilityTest-OnlineMeeting-Minute

1.4. Glossary

Instruction: include the acronyms and key terms used in the document.

Term / Acronym	Explanation
AIT	AIT Austrian Institute of Technology
BayHfoeD	University of Applied Sciences for Public Service in Bavaria, Department of Policing
BFP	Police Federale Belge
COPWIK	COPKIT Knowledge Base
GDCOC	General Directorate Combating Organized Crime
Isdefe	Ingeniería de Sistemas para la Defensa de España, SA-ISDEFE, S.A. S.M.E. M.P.
ESMIR	Ministerio Del Interior
EU	European Union
EW/EA	Early Warning / Early Action
GN	Ministere de l'interieur – Gendarmerie Nationale
KEMEA	Kentro Meleton Asaleias
LEA	Law Enforcement Agency
SPL	State Police of the Ministry of Interior of the Republic of Latvia
TNL	Thales Nederland B.V.
VPN	Virtual Private Network
VTE	Validation Test and Evaluation Environment

Table 1. Glossary

2. General objectives of T4.5

The objective of Task T4.5 is to implement tools for data validation, annotation, and exploration and support LEA partners in applying them.

AIT developed an instance of Recogito¹ named RecogitoJS² to create a web-based tool for the manual verification and correction of the results of automatically annotated texts (and images), as well as for the creation of machine learning ground truth datasets. This will allow integrating relevant data sources and providing a user interface according to the needs of LEA stakeholders.

The LEA users will be instructed how to use the annotation tool for ground truth creation. In the second phase of this task, additional components will be implemented that are not being covered by RecogitoJS alone (expert assessment/correction of information extraction results).

¹ <https://recogito.pelagios.org>

² Modular, client-side web-based annotation tool based on Recogito concepts and technology.

Task leader is BayHfoeD, involved contributors are AIT, ESMIR, GDCOC, GN and BFP.

3. RecogitoJS Usability Test

3.1. Roadmap

No.	Planned steps	completed
01	First consultation between AIT and BayHfoeD about carrying out the virtual workshop	26.03.2020
02	Coordination of planned steps with project coordination	30.03.2020
03	Preparation for the virtual workshop and preparation of the component for the usability test	30.03.2020 till 13.05.2020
04	Virtual Workshop	13.05.2020
05	7-day online usability test (7-day timeframe, desired 2-3 hours testing per LEA)	20.05.2020
06	Evaluation of the results	26.05.2020
07	Complete documentation and evaluation	26.05.2020

Table 2. Roadmap RecogitoJS Usability Test

3.2. Virtual Workshop

3.2.1. Execution of the workshop

The virtual workshop was carried out on 13th of May 2020, 01:00 pm – 02:00 pm (CEST).

The virtual workshop was moderated by AIT and BayHfoeD. In the workshop, after a short introduction of the COPKIT project, necessary background information and the operation of the usability test was explained to the participants. A user manual was distributed to the participants before the workshop.

3.2.2. Participants

Participants	one participant	each	National Police ESP, Guardia Civil, Information Centre (Lativa)
	two participants	each	SPL, KEMEA (GR Police), BFP, BayHfoeD (GER Police)
Presenters	Sven Schlarb	AIT	
	Tobias Mattes	BayHfoeD	
Guests	Emmanouil Kermitsis	KEMEA	
	Franck Mignet	TNL	
	Raquel Pastor Pastor	Isdefe	

3.2.3. *Platform used*

Due to the fact, that the RecogitoJS usability test was not restricted, AIT and BayHfoeD decided to use 'GoToMeeting' as the video conferencing system for the presentation.

3.2.4. *Technical Issues*

During the virtual workshop there were some technical issues (lost connections). For future virtual workshops, all participants should test their hardware (especially microphones) and internet connection in advance. The technical issues were not critical.

3.3. Manual LEA Participants

3.3.1. *Manual for Installation RecogitoJS*

AIT developed a short manual for the installation of the component. It was distributed to the LEA participants prior to the workshop.

3.3.2. *Manual for Using RecogitoJS*

AIT developed a manual with screenshots and user instructions explaining the usage of the component. It was distributed to the LEA participants prior to the workshop.

