An enhanced pre-frontier intelligence picture to safeguard the European borders

D3.4
Web and Social Media Monitoring Services

Project	NESTOR – 101021851
Work Package	WP3 - NESTOR advanced detection capabilities
Editors	CENTRIC – [redacted]
Lead Beneficiary	CENTRIC
Status	☒ Draft ☒ Consortium reviewed ☒ Peer reviewed ☒ Management Support Team reviewed ☒ Project Coordinator accepted
Version	1.0
Due Date	30/11/2022
Delivery Date	02/12/2022
Dissemination Level	CO

NESTOR is a project co-funded by the European Commission under the Horizon 2020 Programme (H2020-SU-SEC-2018-2019-2020) under Grant Agreement No. 101021851

Deliverable	D3.4 – Web and social media monitoring services
Editors	CENTRIC – [redacted]
Contributor	CERTH - [redacted]
Reviewers	HENSOLDT – [redacted] DECODIO – [redacted]
Ethics Assessment	☒ Passed ☐ Rejected Comments (if any):
Security Assessment	☒ Passed ☐ Rejected Comments (if any):

Abstract	The current deliverable reports on the results of Task 3.4 Web and Social Media Monitoring of NESTOR. The respective deliverable with the distinctive title D3.4 – Web and Social Media Monitoring Services provides the background on the concepts of the web and social media monitoring and event detection, before discussing technical details around the methodology and implementation of the tools, followed by legal and ethical considerations for the modules and conclusions.
Disclaimer	The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Communities. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. © Copyright in this document remains vested with the NESTOR Partners

Version	Date	Partner	Description
0.1	03/10/2022	CENTRIC	Table of Contents
0.2	13/10/2022	CENTRIC	Updated ToC based on CERTH feedback
0.3	28/10/2022	CENTRIC	CENTRIC initial input
0.4	07/11/2022	CERTH	CERTH initial inputs
0.5	11/11/2022	CERTH	CERTH final inputs
0.6	15/11/2022	CENTRIC	Additional CENTRIC input, preparation for internal review
0.7	17/11/2022	HEN, DCD	Peer Review Form was integrated
0.8	23/11/2022	CENTRIC	Ethics Review Form was integrated, Ethics and Security assessment were filled in
0.9	02/12/2022	CENTRIC	Updated Section 7 based on EtAB feedback
1.0	02/12/2022	HP, KEMEA	Final Version – Ready to submit

Executive Summary

This deliverable thoroughly describes the work undertaken as part of Task 3.4 of NESTOR, with the tittle Web and Social Media Monitoring. This particular task aims to develop modules and services capable of crawling the surface, deep, and dark web, and social media sources, with the aim of identifying content related to illegal border activities to enhance the pre-frontier intelligence picture relating to border security. The output of the modules developed in Task 3.4 feeds directly into the Visual Analytics (VA) dashboard developed in Task 4.4, to display the information with the aim of providing situational awareness and assisting in decision support.

The respective D3.4 - Web and Social Media Monitoring Services reports on the near-final status of the modules and tools, including details on the progress, the technical development of the activities taking place in the modules as well as examples of the output results from them. By the time of the submission of the current deliverable, the modules are capable of searching web and social media platform, performing further text-based extraction on acquired data, basic trend and social media detection, while storing the output is a data store ready for display in the VA dashboard. The final step for these modules is the incorporation within the integrated NESTOR platform, which will take place before the use of modules in the final pilot, the Greek-Bulgarian Land & Maritime Pilot, in early 2023.

Table of Contents

List of Figures

List of Tables

Terms and Abbreviations

API	Application Programming Interface
BSON	Binary JSON
CAT	Content Acquisition Tool
CHF	Cryptographic Hash Function
CPU	Central Processing Unit
DPIA	Data Protection Impact Assessment
ECHR	European Court of Human Rights
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HTML	Hyper Text Markup Language
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
NER	Named Entity Recognition
NLP	Natural Language Processing
OSINT	Open-Source Intelligence
PUC	Pilot Use Case
REST	Representational State Transfer
[redacted]	[redacted]
TRL	Technology Readiness Level
UI	User Interface
UK	United Kingdom
UN	United Nations
VA	Visual Analytics
WWW	World Wide Web

1. INTRODUCTION

The D3.4 – Web and Social Media Monitoring Services aims to summarize the progress performed under Task 3.4: Web and Social Media Monitoring. To accomplish the objectives, the task has been broken into a few subtasks that will be described throughout the deliverable. The subtasks are web crawling, social media crawling, event detection, and post-processing. The web crawling module, named the Content Acquisition Tool (CAT), and the post-processing modules are developed by CENTRIC. The social media crawler and event detection modules are developed by CERTH. This deliverable summarizes the technical work carried out under Task 3.4, including the initial status of the tools, the plan for the development of the tools, technical specifications and implementation, and the progress of the tools.

1.1. STRUCTURE OF THE DOCUMENT

The deliverable is structured as follows:

Section 2 discusses the background knowledge relating to the topics covered by the modules, including the surface, deep and dark web. **Section 3** describes the task in relation to the project, including requirements, Key Performance Indicators (KPIs) and Pilot Use Cases (PUCs).

Section 4 and **Section 5** discuss the methodology and implementation of the tools, including the initial and final status of the modules, progress made during the project, and the modules and task's overall architecture.

Section 6 discusses the integration of the task into the wider NESTOR system, from both a technical and a higher-level viewpoint.

Section 7 highlights any Legal and Ethical issues or considerations made as a part of this task, followed by **Section 8** which is the conclusion of the deliverable.

2. BACKGROUND

This section provides information on the background concepts related to the task and modules developed, setting out the aims of using these concepts and technologies within NESTOR.

2.1. SURFACE, DEEP AND DARK WEB

The World Wide Web (WWW) is a network of connected HTML pages that may be accessed online, and it consists of 3 layers, namely the surface web, deep web, and dark web.

The Surface Web contains all the contents that are indexable by search engines. This layer is accessible by everyone, with the websites offering free visits such as news sites, blogs, and open forums.

The Deep Web is an area of the underground, invisible web that a normal user cannot access. It provides information that search engines do not crawl and index. Only those who have the authorization or login credentials can access the Deep Web.

The Dark Web is the deepest layer of the web and a subset of the Deep Web. It is an encrypted section of the web that offers complete anonymity, as it is not indexed by search engines. The communication protocols often use some method of routing to hide the client's location from both web servers and other tracking methods. Dark Web can only be accessed using special software and an anonymizing browser such as Tor.

The intended aim of the surface, deep, and dark web crawling within NESTOR is to identify information relating to suspicious activities occurring near border areas, to enhance the pre-frontier intelligence contained within the NESTOR platform. This information could possibly be found on forums on the surface or dark web, where people might discuss (often anonymously) their plans or intentions. These particular sites are the focus of the deeper extraction performed by the web crawler under this task.

2.2. SOCIAL MEDIA CRAWLING

Social media has become a significant element of our lives. Every year, a growing number of people decide to use social media to interact with others and voice their opinions on topics of their interest. Many of the social media platforms are exploited for the exchange of news and ideas, which may provide important intelligence for a variety of purposes. When a crisis occurs, data obtained from social media platforms such as Twitter, may be crucial for the discovery of almost real-time incidents, discussing the consequences of this event and its influence on society.

