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

D8.5 – STANDARDIZATION OPPORTUNITIES AND ACTION PLAN

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-2101-WP08-005-V1.0-DV-PU
Revision:	V1.0
Date:	29/01/2021

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Revision history

Revision	Edition date	Author	Modified Sections / Pages	Comments
0.1	03/12/2020	TNL	All	Initial structure, clues for content and responsibilities for redaction
0.11	18/12/2020	GN	5	Adding content of section for Web archiving
0.12	18/12/2020	UGR - LTA	7	Adding content of section for knowledge representation
0.13	18/12/2020	UGR – LTA	8	Adding content of section for domain semantics
0.14	18/12/2020	LTA-UGR-TSIX	9	Adding content of section for graph data
0.15	18/12/2020	LIF	13	Adding content of section for ethical aspects
0.16	03/01/2021	IBM-TNL	10	Adding content of section for spatial temporal data
0.17	15/01/2021	TNL – IBM -UGR	12	Adding content of section for AI models
0.18	15/01/2021	TNL	1	Adding content of section for HMI
0.2	18/21/2021	TNL	All	Edition and generic sections. Ready for internal review
0.3	27/01/2021	TNL	All	Incorporation of changes and review comments from contributors IBM, LIF, LTA, TNL, UGR.
0.4	27/01/2021	TNL	3	Incorporation of additional LEA inputs
0.5	29/01/2021	TNL	All	Rework after reviews by HP/KEMEA, BFP, VICESSE and ISDEFE.
1.0	29/01/2021	TNL	All	Finalisation, ready to submit.

Table of Contents

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

List of Figures

List of Tables

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 WP8 "Dissemination, exploitation and communications" includes tasks related to the uptake and exploitation of the results of the COPKIT project. In particular Task T8.6 "Initiate standardisation, explore certification potential" considers standardisation and certification as an activity carried out throughout the project in which awareness is maintained for the existing standards, de-facto standards, best practices and widely spread formats applicable for the results (in particular of technical nature) of the project. Task T8.6 results in Deliverable D8.5 "Standardization Opportunities and action plan" which summarises the investigations carried out during the task with respect to applicable standardisations and certification and build upon it to define possible actions in the fields for which it appears relevant.

1.2. Purpose and Scope

The present document presents the result of the analysis carried out during the project while developing the results (in particular of technical nature) with respect to existence and applicability of standards to (parts of) the products. Out of the numerous products or sub-parts thereof, certain aspects were identified for which the application of standards, certification process or the application of other practices recognised by the community seemed to have the highest benefits. This document provides details of the standardisation and best practices for these aspects and their potential. Based on this analysis an action plan is proposed.

COPKIT Objective 2 is to '*Develop a toolkit for knowledge production and exploitation in investigative and strategic analysis work […] from knowledge discovery via situation assessment to forecasting*". As a consequence, the results constituting the core of COPKIT innovation touch to a wide range of technologies, and often rely on even more technologies to support the functionalities (think of Relational database, virtualisation approaches and other IT supporting components). The COPKIT Team decided to focus on the areas that directly impact the produced tools and did carry out a deep investigation regarding supporting technologies. Moreover, LEAs (the expected end-users of the COPKIT results) have a significant number of IT systems in place which modification is likely to involve large costs. Therefore, the COPKIT results are more likely to be adapted than to impose modification to current practices, rendering the investigation of potential modification of standards futile for established IT technologies.

It should be noted that, due to the focus on innovation, the technology used in the COPKIT tools is often relatively recent and the field itself (Artificial Intelligence and Machine Learning) not stabilised. "Leading" approaches, and practices tend to change rapidly, and the community consensus is unstable. In many relevant areas, standardisation work is non-existent. This fast pace of change, combined with the presence of large commercial actors (Google, Amazon etc.) limits the ability of the COPKIT Team to realistically influence standardisation work and (national) standardisation organisations. The current document focuses on the propositions that seemed realistically achievable for the project.

1.3. Document Structure

This document is structured in the following way:

- Section 2 presents background information on the nature of the (technical) areas relevant in the COPKIT project, the objectives of the analysis, its scope and limitations.