3.4. Usability Test

3.4.1. *Access to RecogitoJS*

The participants of the online workshop got a test version of RecogitoJS. For the download of the component, a link to the files stored in the own cloud of the BayHfoeD was provided. The testing period was from the 13th of May to the 20th of May. The participants were able to test the component on their own hardware.

3.4.2. *Submission of results*

In the usability Test 12 LEA participants were involved. Seven of them sent back the JSON result file (created by the component) and the evaluation form.

4. Results of Usability Test Evaluation

In advance of the virtual workshop, the participants got an evaluation form that was developed by AIT and BayHfoeD. The participants had to fill it during the seven-day testing timeframe. Attached to this document is a copy of this questionnaire.

4.1. Structure of the Questionnaire

The evaluation form was structured into seven parts. The results of the evaluation will be present in chapter 4.2. The participants were asked about their personal experience, the virtual workshop, the offered manuals, about the functionality and usability of the component. At the end of the questionnaire there was room for additional remarks.

4.1.1. General information about the participants

As mentioned, seven participants sent back the evaluation form [two from KEMEA (incl. Hellenic Police), two from ESMIR (incl. Guardia Civil, Guardia National), two from BayHfoeD (incl. German Police) and one from BFP]. Six of them were aged 31 – 40 and one was 41 – 50. Apart from one participant without experience in the field of cyber-criminal investigations, the other participants each had 1 – 5 years of experience. Three of the participants spent more than three hours in testing the component, three between 2 – 3 hours and one between 1 – 2 hours.

4.1.2. Evaluation of the virtual workshop

The online workshop was evaluated with following questions (1 = I totally disagree; 5 = I totally agree):

Questions	result
The online workshop was well organized	4,1
The online platform was suitable for the intended purpose	4,3
The chosen timeframe for the workshop was enough	4,4
There was enough space to ask questions during the online workshop	4,6
The presentation was well tailored to the audience	4,3

Table 3. Evaluation of the virtual workshop

Additional remarks of the participants:

- The internet connection of the participants should be tested in advance.
- Test microphone before meeting
- Still not sure about the targeted audience of some of the technical slides and explanations, but overall it showed the general idea behind the tool, as well as the usability test was technically easy to do--even without the manual--, no further help needed.
- Technical problems led to misunderstanding of the task and approach to the use of the tool.

4.1.3. Evaluation of the RecogitoJS manual

The offered manual was evaluated with following questions (1 = I totally disagree; 5 = I totally agree):

Questions	result
The needed steps for using the AIT tool were explained step by step in the manual	4,9
I was able to test the functions of the AIT tool by using the manual	4,4
I got support in case of technical issues from the AIT staff	4,7
I got the AIT manual adequate advance	4,6
I got support in case of technical issues from the AIT staff	N/A
I got the AIT manual adeqate advance	4,9

Table 4. Evaluation of the manual

Additional remarks of the participants:

- No suggestions. The manual was very suitable and helpful.
- That was great idea to have the manual beforehand. It was well explained and useful to understand the tool.

4.1.4. Short implemented SUS part

To get a common overview over the usability of the component in the questionnaire a short System Usability Scale (SUS) part was implemented. The SUS provides a “quick and dirty”, reliable tool for measuring the usability. It consists of a 10-item questionnaire with five response options for respondents; from ‘strongly agree’ to ‘strongly disagree’. Originally created by John Brooke in 1986, it allows you to evaluate a wide variety of products and services, including hardware, software, mobile devices, websites, and applications.³ The SUS is not diagnostic, but it gives an easy and short overview, how the participants rate the tool. The maximum possible score is 100, the average SUS score is 68. The result is determined using a special matrix.

The SUS score for RecogitoJS was 81,84 – this result is to be interpreted as **good** (score over 73).