Given the enormous amount of data transmitted on social media platforms, the tracking of content of interest and manually processing can be an overwhelming task and would require many hours of human work, thus it could be proven expensive. Therefore, it is vital to develop efficient ways for highlighting and identifying early valuable information regarding events of interest. The tools that are able to monitor, gather and analyze data from open sources are referred to as Open-Source Intelligence (OSINT) tools.

The social media monitoring service, divided into the Social Media Crawler (data acquisition) and Event Detection (data analysis), is an OSINT tool with the purpose of tracking, extracting, and analyzing relevant information based on the domains of interest of the NESTOR project. More particularly, it aims to identify content related to suspicious border activities (i.e., smuggling, illegal crossing). The service will intensify the awareness of the NESTOR platform by leveraging pro-active and real-time searches on Twitter.

3. PROJECT DOMAIN

This section will discuss the project domain, including user and technical requirements, KPIs, and the modules applications to Pilot Use Cases (PUCs), essentially describing the aims of the modules and how they relate to the NESTOR project.

3.1. USER AND TECHNICAL REQUIREMENTS

[redacted]

Table 1: Functional Requirements relating to the Web and Social Media Monitoring Services

Req. ID	Requirement Description	MoSCoW Rating
[redacted]		

Table 2: Functional Requirements mapped to Technical Specifications relating to the Web and Social Media Monitoring Services

Req. ID	Spec ID	Technical Specification
[redacted]		

[redacted]

[redacted]

3.2. KPIS

KPI No	KPI Name	KPI Description	Req. IDs	Measures of Effectiveness (MoEs)	Metric	Trial KPI related to
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 3: KPIs relating to the Web and Social Media Monitoring Services

KPI2.31 is met by the web crawling module, with the capability of crawling the surface, deep and dark web based upon given configuration. [redacted]

KPI2.32 aims at detecting events via the processing of posts coming from the Social Media Crawler. If an alert is given for an event, it is considered as detected for the measurement of this KPI. Currently the accuracy of the model is being tested while its parameters are being tuned.

3.3. APPLICATION TO PILOT USE CASES

The modules developed in Task 3.4 are involved in the Greek-Bulgarian Land & Maritime Trial. The area of interest of the trial is the triple border point of Greece-Bulgaria-Turkey, following the course of Evros river to the Thracian Sea and extending up to Samothrace. The overall intention of the modules in the pilot is to monitor the required border territory, by performing analysis on the web and social media to detect potential illegal activities in the surrounding areas using web and social media data.

4. METHODOLOGY

This section includes details on the modules from the beginning of the timeline of NESTOR and the work that will be carried out in the duration of the task, so to achieve the aims set out for the modules within NESTOR. This section describes previous work, existing approaches to the tasks at hand, and the improvements made to the tools for and during NESTOR.

4.1. CONTENT ACQUISITION

At the beginning of NESTOR, CENTRIC provided an initial version of the web crawling module, the Content Acquisition Tool (CAT), capable of crawling the surface, deep, and dark web based on provided configurations. Given that CAT is an existing module at the start of the project, the aim is to improve its stability, performance, and extraction capabilities to increase its Technology Readiness Level (TRL) and provide a more robust content extraction module to the platform.

Common approaches to basic collection include depth-first and breadth-first crawling [1], where an initial web page is crawled and then links on this page crawled to a depth determined by the user. Information collected through these simple approaches can unfortunately be irrelevant to the user or too much data for a user to sort through manually.

Fortunately, content extraction techniques aim to mitigate this problem by extracting content determined as appropriate for a user. Techniques can include modelling a HTML page as a tree and identifying key content [2], statistical modelling of web elements [3], or using Machine Learning to interpret relevant content from a tagged dataset [4].

CAT draws from existing crawling and extraction techniques, by performing depth-based web crawls, then identifying content relevant to a user through targeted extraction on supported platforms and domains to extract meaningful post content from raw HTML pages.

As the intention of the module is to identify suspicious activities occurring near the borders, and not to identify persons that might be performing these activities, a major inclusion to the Content Acquisition Tool will be an initial approach to pseudonymizing any personal data collected during the acquisition process. This will use [redacted] to hash personal data into unique hashes ensuring no personal data relating to persons performing the activities are shared or further processed.

4.2. POST-PROCESSING MODULES

Post-processing at the start of NESTOR was at a low TRL, providing prototype Natural Language Processing (NLP) tools such as Named Entity Recognition (NER), keyword extraction, keyword geocoding to extract “indicators”, and combining these various NLP tools in aggregate to identify trends of indicators.

Indicators extracted by NLP are concepts extracted by performing textual analysis on content collected by the Content Acquisition Tool that may be of interest to the users of NESTOR in pilot use cases, such as detection of illegal border crossing keywords. These indicators when extracted automatically can significantly reduce the need for Law Enforcement Agencies (LEAs) to manually parse online content, prioritizing content based on indicators and using these indicators in downstream alerting and analysis modules, such as trend detection.

At the start of NESTOR, these NLP tools and trend detection are at a low TRL level but functional, being based on open-source models and requiring a large amount of maintenance and resources to run. This means the development goal of T3.4 with post-processing modules is to streamline the NLP modules to make them more robust and easily deployable, more stable over long periods of time, and tweak to users in NESTOR pilot scenarios to make information actionable to users.

Trend detection will require user feedback and modification during and after the pilots, due to "trends" requiring a consistent definition between users and developers and appropriate technical modification so this interpretation of trends is represented in T3.4 outputs.

[redacted] T3.4 analysis will take a similar approach, starting from the open-source prototypes and using user feedback in pilot scenarios to adapt these solutions to the NESTOR domain.

The focus of development on the post-processing modules during this task is ensuring their relevance to the domain of border intelligence and improving the multilingual abilities of the modules to ensure they will be applicable to the domain and PUCs. As they are included only in the final PUC, the Greek-Bulgarian trial, the target languages for multilingual processing should include Greek at a minimum.

4.3. SOCIAL MEDIA CRAWLER

Social media platforms, such as Twitter, are Web sources referred to as Deep Web since the search engines cannot discover and index most of their content. Additionally, the employment of traditional Web crawlers is against the Terms of Services and Conditions of the providers and would require illicit means (i.e., impersonating that the Web crawler is an actual human user) as to scrape the content from the Web pages. Moreover, the servers that host the platforms are enhanced with anti-scraping tools and will block such bots.

In order to overcome all the aforementioned issues, the Social Media Crawler, initially presented in NESTOR, was capable of acquiring data from Twitter leveraging its official Application Programming Interface (APIs), hence in compliance with the Terms of Use and Privacy Policy of the platform. However, the usage of the Twitter APIs comes with some rate and resource limitations enforced by the provider [4].

[1] https://huggingface.co/

[2] https://keras.io/

[3] https://spacy.io/

[4] These limitations are illustrated in details in Section 5.4.3 as well as the progress made to overcome them.

Our work, in the context of T3.4, focuses on: (i) the further improvement of the performance and the stability of the component (i.e., ways to overcome the limitations) , (ii) the successful integration of the component as part of the NESTOR platform based on the architecture as described in D5.1, (iii) the fulfilment of the user requirements (see APPENDIX I: User Requirements) and (iv) finally the adjustment of the pseudonymization and data minimization techniques according to the Ethical and Legal framework of the project.