- Section 3 presents the methodology followed to gather information and select (technical) aspects for further evaluation.
- Section 4 presents a high-level overview of the (technical) areas identified as promising for a standardisation effort.
- These areas are further analysed in depth in sections 5 to 13. These sections include proposed actions in the corresponding areas.
- Finally, section 14 presents the conclusions drawn and a summary of the proposed actions.

1.4. Applicable and Reference Documents

- Grant Agreement number 786687 - COPKIT - H2020-SEC-2016-2017/H2020-SEC-2016-2017-2.
- D2.5 – "COPKIT toolkit definition and EW/EA eco-system description", CPK-2005-WP02-005-V1.0-DV-EURES
- D3.6 - Prototype HMI for analysts for usage of multi-level intelligence v3, CPK-2007-WP03-006-V1.0-DV-CO
- D3.7 - "Demonstrator of security, privacy and uncertainty management handling for COPKIT eco-system v2" in preparation
- D1.4 - Data and Knowledge Management Plan, CPK-2105-WP1-008-V1.0-DV-CO, in preparation.
- D4.3 - Extraction Components Final Release, CPK-2011-WP4-005-V1.0-DV-CO
- D4.4 - Annotation Tool for the Law Enforcement Domain, CPK-2005-WP04-004-V1-DV-PU
- D4.5 - Darkweb data collection tool, CPK-2011-WP4-002-V1.0-DV-CO
- D5.2 - "COPWIK repository: final version, including Uncertainty and imperfection mechanisms and access tools" in preparation
- D6.5 - "Software toolset addressing knowledge discovery v2" in preparation
- D6.6 - "Software toolset addressing situation assessment and fusion v2" in preparation
- D6.7 - Software toolset addressing context-aware forecasting v2, in preparation

1.5. Glossary

Acronym / abbreviation	Definition
AI	Artificial Intelligence
AIT	Austrian Institute of Technology
BAYHFOD	Hochschule fur den Offentlichen Dienst in Bayern
BFP	Belgian Federal Police
CaaS	Crime as a Service
CASTF	Context Aware Spatial Temporal Forecasting, a COPKIT Tool for spatial – temporal forecasting and analysis (Deliverable D6.6)
CKNER	Copkit Named Entity Recognition a COPKIT tool for information extraction from textual data (Deliverable D4.3)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Acronym / abbreviation	Definition
CKRELEXT	COPKIT Relation Extraction, a COPKIT Tool for information extraction from textual data (Deliverable D4.3)
COPLAB	COPKIT Live Lab, one of the exploitation mechanism envisioned in the COPKIT project.
COPWIK	Knowledge base of the COPKIT eco-system
CoU	Community of Users (an initiative supported by EC DG-HOME to foster the creation of a community of participants in research project on different topics. In this document we refer to the community for FCT projects)
CSV	Coma Separated Value
CTSAE	Contextualised Threat and Situation Assessment Estimator, a COPKIT Tool for Situation assessment
DB	DataBase
DNA	Deoxyribonucleic Acid
DoW	Description of Work section of the Grand agreement (Annex 1, part A) describing the work to be realized in the COPKIT project.
DSR	**D**ata**s**et **R**epository, a COPKIT tool for persistence of text corpus datasets, annotated corpuses and NLP models (Deliverable D4.3)
EA	Early Action
EIS	Europol Information System
ELP (Team)	Ethical, Legal and Privacy (Team), the team of partners specialized in Ethical, Legal and Privacy aspects.
EMPACT	European Multidisciplinary Platform Against Criminal Threats
ESMIR	Ministerio del Interior
ESTF	Explainable Spatial-Temporal Forecaster, a COPKIT Tool for Spatial Temporal Forecasting and analysis (Deliverable D6.7)
EU	**European Union**
EUROPOL	European Police Office
EW	Early Warning
FCT	Fighting Crime and Terrorism
FIS/ARD	**F**requent **I**tem**s**et / Association Rules Discovery, a COPKIT Tool for Knowledge Discovery (Deliverable D6.5)
GDCOC	Glavna Direktsia Borba s Organiziranata Prestupnost
GDPR	General Data Protection Regulation
GIS	Geographical Information System
GN	(French) Gendarmerie Nationale
GPS	Global Positioning System
GUCI	Guardia Civil
HMI	Human Machine Interface
I/O	Input / Output
IBM	International Business Machines Corporation
ICT	Information and Communications Technologies