4.1.5. Evaluation Usability of RecogitoJS in Detail

The usability of RecogitoJS was evaluated with following questions (1 = I totally disagree; 5 = I totally agree):

Questions	result
Manual and assistance - the functions and purpose of the system have been sufficiently explained to me	4,7
Manual and assistance - I know what actions I am expected to take to use the system	4,6
Web application in general - the web application appearance is clear and inviting	4,4
Web application in general - the way how the web application is structured makes sense to me	4,4
Navigation on the website - the navigation offers me good guidance	4,2
Navigation on the website - I always know where I am on the page	4,3
Functionality - I know if I have successfully carried out an action	4,0
Functionality - if an error occurs, I am clearly shown what exactly happened and what action is expected of me	3,5
Annotation user interface - when symbols are used, I understand what they mean	4,6
Annotation user interface - visual elements (fonts, screen, icons, buttons) are big enough for me	4,4
Annotation user interface - I find the arrangement of the page content clear	4,7
Annotation user interface - the feedback from the user interface is understandable to me	4,6
Annotation user interface - I recognize when elements (e.g. buttons) have a certain function	4,6
Annotation user interface - I find the features I'm looking for	4,6
Annotation user interface - I know how to load documents and start annotating them	4,0

³ Jordan, P. W. et al. (1996)

Annotation user interface - it's easy for me to add annotations	4,7
Annotation user interface - it's easy for me to change annotations	4,7
Annotation user interface – it's easy for me to delete annotations	4,6
Annotation user interface - my annotations are clearly shown	4,0
Annotation user interface - I know how to proceed after I finished annotating a document	4,0

Table 5. Evaluation of the usability in detail

4.1.6. Additional Questions about RecogitoJS

During creating the test procedure following additional questions from AIT arose and were implemented into the evaluation (1 = I totally disagree; 5 = I totally agree):

Questions	result
Were too many steps (clicks) required to process the content or carry out the annotations?	2,6
Are Shortcuts required for more efficient data processing in the component?	3,4
The annotation of entities was a too complex process	1,9
The drawing of relations was a too complex process	2,4

Table 6. Additional Questions

4.1.7. Open Space for Additions

At the end of the questionnaire there was open space for additions from the participants. They were requested to structure their answer in a MoSCoW (Must have / Should have / Could have / Won't have) prioritization. This prioritization method developed by Clegg & Barker⁴ is commonly used for catching end user requirements during development processes.

Not all participants gave the feedback structured, due to this fact here in this documentation the category 'additional information' is added – RecogitoJS...

Must have:

- When using a second screen - problem with right edge annotation - the annotation window is not in the visible area – please change
- short cut for further annotations (strg + '+') or the possibility to write more separated with a semicolon
- example 4279-4279 credit card ranges mega list – credit card number in two lines; wrote wrong annotation for entity, want to erase – tool only erased the part in the first line, not in the second – bugfix necessary
- "Drug offer" – long examples – problem control panel disappears – you have to scroll – not efficient
- I found a bit hard to draw relations between entities, as every entity can be somehow related to many others, and the process can last a long time. I think users need to know what is expected from them from a semantic point of view. I.e. a phone number can be of great interest for investigate a particular crime, but I didn't know if the purpose was to get operative data from the advertisement, or create a more general basic intelligence, such as a barrel is an essential component (and therefore it is considered a firearm) or the magazine it is not an essential component.

⁴ Clegg & Barker (1994)

- I had doubts for example when labelling weapons, models and calibers. I did not know if I had already labelled a particular brand, or sometimes the way I linked the relations between them (brandOfThisModel, ModelOfThisBrand,...). From what I understood during the workshop that will not happen after the tool is fully developed, but currently it is very confusing for the user.
- Further explanations are needed to be more efficient. The examples in the guideline were easy and clear but, what criteria must the user follow when creating entities? What is the exact definition of an entity (semantically speaking)? In which way relations are drawn between entities? Are there key words that must be labelled (i.e. smw –JargonSemiautomaticWeapon-)? Hacking is a Modus Operandi, should we label it as such? More complex examples must be provided to improve the user's perspective of the tool, in my opinion.
- About relations, links must be drag-able, so at the end of annotations and relations, before downloading the json file, user can have a late look, which is difficult to do sometimes if many annotations and relations are settled.

Should have:

- Confirmation of the annotation entry by using "Enter" twice
- When drawing relations (if we have a lot) some lines overlapped and there is no way to move them, in order to clearly visible.
- Feature maybe not possible in this version, but an auto-fill option "should" be added using the already existing annotations. This could decrease the number of typos, duplicates because of upper/lower case differences and having to go back to other articles to see how we described an annotation or a relation.
- Be more automatic during the process. Tell the user if the word has been already labelled with an entity beforehand, and give more hints during the process.
- About annotations, it should include the option that once an annotation is settled, you get for the next time the list of annotations already used, so it comes easier and faster to work, when settling repeated annotations for other entities.