4.4. EVENT DETECTION

Event Detection tools for social media is a very active area of research in AI and statistics. The fact is that nowadays, social media is an integral part of our daily life. Every year, an increasing number of individuals opt to use social media to interact with others and share their thoughts on current events and circumstances. Hence, the need for tools capable of extracting useful intelligence out of vast amounts of unstructured, noisy content and interaction is particularly useful. Specifically, social media platforms designed for news dissemination and opinions exchange (such as microblogging services) can contain useful information that can be utilized for different purposes.

However, as state of the art AI tools need an immense amount of labeled data in order to be trained properly, training an Event Detection tool with such a data-driven approach for a very specific use case is rendered difficult, due to data availability issues. On the other hand, statistical time series models allow for a data free approach for detecting suspicious and unlikely trends based on distribution changes in data streams.

The Event Detection tool utilizes a model based on the Poisson distribution, which refers to the probability of observing some quantity (posts), when there are many possible individual contributors to this quantity, each with a low probability of contributing. Such is the case in social media platforms like Twitter, where the overall userbase leads to huge quantities of posts regarding a trend, when each individual user has a small probability of tweeting about that trend. Hence, Poisson distribution seems to be a suitable assumption for such a use-case. Mathematically, the Poisson distribution is expressed as follows,

$$P(c_i; \nu) = \nu^{c_i} \cdot e^{-\nu} / c_i!$$

where P is the probability of observing c_i counts of something, where the expected count is ν . The value of ν serves as a background model for the estimation of the probability, but since there is no way of knowing the true value of it, a reasonable assumption is the selection of the previous count c_{i-1}(count of posts measured in the previous time bin) or an average over many previous points (cycle correction).

For the detection of trends, a model for quantifying the unlikeliness of an observed count should be chosen. Based on the trend detection library used[5], the difference between the

[5] https://github.com/twitterdev/Gnip-Trend-Detection

observed count c_i and the backgound model ν (in the basic configuration: c_{i-1}) can be measured in multiples (η) of confidence intervals (CI) with a preset confidence level α :

$$c_i - \nu = \eta \cdot CI(\alpha, \nu)$$

Therefore, by choosing a threshold η_c we can detect when new counts produce bigger eta values.

Overall, the Event Detection tool in the context of T3.4 is designed to analyze incoming streams of social media data based on Poisson distribution changes, measure the unlikeliness of changes, and detect suspicious events related to border activities in order to notify competent authorities in a timely manner, allowing for better decision-making and preparation for such occurrences.

5. IMPLEMENTATION

In this section, the architecture and technologies utilized to implement the modules within Task 3.4, including the data models employed, their structures, and the flow of data within the system will be covered in detail. This section will also include details around the completion of the task and what work was undertaken on the modules.

5.1. DATA STORAGE

[redacted]

5.2. DATA MODEL

[redacted]

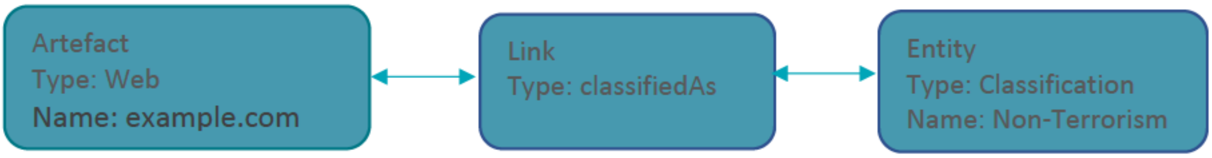


Figure 1: Entity-Link Model

[redacted]

5.3. ARCHITECTURE

[redacted]

Figure 2: The module-level architecture of the tools in Tasks 3.4 and 4.4, both tightly interconnected and led by CENTRIC

Figure 3 below shows a sequence diagram that includes module interactions per-module. Please note that this sequence diagram does not represent the flow of data through the task in order, but shows the individual sequences per-module, and data will not necessarily flow through every module or in the order shown in the diagram.

Figure 3: Interactions of the modules of the Web and Social Media monitoring tool

5.4. TECH STACK

This section provides an overview of the technologies used in the development process of Task 3.4, including languages, frameworks, libraries and more.

5.4.1. Web Crawler

[6] https://www.java.com/en/
[7] https://www.playframework.com/
[8] https://spring.io/projects/spring-data-mongodb

[9] https://www.selenium.dev/documentation/webdriver/

[10] https://www.docker.com/

[11] https://www.json.org/json-en.html

[12] https://www.mongodb.com/

[13] https://github.com/ttezel/twit

[14] https://kafka.js.org/

[15] https://expressjs.com/

For communication purposes and as the NESTOR project's interoperability layer is based on Apache Kafka, a Kafka client is used to interact with the message broker by consuming messages when new data are available for analysis and producing alerts when suspicious events are detected. More specifically, the confluent-kafka library is employed as a python client.

Finally, Docker was also used with this tool for virtual containerization of the software, enabling easier deployment and maintenance.

5.5. PROGRESS

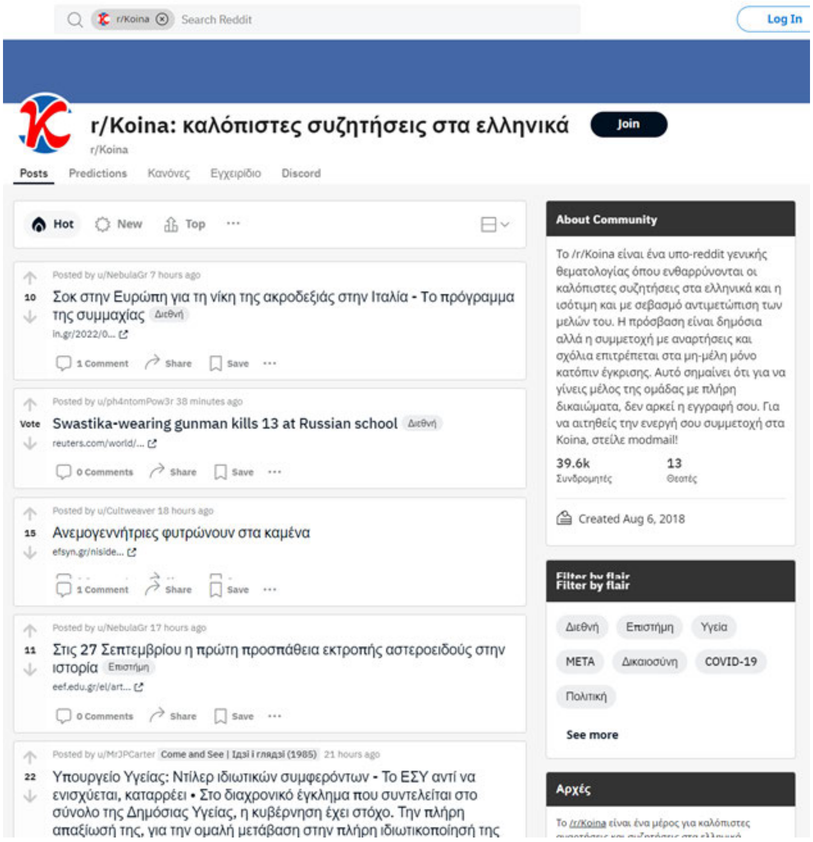
The following section describes the advancements made by CENTRIC and CERTH to each of the tools during the task and displays the status and functionalities within the NESTOR system.