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Acronym / abbreviation	Definition
IETF	Internet Engineering Task Force
IGPR	Inspectoratul General al Politiei Romane
IT	Information Technologies
JSON	JavaScript Object Notation (a frequently used syntax)
KB	Knowledge Base
KD	Knowledge Discovery
LAU	Local Area Unit (EU acknowledge administrative regional divisions)
LEA	Law Enforcement Agency
LTA	Legind Technology
ML	Machine Learning
NA	Not Applicable
NIST	National Institute of Standards and Technology (US standardisation body)
NLP	Natural Language Processing
NSB	National Standardisation Body
NUTS	Nomenclature of Territorial Units for Statistics (EU defined administrative regional divisions)
OCG	Organised Crime Group
OCT	Organised Crime and Terrorist
ONNX	Open Neural Network Exchange (a format to exchange models, in particular Deep Neural Networks)
OSINT	Open Source Intelligence
PESTLE	Political, Economic, Social, Technological, Legal and Environmental
PFA	Portable Format Analytics (a format to exchange computation graphs)
PMML	Predictive Model Markup Language (a format to exchange models)
PN	Policía Nacional
SIENA	Secure Information Exchange Network Application
spaCy	A framework for Natural Language Processing components
SPL	Iekslietu Ministrijas Valsts Policija State Police of the Ministry of Interior
SQL	Structured Query Language
THB	Traffic of Human Being
TNL	Thales Nederland
UGR	University of Granada
UK	United Kingdom
UNICRI	United Nations Interregional Crime and Justice Research Institute
USA	United States of America
W3C	World Wide Web Consortium (standardisation body)
WP	Work Package

2. Objectives and limitation of the exploration of standardisation in COPKIT

2.1. Objectives

To approach the activities of the task T.8.6 “Initiate standardisation, explore certification potential”, it is useful to examine the relationships between standardisation and research projects in general, and EU and H2020 projects in particular. The EC funded project BRIDGIT (and its follow up BRIDGIT2) which consortium included the CEN-CENELEC management centre and 9 National Standardisation Bodies (NSBs), produced a set of reports including “How to link standardization with EU research projects: Standards to support research and innovation”[1]. The document can be seen as a guide for research projects wishing to develop their activities with respect to standardisations and certifications. Their views are based on an integrated approach developed by the Joint Working Group “Standardisation, Innovation and Research (STAIR)” of CEN and CENELEC. The guide summarises the potential benefits of incorporating activities related to standards in research projects. The objectives mentioned can be summarised as follow:

- Efficiency gains during the execution of the project. The usage of standards can facilitate various phases of the project at the early stage by supporting a common view between partners, during technical development by providing concepts or implemented stable interfaces and in the evaluation phase by supporting the testing and the compatibility with environment of the end-users.
- Facilitating the introduction of products on the market. Standards can facilitate the take-up by targeted end-users or organisations of results of the project activities (whether in a commercial form or not).
- Increase the impact and exploitation (in addition to the already mentioned products) of the knowledge resulting from the project activities by:
  - Facilitating reuse of the results by other researchers
  - Increasing the maturity of the field by contributing to the standardisation process via different actions adapted to the maturity.

The BRIDGIT project also produced an equivalent report[2] presented from the point of view of NSBs, which provides interesting information also for researchers.

The COPKIT project subscribes fully to the objectives defined above with respect to its activities related to Standardisation and Certification. However, the COPKIT project proposes innovation in a large scope of processes, encompassing organisational processes, methodology and training for (intelligence) analysis of criminal activities, and a large scope of technologies for data analytics tools and for IT environment supporting such analysis. Prioritisation and focus were therefore necessary.

In addition, the technologies explored in COPKIT such as Data science, AI and Machine Learning are relatively recent and for most applications, no consensus on approaches exists. Given the maturity, very few standards exist. The analysis therefore extends the scope to de-facto standards and dominating formats.