Could have:

- Separate the links cancel and ok (left and right with spacing) - sometimes it happens that you click cancel instead of ok
- It would be useful if there was the ability to copy and paste part of the text (mostly during the usability test).
- A recap of annotation/relations "could" be displayed below the article for easier overview as some busy article could become messy very fast.
- Put the information you are labelling and relations you are drawing in some graphic, so that it is easier for the user to understand what you are introducing and perhaps to make some modifications.

Won't have: N/A

Additional information:

- Some minor issues in UI, e.g. I can add stacked entities (like "brand model" and stacked on-top a "product" covering both, but neither can remove/edit them nor connect. Also, adding an Entity requires an "ok" click when I already pressed enter. This "ok" cannot be called via enter key; it is possible however when naming the links.
- By the way, a problem was the missing import button, my browser had to restart, and unfortunately, I only activated the saving per session ... But what I exported in between is still there
- What I noticed, but was not asked in the survey: The rule that entities should be capitalized at the beginning, but may also be capitalized in exceptional cases, is not so 100% consistent. whether you can't just delete it without replacement. Of course, this can also be specified in the end using an ontology, but by then you can throw it out

5. Ethical issues

5.1. Access to the component

To avoid ethical issues during the usability test performed by the LEA's participants, the test was first planned to be carried out in the VTE hosted by KEMEA, with a VPN access for the participants.

But AIT built up a release of the tool that could be used by the participants on their own devices. To distribute this in a secure way the tool was sent with ZED!Pro encrypted to BayHfoeD.

All participants got a secured link to BayHfoeD's own cloud – for the download.

5.2. Evaluation

After the testing timeframe the participants sent back a JSON file, produced of the component, and an evaluation form. Both were anonymized, so that no conclusions can be drawn about the participants.

6. Risks for RecogitoJS Usability Test and Mitigation

For the provision and finalisation of T4.5 RecogitoJS usability test only one actual risk was identified – **“Coronavirus (COVID-19)”**

Risk description: (1) Because of the worldwide pandemic spread all over Europe, there are state orders of “ban on going out” (curfew) in the most European countries. It is not possible to perform face-to-face workshops (to test the usability of the tool). (2) The representatives of the LEAs are very engaged, due to the current pandemic situation. The support of H2020 projects, moves into the second row.

Risk Level: medium

Mitigation actions: AIT in cooperation with BayHfoeD decided to perform the usability test in all steps with none physically attendance (a kick-off virtual workshop, online test in own IT system environments, evaluation by template). The number of participants of the LEAs was limited to 10 – maximum 2 participants per LEA (a number that was possible to handle for the LEAs in the current situation).

7. References

Clegg, D., & Barker, R. (1994). CASE Method Fast-track. Zaltbommel, Niederlande: Van Haren Publishing.
Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (1996). Usability Evaluation In Industry. Abingdon, Vereinigtes Königreich: Taylor & Francis.

Bibliography

- Baker, C. F., C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. Pages 86–90 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1. . Association for Computational Linguistics, Montreal, Quebec, Canada.
- Deng, L., and J. Wiebe. 2015. MPQA 3.0: An Entity/Event-Level Sentiment Corpus. Pages 1323–1328 Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. . Association for Computational Linguistics, Denver, Colorado.
- Gwern Branwen et al. 2015. Dark Net Market archives, 2011-2015. Accessed: 2019-01-23. July 2015. url: <https://www.gwern.net/DNM-archives>.
- Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19:313–330.
- Montiel-Ponsoda, E., D. Vila-Suero, B. Villazon Terrazas, G. Dunsire, E. Rodriguez, and A. Gomez-Perez. 2011. Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web 0.
- Pustejovsky, J., and A. Stubbs. 2012. Natural Language Annotation for Machine Learning. . O'Reilly Media, Incorporated.
- Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. Pages 2383–2392 Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. . Association for Computational Linguistics, Austin, Texas.
- Saldaña, J. 2009. Chapter 1: An introduction to codes and coding. The coding manual for qualitative researchers:3–21.
- Weischedel, R., E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, and N. Xue. 2011. OntoNotes: A Large Training Corpus for Enhanced Processing.