5.5.1. Content Acquisition Tool (CAT)

During the development of this task, several improvements were made to the Content Acquisition Tool to enhance the capability and offer extra functionalities to cover the needs of the NESTOR project. These include enhancement in a few categories such as general improvement for stability and performance, increased access to services and platforms, and additional functionalities and improved extraction from these services and platforms.

The major milestone was the refinement of a refactored browser architecture. This required switching from CAT's headless browsers to multiple instances in Docker containers. This implementation allows all services such as the surface web, deep web, and dark web to do the same by simply calling their browser to access them. If any problems occur, without restarting CAT, each browser can be restarted individually. Another feature added to CAT is the implementation of cookie-based authentication, which allows authenticators to crawl pages that require authentication, bypassing some captcha mechanisms often presented to unauthenticated users. It works by allowing the user to extract cookies from existing login sessions and inject them into a headless browser before crawling, allowing the crawler to act as the logged-in user for the rest of the crawl.

Figure 4 shows an example Greek web page tagged for collection from Reddit. Starting from collection of the web page via accessing the requested URL through a headless browser, CAT collects this web page and begins processing for alignment with the entity-link model described in 5.2.



r/Koina: καλόπιστες συζητήσεις στα ελληνικά

r/Koina

Posted by u/NebulaGr 7 hours ago
10 Σοκ στην Ευρώπη για τη νίκη της ακροδεξιάς στην Ιταλία - Το πρόγραμμα της συμμαχίας Διεθνή
in.gr/2022/0...
1 Comment Share Save

Posted by u/ph4ntomPow3r 38 minutes ago
Vote Swastika-wearing gunman kills 13 at Russian school Διεθνή
reuters.com/world/...
0 Comments Share Save

Posted by u/Cultweaver 18 hours ago
15 Ανεμογεννήτριες φυτρώνουν στα καμένα
efsyn.gr/niside...
1 Comment Share Save

Posted by u/NebulaGr 17 hours ago
11 Στις 27 Σεπτεμβρίου η πρώτη προσπάθεια εκτροπής αστεροειδούς στην ιστορία Επιστήμη
eef.edu.gr/el/art...
0 Comments Share Save

Posted by u/MrJPCarter Come and See | Iдзі і глядзі (1985) 21 hours ago
22 Υπουργείο Υγείας: Ντίλερ ιδιωτικών συμφερόντων - Το ΕΣΥ αντί να ενισχύεται, καταρρέει • Στο διαχρονικό έγκλημα που συντελείται στο σύνολο της Δημόσιας Υγείας, η κυβέρνηση έχει στόχο. Την πλήρη απαξίωσή της, για την ομαλή μετάβαση στην πλήρη ιδιωτικοποίησή της

About Community

Το /r/Koina είναι ένα υπο-reddit γενικής θεματολογίας όπου ενθαρρύνονται οι καλόπιστες συζητήσεις στα ελληνικά και η ισότιμη και με σεβασμό αντιμετώπιση των μελών του. Η πρόσβαση είναι δημόσια αλλά η συμμετοχή με αναρτήσεις και σχόλια επιτρέπεται στα μη-μέλη μόνο κατόπιν έγκρισης. Αυτό σημαίνει ότι για να γίνεις μέλος της ομάδας με πλήρη δικαιώματα, δεν αρκεί η εγγραφή σου. Για να αιτηθείς την ενεργή σου συμμετοχή στα Koina, στείλε modmail!

39.6k Συνδρομητές
13 Θεατές

Created Aug 6, 2018

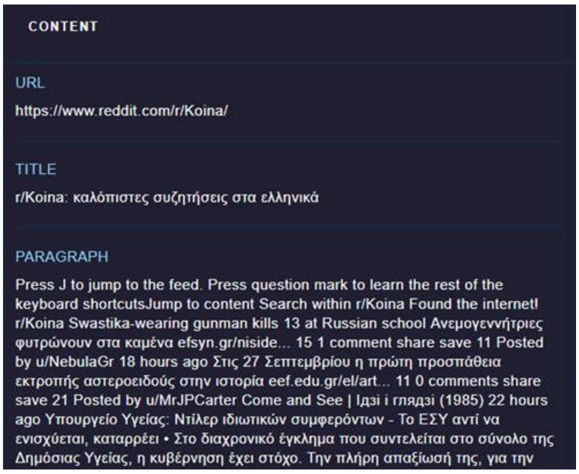
Filter by flair
Filter by flair

Αρχές

Το /r/Koina είναι ένα μέρος για καλόπιστες

Figure 4: Web page collected by the Content Acquisition Tool

Figure 5 shows the Reddit webpage after removing HTML formatting to adapt the useful content of the page into an Artefact. This model is stored with associated metadata such as the URL it was acquired from, the collection and processing times, and the title of the page. The raw textual content of the page after removing HTML formatting is stored in the content of the artefact model.



CONTENT

URL
https://www.reddit.com/r/Koina/

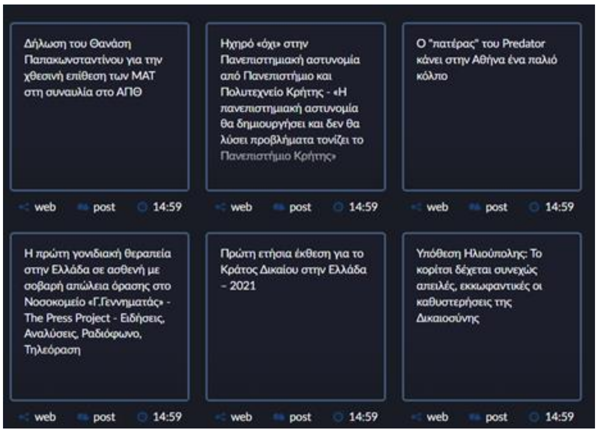
TITLE
r/Koina: καλόπιστες συζητήσεις στα ελληνικά

PARAGRAPH
Press J to jump to the feed. Press question mark to learn the rest of the keyboard shortcutsJump to content Search within r/Koina Found the internet! r/Koina Swastika-wearing gunman kills 13 at Russian school Ανεμογεννήτριες φυτρώνουν στα καμένα efsyn.gr/niside... 15 1 comment share save 11 Posted by u/NebulaGr 18 hours ago Στις 27 Σεπτεμβρίου η πρώτη προσπάθεια εκτροπής αστεροειδούς στην ιστορία eef.edu.gr/el/art... 11 0 comments share save 21 Posted by u/MrJPCarter Come and See | Iдзі і глядзі (1985) 22 hours ago Υπουργείο Υγείας: Ντίλερ ιδιωτικών συμφερόντων - Το ΕΣΥ αντί να ενισχύεται, καταρρέει • Στο διαχρονικό έγκλημα που συντελείται στο σύνολο της Δημόσιας Υγείας, η κυβέρνηση έχει στόχο. Την πλήρη απαξίωσή της, για την

Figure 5: Web page content extracted by the Content Acquisition Tool

Figure 6 shows posts extracted from the main Reddit page, utilizing an automatically generated schema of the Reddit domain to identify where these posts are in the HTML content. Web pages expected in pilot scenarios such as Reddit are added as supported platforms through the automated generation of schemas where possible, with post-processing being applied to web pages if posts cannot be extracted automatically.