[1] Accessible on the site of CEN-CENELEC https://www.cencenelec.eu/research/news/publications/Publications/BRIDGIT-researcher-guide.pdf (last accessed on January 13th, 2021)

[2] Accessible on the site of CEN-CENELEC, https://www.cencenelec.eu/research/news/publications/Publications/BRIDGIT-members-guide.pdf (last accessed January 13th, 2021.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

The initial step to enable a link between standardisation and research activities is to screen existing standards[3]. Given the maturity of the technologies covered, the screening focused on:

- Identify areas in our results with potential I/O exchange with existing products.
- Identify and use well spread "approaches" and formats for this I/O.
- When sub-components were required that were not part of the core innovation objectives of the COPKIT project, identify and use (when reasonable) well spread and open-source sub-components.

Section 2.2 presents the limitations in our study and screening related to the scope of the projects. Section 2.3 presents the guidelines for component implementation followed in the COPKIT project to overcome the absence of standards applicable to the technical fields addressed, while maintaining the testability, reusability and exploitation opportunities.

2.2. Limitations

As indicated in its objectives, the COPKIT project proposes "an intelligence-led Early Warning (EW) / Early Action (EA) system for both strategic and operational levels". This means that the scope of interest of the COPKIT project is very large as it encompasses:

- Methodological aspects and processes of analytical work in Fighting Crime both at the case investigation and at the strategic level.
- Data analytics tools for the entire analysis cycle, relying on a large range of technologies including AI and ML.
- IT environment facilitating the secure and responsible usage of the tools.
- Ethical Legal and privacy issues related to data processing in the domain of Law Enforcement.
- Corresponding trainings for LEA analysts.

The span of the fields addressed, particularly in terms of technologies is very large. As indicated in the proposal and the Description of Work, the COPKIT project focuses on certain types of data and approaches. Such a focus does not limit the applicability of the innovative concepts and allows the project to demonstrate them along the entire analysis cycle. The activities relative to Standardisation and Certification opportunities are aligned with the focus of the COPKIT project and the tools produced and should be kept in mind:

- Regarding data collection, the COPKIT project focuses on a tool for collecting data from the dark net. Collection of other types of data is **not** investigated.
- Regarding information extraction and structuring, the COPKIT project focuses on textual data, particularly short texts using informal language. Images, video and audio data are **not** investigated.
- An existing IT environment is assumed to be in place, including its legacy. It is **not the intention** of the proposed concepts and solutions to supersede the existing environment but to co-exist with it, including legacy systems and databases.
- The proposed IT environment address the analyst daily work and the "sharing" is intended as internal. In particular, it is **not the intention** to:
  - Address communication with judicial authorities (and the corresponding formal process).

[3] See for instance "How to link standardization with EU research projects: Standards to support research and innovation" (page 9), published by the BRIDGIT project, Accessible on the site of CEN-CENELEC https://www.cencenelec.eu/research/news/publications/Publications/BRIDGIT-researcher-guide.pdf (last accessed on January 13th, 2021)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

  - Address the judicial requirements for court admissibility of (forensic) evidences (such as the chain-of-custody).
  - Replace international cooperation and information exchange mechanism and tools.

The reader is invited to examine the scope and limitations of the technical results of the COPKIT projects that are included in the technical deliverables (Deliverable D2.5, D3.6, D3.7, D4.3, D4.4, D4.5, D5.2, D6.5, D6.6 and D6.7).

2.3. COPKIT approach for architecture and choice of “generic” sub-components

In section 2.1 we showed that one of the potential benefits of screening and using standards during the implementation phase of the project was to facilitate the testing and evaluation for end-users. Further, using standards is also likely to facilitate take-up.