Individual posts are also created as artefact data models, making posts from a web page and posts from social media domain agnostic for downstream analysis processing and easier to interpret and filter for users via the “web” type associated with the artefact. Individual posts extracted from whole web pages are linked to the URL they were extracted from, so all the content extracted by CAT is accessible via this extracted URL.



Δήλωση του Θανάση Παπακωνσταντίνου για την χθεσινή επίθεση των ΜΑΤ στη συναυλία στο ΑΠΘ

web post 14:59

Ηχηρό «όχι» στην Πανεπιστημιακή αστυνομία από Πανεπιστήμιο και Πολυτεχνείο Κρήτης - «Η πανεπιστημιακή αστυνομία θα δημιουργήσει και δεν θα λύσει προβλήματα τονίζει το Πανεπιστήμιο Κρήτης»

web post 14:59

Ο "πατέρας" του Predator κάνει στην Αθήνα ένα παλιό κόλπο

web post 14:59

Η πρώτη γονιδιακή θεραπεία στην Ελλάδα σε ασθενή με σοβαρή απώλεια όρασης στο Νοσοκομείο «Γ.Γεννηματάς» - The Press Project - Ειδήσεις, Αναλύσεις, Ραδιόφωνο, Τηλεόραση

web post 14:59

Πρώτη ετήσια έκθεση για το Κράτος Δικαίου στην Ελλάδα - 2021

web post 14:59

Υπόθεση Ηλιούπολης: Το κορίτσι δέχεται συνεχώς απειλές, εκκωφαντικές οι καθυστερήσεις της Δικαιοσύνης

web post 14:59

Figure 6: Posts extracted from web page using the Content Acquisition Tool

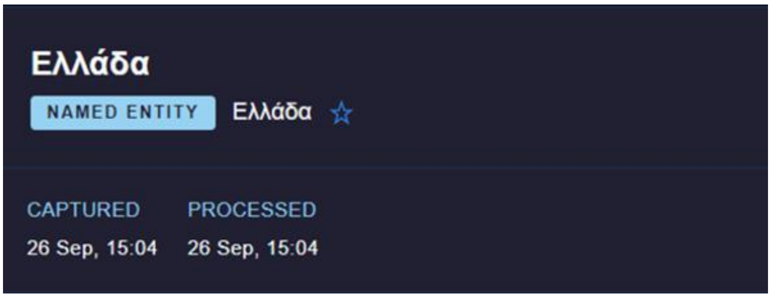
The visualization of artefacts and entities extracted by CAT through the VA dashboard will be covered in more detail in D4.4.

5.5.2. Post-Processing Modules

Post-processing modules have been developed from the baseline NLP tool prototypes discussed in Section 4.2 have been adapted to the NESTOR pilots by adding Greek and Russian language models for NER to support additional pilot languages.

NLP tools have been grouped into an “Entity Extraction” deployable module for ease of use, and the robustness of these individual NLP components improved by testing against example Greek twitter data for edge cases in extraction, and to appropriately format NLP outputs for downstream analysis and trend detection.

Figure 7 shows the individual entity outputs of NER processing as part of entity extraction. All entities store both a captured and processed time in line with the Entity data model described in Section 5.2 to link to future occurrences.



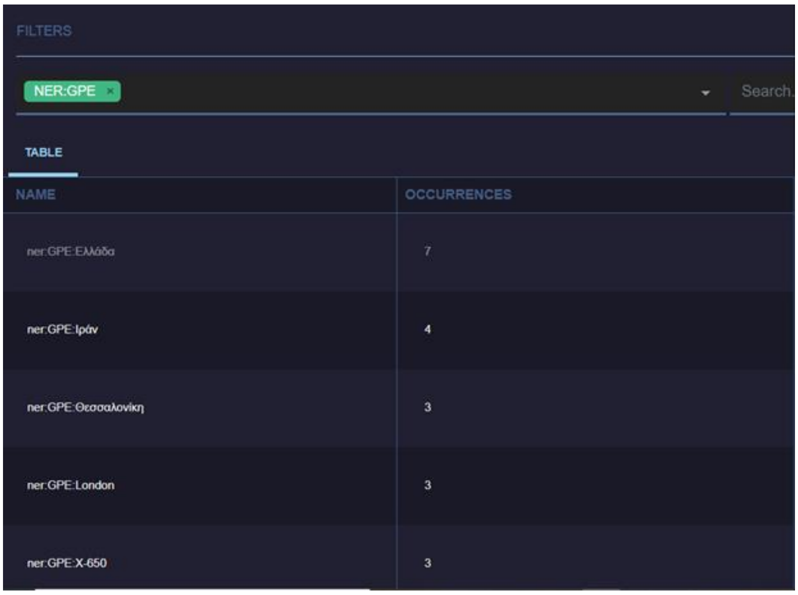
Ελλάδα

NAMED ENTITY Ελλάδα ☆

CAPTURED	PROCESSED
26 Sep, 15:04	26 Sep, 15:04

Figure 7: Individual Entity extracted via Named Entity Recognition

These occurrences increase as the entity is identified in future content as shown in Figure 8, to sort by and identify the most commonly occurring named entities of interest. This list view takes advantage of the Entity data model through the “type” field, allowing for filtering by Named Entity type such as “Geo-Political Entity (GPE)”.



FILTERS

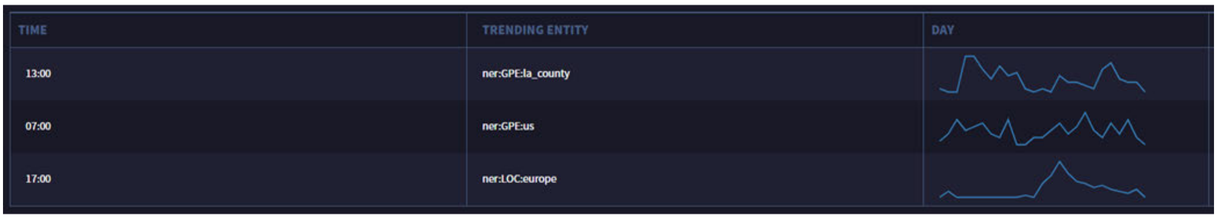
NER:GPE ×

NAME	OCCURRENCES
ner:GPE:Ελλάδα	7
ner:GPE:Ιράν	4
ner:GPE:Θεσσαλονίκη	3
ner:GPE:London	3
ner:GPE:X-650	3

Figure 8: List of Entities extracted via Named Entity Recognition

Much of the trend detection development was oriented towards how these trends are leveraged by a user, focusing on aggregating these NLP outputs and visualizing the outputs as shown in Figure 9. These visualizations show the entities’ name, the timeline of occurrences throughout a day period, and the time at which the entity was shown to be trending.

The details of visualization will be covered further in D4.4, but it is mentioned here as much of the development on trends for NESTOR focused on how extracted indicators from entity extraction are aggregated and trend detection is performed to end with this visualization to users to leverage these indicators in pilot scenarios.



TIME	TRENDING ENTITY	DAY
13:00	ner:GPE:la_county	
07:00	ner:GPE:us	
17:00	ner:LOC:europe	

Figure 9: List of trending Entities

5.5.3. Social Media Crawler

[redacted]

[16] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

[17] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/api-reference/post-statuses-filter

[18] Each request to the Search API returns up to 100 posts ,hence the crawler can potentially gather up to 180,000 posts

[redacted]

Figure 11: Social Media Crawler architecture

Next, the collected content is pseudonymized leveraging the SHA-512 method (see Section **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** for more details), consequently any personal information pertaining to social media accounts is concealed. At the same time, as an extra security measure, based on the principle of data minimization, any redundant data is immediately discarded and only the content necessary for the analysis is kept. Prior to the storage, the acquired intelligence is structured as the data model (described in Section **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**) demands, enabling the successful retrieval and depiction of the collected content into CENTRIC's Visual Analytics Dashboard (Task 4.4). Lastly, the component publishes messages to Kafka, indicating the successful storing of social media data so that the consumer (i.e., Event Detection) can start performing the analysis.

The crawling process can run infinitely until it is manually terminated by the operator. New crawling tasks can start, with different keywords, but all the derived content will be aggregated and presented in a unified manner, into the aforementioned Dashboard. A use case example is presented in Figure 12.

5.5.4. Event Detection

Currently, the Event Detection module can analyze collections of tweets using the point-by-point Poisson model and detect unlikely changes in tweet count (based on some keyword related crawl) distributions. The algorithm is parametric, meaning that certain parameters have to be adjusted in the use case domain in order to accurately detect events.

More specifically, the preset confidence level α and the eta threshold for new detections η_c are crucial for the optimal operation of the algorithm. For a given confidence level, increasing η_c increases the precision [5] of the algorithm, as it will only detect events when a new count surpasses the background ν by multiple confidence intervals. In contrast, if the desirable output is to detect all events with a tolerance in false alarms, then by lowering η_c we are effectively increasing the recall [5] capabilities of the system. In a similar fashion, we can tune the bin_size parameter of the model, which refers to the time period for a single data point (counts of posts). Small bin_size leads to lower precision but leads to rapid detection of events. Finally, if cycle correction is utilized for the background model (using more data points rather than just the latest one for ν), the time frame across which the algorithm will examine the previous data points must be optimized.

In Figure 13 the variation of eta values is shown for a simple example search for #scotus. The example gathers counts of that hashtag in hourly bins. For $\alpha = 0.99$ the red spikes sugggest larger values of η_c that are potentially linked to some event in that topic.

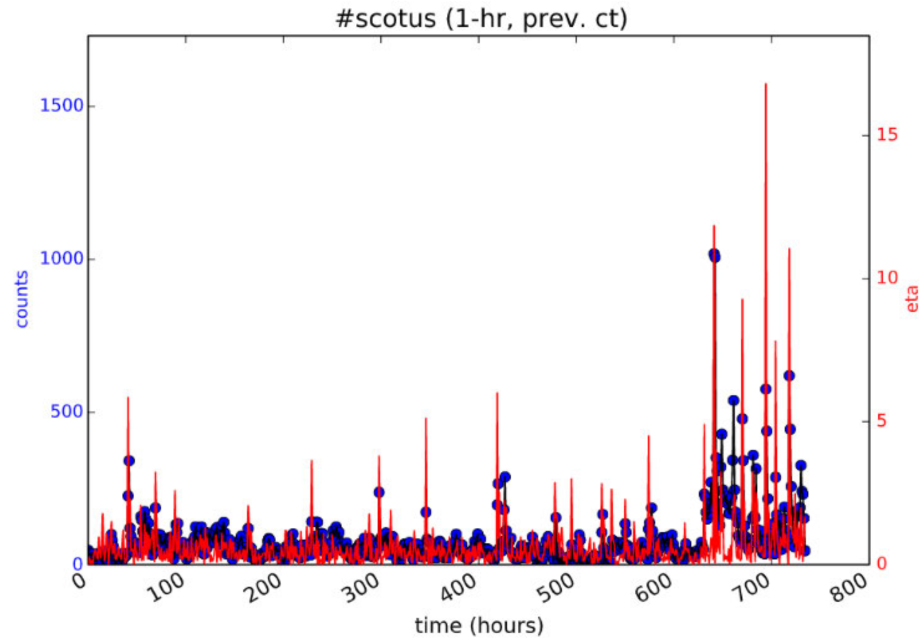


Figure 13: Point by Point Poisson example

The pipeline of the Event Detection tool works in parallel with the Social Media Crawler and is illustrated in Figure 14. When there is a search query for crawling, the Event Detection module gets activated alongside the Social Media Crawler to analyze the acquired data. Then, is consumes the related Kafka topics from the crawler, with pointers to the DB for the collected data. When there is a batch of tweets ready for analysis, they are retrieved from the DB and the Event Detection pipeline is performed on those tweets. First, the data are converted to a time bin with counts format. Afterwards the η_c values are calculated, and the detection of events is performed by comparing the calculated values with the predefined threshold η_c. If an event is detected, the module produces an alert on the Kafka bus with all the relevant posts, to be consumed by other modules.

All in all, the Event detection tool can operate in different time granularities (e.g., minutes, hours, days) regarding the tweet count depending on the needs of its use and the characteristics of the specified search. It can be adjusted for the trade-off of precision and recall depending on the circumstances by tuning its parameters and it can also use various granularities in the background model, used for cycle correction (see 4.4), hence changing the amount of lookback that the model is taking into account for the next estimation.

6. DEPLOYMENT AND INTEGRATION

This section talks about the integration of Task 3.4 within the wider NESTOR system explaining the overall flow of incoming and outgoing data within the task.

All tools developed in this task are containerized to facilitate easy deployment and integration, as mentioned in Section 5.4, and communicate with each other via REST APIs and the Kafka message bus included as a part of the Interoperability Layer. This container-based deployment architecture allows for the modules to be easily deployed to a server for the final integrated platform and piloting of the tools, as well as allowing for easy updates and maintenance, with each module being able to be updated and redeployed individually should issues arise.

Figure 15 below depicts the overall architecture of the NESTOR platform as developed during Task 5.1. As a high-level architecture diagram, this does not include every module developed under Task 3.4 individually. The modules developed during this task, as shown with their interactions in Figure 2, all fit under the 'Web Crawler CENTRIC' component towards the top left.

As mentioned previously in this deliverable, the data outputs of the modules developed under this task feed into the data store provided by CENTRIC, which is used to display data within the Visual Analytics dashboard developed under Task 4.4. The VA dashboard facilitates the integration of these modules into the BC3i platform from a user-facing perspective. This integration will be discussed in D4.4.

7. LEGAL AND ETHICAL CONSIDERATIONS

The functionalities of the NESTOR services must coincide with the Legal requirements of the product plus the possibility of their being any ethical implications that could be presented towards the consortium regarding the plausibility of using such a technical response. The NESTOR system: produced by CENTRIC and CERTH, aims to deploy a Web and Social Media monitoring service designed to provide any information regarding threats to security that can be found online in any format; relating to the Surface / Dark / Deep Web plus any threats made on any Social Media platforms. The following section will dissect the requirements of the components this task used across the NESTOR solution by determining the legislative controllers that have an impact on the components that are being used; this relates to the transfer of data and the functionalities of social media / web crawlers.

The social media monitoring service, is built upon privacy and GDPR aware architecture, having as a lawful basis Article 6(1) (f) GDPR that permits the processing of personal data when necessary (i.e., fulfill the objectives of a scientific research project). Any collected content is publicly available while the framework complies with the relevant Terms of Use, Privacy Policy, and licenses of the data providers (i.e., Twitter).