The COPKIT strategy with respect to technical development is to prepare a set of tools that can be used independently. Integration requirements are taken into account, among others by proposing a platform for integration (the SSAP tool developed in Task T3.2 and reported about in Deliverable D3.7). However, the project efforts to realise integration between the tools is only for demonstration purposes and no attempt is made to integrate with existing LEA systems. The rationale is that LEA existing systems are too heterogeneous and include many legacy systems leading to ad-hoc integration strategies that can be more efficiently executed in the productising phase (TRL7-8). This assumption was validated during the project. This is also visible in the answers provided by LEA representatives in the questionnaire addressed for the purpose of Task T8.6 indicating that many systems and tools already co-exists in LEAs IT environment (see section 3.2 and in particular 3.2.6). Consequently, there is no standard architecture, test environment or production environment that can be used as a reference. The reader is invited to refer to Deliverable D2.5 “COPKIT Toolkit definition and EW/EA eco-system description” for more information regarding the COPKIT proposal for an eco-system.

This strategy has consequences in the implementation of COPKIT tools if the goal is still to attempt to facilitate the testing and ultimately the take-up. In the absence of a widely agreed-upon approach, attempting to maximise flexibility is the best strategy. This led to recommendations for the development of components and the design and implementation of test environment.

The following recommendations for architecture of tools were used in the project:

- Use a service-oriented approach for each tool. Tools should be presented as web services.
- Separate as much as practically possible the core functionalities of the tool (e.g., the innovative functionalities developed for the COPKIT project) from parts that are off-the-shelf (storage, HMI etc.). In this way, sub-components can be changed when in the productising phase (TRL7-9) to suit the specific practices and recommendations of the acquiring LEA.
- Use open-sources software when possible and especially for the components outside of the core functionalities.
- Inputs and outputs formats are kept as simple as possible and human readable to support easy testing and transformation of LEAs own data.

This approach provides flexibility for the productising phase.

The same goal of flexibility led to the design of the Validation, Test and Evaluation environment (VTE, reported about in Deliverable D7.1) based on virtualisation techniques. The virtualisation takes place at two levels:

- A virtual machine acting as a server using open-source OS and software.
- An application-level virtualisation using Docker, a modern, flexible and lightweight virtualisation for micro-service. This enables the encapsulation of each tool in one (or several if applicable) Docker container(s).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

- Usage of the tools by end-users is done using a web browser, such as Firefox or Chromium.

Such an environment is flexible enough to ensure that tools can be installed and configured in many environments. This was demonstrated in the context of Task T7.3 with the installation of the VTE on the infrastructure of the Spanish Guardia Civil (one of the LEAs that are partners in the COPKIT project).

With these actions, the COPKIT project mitigated the effect of heterogeneity and lack of standards or even of dominating practices in the field of IT systems for LEA organisations.

3. Methodology

3.1. Overview

As discussed in section 2.1, a very large number of fields addressed in the COPKIT project are potentially relevant for standardisation and certification opportunities. Identifying the most promising fields was therefore a primary goal of activities carried out in Task T8.6. This identification was carried out in parallel along two lines:

- Identifying the challenges experienced by LEAs regarding the (lack of) standardisation or certification processes. The areas in which LEAs estimate that the lack of a standard approach negatively impacts the realisation of their objectives will receive special attention and the intensity of the impact is used as an element to determine priorities.
- Identifying during the technical development of tools the situation with respect to technical standards or certification for the relevant areas, technologies and data types. The information gathered provides indications of the maturity of the technical and applicative area. The maturity is an important element to identify realistic possibilities of actions that can impact the community.

Figure 1 provides an overview of the process followed during the execution of the task.