Additionally, the service respects the fundamental rights of freedom of the data subject. From an ethical perspective, any personal data pertaining to sensitive information such as political beliefs, religion, racial origins, and vulnerable groups is not collected (at least to our knowledge) as such data is not available by the social media providers. Moreover, in accordance with Article 89 (1) GDPR, social media framework employs a pseudonymization mechanism capable of concealing any personal information derived from the social media profiles (e.g., ids, usernames). [redacted]

Finally, yet importantly, with respect to the data minimization principles, only the part of the collected content that is necessary for the analysis is processed and stored within CENTRIC's Database. Consequently, any surplus information is immediately discarded.

The type of data CENTRIC processes includes the surface, deep and dark web data, and other non-social media online data. The web crawling monitoring service is designed with privacy and GDPR compliance in mind, and it has Article 6(1)(e) of the GDPR as its legal justification for processing a task in the public interest. If name identifiers are collected while data collection contains personal information, attempted pseudonymization will be done using an appropriate hashing technique. To ensure compliance with the GDPR's data minimization principles, these names will be pseudonymized during collection by using a hash function or an encryption method. The web crawler component will not process criminal offence data to

[19] The only way to find the original data that produces the hash value is by brute-force search of all potential inputs and conclude to the matching one. This procedure requires enormous computational power.

protect the rights of the data subject as prescribed by Regulation 2016/679. This functionality will not any point infringe any special categories of data such as racial or ethnic origin; political opinions; religious or philosophical beliefs; data concerning health; a person's sex life; and sexual orientation. In accordance with the principles of data minimization, only the portion of the content that has been collected that is necessary for the analysis is processed and retained within CENTRIC's Database. All data is protected and cannot be accessed by an outside source without prior internal authorization and authentication. Moreover, the data we process complies with Article 5 (1)(e) of the GDPR (storage limitation), as the data is periodically reviewed and erased or anonymized when the data is no longer needed.

Web Crawlers are powerful tools designed to appropriate large amounts of data which can be utilized to provide patterns and knowledge on potential threats to the security of the European border security. Web Crawlers are a legal capability[20] of developers and security agencies to deploy, however, developers, LEAs and NGOs should be aware of copyright implications of scraping an individual's copyrighted data. Any data that is copyrighted should not be stored as under Copyright Law of the EU and UK can lead to financial loss for either party – this being through damages or financial loss due to a copyright infringement. The CENTRIC SDS is constantly reviewed for data relevancy and any copyrighted information that cannot be appropriately stored. In a research setting, however, exemptions are present but limited to copies for computational analysis.[21] Research institutes and researchers are still required to buy the appropriate subscription or license to access the material. In the application of the NESTOR Solution – exemptions to copyright can be found within Legal Proceedings, however, the appropriate applicating end-user should consult with their legal department before deploying such capabilities within their third country or member state.

7.1. GENERAL DATA PROTECTION REGULATION

The NESTOR solution is a European based response to any threats to the security of the border, therefore, for the instance of Data Protection the schema that is GDPR; Regulation 2016/679, is utilized to protect the data subjects and collated information from the appropriate tools. This is to ensure that the infringements of rights do not occur and that the appropriate methods that are taken to mitigate a risk to data breach are implemented on a project wide basis. General Data Protection Regulation (GDPR) is a regulatory EU Law which requires individuals to be accountable for their processing of their data while following the guidelines of the regulation. Due to the components in question being developed by a UK partner (CENTRIC), the guidelines that function from the Data Protection Act 2018; Part 2, contain a mirrored UK version of the GDPR which comes from the enshrined European Law. The Data Protection Act 2018 was a product of Brexit to ensure more accessibility for the UK to continue to function with European partners.

The following section will detail the principles of protecting data that are paramount to the functionalities of the NESTOR solution; specifically, towards the functions of D3.4.

[20] hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985 (American Case)

[21] https://www.gov.uk/guidance/exceptions-to-copyright

7.1.1. Principles of Processing Data

[redacted]

7.1.2. Storage of Data

[redacted]

7.1.3. Data Protection Impact Assessment (DPIA)

Due to the impact which the services would have against an individual’s fundamental rights and their rights as a Data Subject. A DPIA was conducted to find any areas of possible infringements or weakness – in fact, two DPIAs were conducted separately from the creators of the modules (CENTRIC and CERTH). A DPIA is a fundamental tool to highlighting any areas of discrepancies within the protection of data and any issues regarding the compatibility of data protection principles functioning within each component of the Web and Social Media Monitoring Service.

7.2. ETHICAL CONSIDERATIONS

The NESTOR consortium endorses and implements the Charter of Fundamental Rights of the European Union across its functioning services. The modules that are found within this deliverable are required to adhere to the following rights: *dignity, freedoms, equality, solidarity and citizens' rights and justice.* The data that is collated within this module is expected to be handled with respect to the Data Subject when regarding their data. The European Union is controlled regarding Human Rights through the UN's Deceleration on Human Rights and the ECHR – European Court of Human Rights which enforce the Charter of Fundamental Rights of the European Union. The United Kingdom (CENTRIC) are required to adhere to the rulings found in the Human Rights Act 1998 and the Universal Declaration of Human Rights (UN). Data Processors and Controllers should be aware of an Ethical approach to Web Crawling and scraping data – these ethical practices should include respecting *robots.txt* when implementing a crawler, understanding the legal groundings for web crawlers which have been defined within their terms of Services, handling data with respect and integrity.

The CENTRIC developers and CERTH developers understand the ethical implications and their purpose of producing a tool that does not infringe upon any data subject. A transparent and responsible approach to protecting the fundamental rights and freedoms of data subjects has been considered to create a fair but functioning system for the NESTOR solution.

8. CONCLUSIONS

Task 3.4 of the project as well as the information collected from the web and presented within the NESTOR platform, indicate a key role in the pre-frontier aspect on the aims of NESTOR, providing a pre-frontier intelligence picture of the European Border.

The respective D3.4 – Web and Social Media Monitoring Services began by outlining the importance of the development of tools, capable of performing crawling activities and extracting relevant data from the web and social media platforms, as well as analyzing this data to derive usable and timely information. This was followed by the envisioned tools' relation to the project, including relevant requirements, KPIs and applications to the PUCs.

In the methodology section, the intended aims for the web crawling and post-processing modules, and the difficulties of extracting and analyzing data from social media and the final choice of methods (i.e., usage of official APIs, point by point Poisson model) were outlined.

Considerations regarding the legal and ethical aspects relating to the modules, most importantly the personal data aspects and mitigations, and the relevance to the architectural framework of the NESTOR project were included throughout the deliverable. This showcases that the modules were designed and developed with the legal and ethical considerations and ramifications of acquiring data from online sources, including the acquisition and processing of personal data.

Finally, during the implementation phase, the process of overcoming difficulties (enforced by the social media provider) and the complete integrated solutions both for Social Media Crawler and Event Detection analysis were discussed. The final status of the web crawler and post-processing modules was also discussed, including example outputs, with further details on the visual display of these outputs to follow in Deliverable 4.4: Visual analytics and decision support system.

9. REFERENCES

[1] R. Janbandhu, P. Dahiwale and M. M. Raghuwanshi, "Analysis of web crawling algorithms," *International Journal on Recent and Innovation Trends in Computing and Communication,* pp. 488-492, 2014.

[2] A. F. R. Rahman, H. Alam and R. Hartono, "Content extraction from html documents," *1st Int. Workshop on Web Document Analysis,* pp. 1-4, 2001.

[3] T. Weninger, W. H. Hsu and J. Han, "CETR: content extraction via tag ratios.," *Proceedings of the 19th international conference on World wide web,* pp. 971-980, 2010.

[4] S. Wu, J. Liu and J. Fan, "Automatic web content extraction by combination of learning and grouping," *Proceedings of the 24th international conference on World Wide Web,* pp. 1264-1274, 2015.

[5] K. Ting, "Precision and Recall," *Encyclopedia of Machine Learning,* 2011.

[6] J. Smith, "How to create references using the bibliography tool in ms word," *A nice journal,* pp. 12-14, 1990.

Appendix A: Quality Review Report

NESTOR Consortium uses this Quality Review Report process internally in order to assure the required and desired quality assurance for all project's deliverables and consequently the consistency and high standard for documented project results.

The Quality Review Report is used individually by each deliverable's peer reviewers with allocated time for the review to be 7 calendar days. The author of the document has the final responsibility to reply on the comments and suggestions of the peer reviewers and decide what changes are needed to the document and what actions have to be further undertaken.

1.1 Reviewers

Project Coordinator	HP - [redacted]
Management Team Member	CERTH - [redacted]
Internal Peer Reviewers	HENSOLDT – [redacted] DECODIO – [redacted] [redacted]

1.2 Overall Peer Review Result

The Deliverable is:

☐ Fully accepted

☒ Accepted with minor corrections, as suggested by the reviewers

☐ Rejected unless major corrections are applied, as suggested by the reviewers

1.3 Consolidated Comments of Quality Reviewers

General Comments	
Deliverable contents thoroughness	Reviewers' comment: Really good and holistic report Author's reply:
Innovation level	Reviewers' comment: Author's reply:
Correspondence to project and programme objectives	Reviewers' comment: Very good positioning of this task in the overall project In Section 3 Author's reply:
Specific Comments	
Relevance with the objectives of the deliverable	☒ Yes ☐ No ☐ Partially ☐ Not applicable Reviewers' comment: Author's reply:
Completeness of the document according to the its objectives	☒ Yes ☐ No ☐ Partially ☐ Not applicable Reviewers' comment: Author's reply:

Methodological framework soundness	☒ Yes ☐ No ☐ Partially ☐ Not applicable Reviewers' comment: Author's reply:
Quality of the results achieved	☒ Yes ☐ No ☐ Partially ☐ Not applicable Reviewers' comment: Author's reply:
Structure of the deliverable with clear objectives, methodology, implementation, results and conclusions	☒ Yes ☐ No ☐ Partially ☐ Not applicable Reviewers' comment: Abstract and Executive Summary missing Author's reply: Added after internal review
Clarity and quality of presentation, language and format	☒ Yes ☐ No ☐ Partially ☐ Not applicable Reviewers' comment: Minor formatting issues Author's reply:

Detailed Comments (please add rows if needed)

No.	**Reference**	**Remark(s)**
1		Section 5.5.1 is titled the same as 4.1, can one be rephrased? Centric response: Resolved
2		
3		

Appendix B: Deliverable Ethics Review

Ethical and Legal Issues	Yes/No by Partner & EtAB comments (if needed)
General	
This deliverable includes the opinion/input of a DPO, Legal or Ethics Advisor.	**No** **EtAB comments:**
Human Participation in research activities (questionnaires, workshops, pilots or other research activities)	
This deliverable is based on research activities (questionnaires, workshops, pilots or other tasks) that involve human participants.	**No** **EtAB comments:** The individuals involved in the data processing operations of the web and social media monitoring services are not consented human participants due to the nature of these data processing operations. However, they are data subjects according to the GDPR.
This deliverable is based on research activities (either during pilots or during the execution of other tasks) that may involve children or adults unable to give informed consent or vulnerable individuals/groups.	**No** **EtAB comments:**
Informed Consent Forms for the participation of humans in research have been/will be signed.	**N/A** **EtAB comments:**
Measures for the protection of vulnerable individuals/groups have been/will be implemented.	**N/A** **EtAB comments:**
Incidental findings, i.e. findings that are outside the research's scope, may be detected as part of the research activities described in this deliverable (criminal activity or personal data of non-volunteers during trials).	**N/A** **EtAB comments:**
Data Protection	
This deliverable is based on research activities that involve processing of personal data.	**Yes** **EtAB comments:** See in detail Section 7.
This deliverable is based on research activities that involve processing of special categories of personal data according to Article 9 GDPR. Special categories of personal data means personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation).	**No (maybe)** **EtAB comments:** See in detail Section 7 and the relevant DPIAs conducted by CENTRIC and CERTH.
This deliverable is based on research activities that involve further processing of previously collected personal data or publicly available personal data.	**Yes** **EtAB comments:** Web crawling is considered as further processing of previously collected personal data as described in D8.3 POPD-Requirement No.3.
Informed Consent Forms for the personal data processing have been/will be signed and data subjects have been duly informed about their rights.	**No** **EtAB comments:**
The conditions for consent cannot be fulfilled. Another legal basis exists.	**Yes** **EtAB comments:** See in detail Section 7 and the relevant DPIAs conducted by CENTRIC and CERTH. Article 6(1)(f) GDPR is the applicable lawful basis for CERTH's operations, while Article 6(1) (e) GDPR is the applicable lawful basis for CENTRIC's operations.

This deliverable is based on research activities that involve transfer of personal data from/to non-EU/EEA countries (non-EU/EEA partners or advisory board members from non-EU/EEA countries) or processing of personal data during the use of platforms regulated by non-EU/EEA law.	**Yes** **EtAB comments:**
This deliverable implements appropriate technical measures that constitute safeguards (encryption or anonymisation or pseudonymisation).	**Yes** **EtAB comments:** See for instance p.16, 21, 30 and Section 7.
This deliverable implements other security measures for the prevention of unauthorized access to, unauthorized transfer of, loss or erasure of personal data.	**Yes** **EtAB comments:** See Section 7 and the relevant DPIAs.
This deliverable is based on research activities that involve profiling of data subjects. Profiling means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.	**No** **EtAB comments:**
Health and Safety procedures (for the staff and the participants in the pilots or other research activities)	
This deliverable refers to activities that may raise health and safety concerns (e.g. from the use of UAVs or from other risks during the pilots).	**No** **EtAB comments:**
This deliverable integrates the measures and mitigation actions presented in D8.5 EPQ-Requirement No.5.	**No** **EtAB comments:**
Dual use	
This deliverable refers to research activities that involve dual-use items in the sense of Regulation (EC) 428/2009, or other items for which an authorization is required.	**No** **EtAB comments:**
Potential misuse of the research findings	
This deliverable includes methodology, knowledge or references to tools and technologies that could be misused if they ended up to the wrong hands or could lead to discrimination and stigmatization of humans.	**Yes** **EtAB comments:**
This deliverable integrates the mitigation actions presented in D8.7 M-Requirement No.7.	**Yes** **EtAB comments: The deliverable is disseminated only amongst the Consortium partners and the European Commission. Information provided in this deliverable about specific NESTOR technologies and their development included in this deliverable will be filtered prior to any publication or communication to the public.**