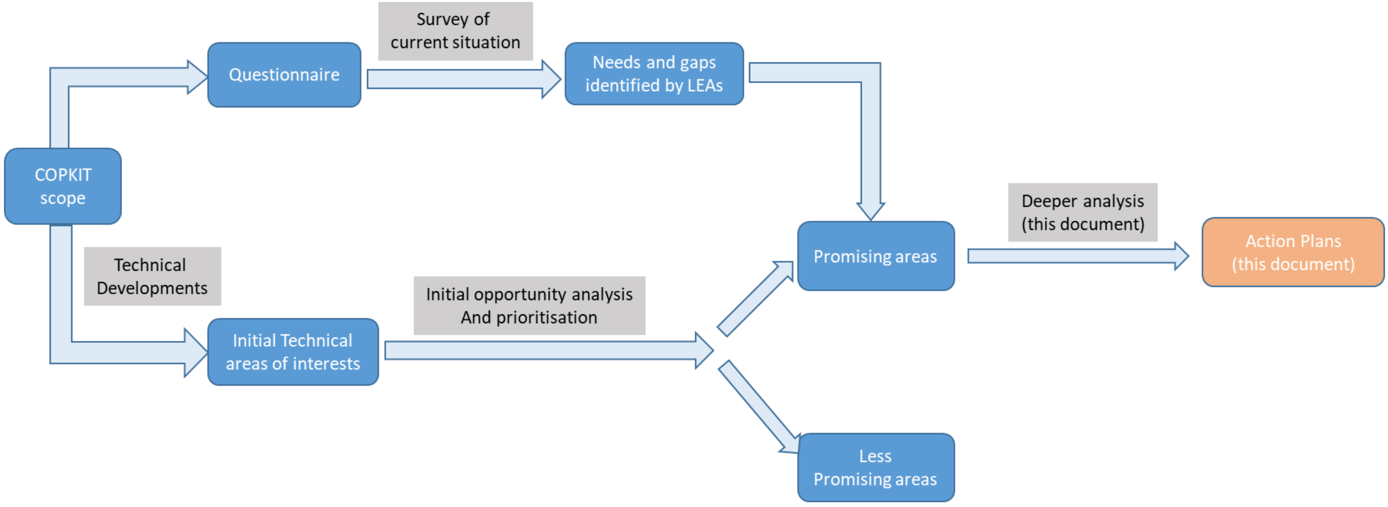

Survey of current situation
Questionnaire
Needs and gaps identified by LEAs
COPKIT scope
Technical Developments
Initial Technical areas of interests
Initial opportunity analysis And prioritisation
Promising areas
Deeper analysis (this document)
Action Plans (this document)
Less Promising areas

Figure 1: Overview of the process followed to analyse opportunities for Standardisation and Certifications and produce action plans.

The approach for these two axes is described in the following sub-sections.

3.2. LEA inputs

An important element to identify opportunities for standardisation and certification is to identify "pains" experienced by LEAs. Areas in which the lack of standardised approach negatively impacts the execution of their tasks should receive special attention. For this purpose, a questionnaire was developed and sent to each LEA partner in the COPKIT project. The questionnaire is provided for information in "Annex I: Questionnaire sent to LEA to collect information on experienced challenges related to Standardisation and certification". The questionnaire contained 6 questions (of which one optional) covering various types of activities (based on the experience in COPKIT but also on information describing the tasks of analysts

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

made available by various LEAs for their training or specialised training institutions[4]). The LEAs representatives were asked to provide information on the usage of standard or the establishment of well-defined practices in their organisation regarding:

- (intelligence) analysis process.
- (Intelligence) analysis techniques, methodology and tools.
- Training of analysts.
- Data format, data exchanges and input / output of analysis tools
- (optionally) IT systems and other digital tools (not dedicated to the analysis tasks).

In addition, one open question was dedicated to the identification of challenges related to the lack of standardisation and prioritisation thereof.

It should be noted that answering such broad questions is a difficult task for LEAs representatives in the COPKIT project for several reasons. First, the LEA representatives in the COPKIT project are practitioners, e.g., case / operational analysts or strategic analysts. A number of aspects of standardisation are beyond their expertise. While they reached out to other members and roles in their organisation, they were not always able to provide full answers to the questionnaire. Second, LEA organisations are very large organisation and “analysis work” spans across the organisation. For instance, while processes may exist, their implementation may differ in different parts of the organisation. A statistically well-founded view of the processes in places and standardisation challenges would require a full-blown audit which is beyond the possibilities and scope of Task T8.6.

The answers provided were therefore analysed and used knowing that they reflected the views of the respondents, in their area of expertise and in the parts of the organisation (department) that they have experience with. In this sense, they should be seen as elements of a field study that can provide qualitative insight on the situation but not quantitative information.

The views do not necessarily reflects the views of the organisations as a whole.

The following sections provide a synthesis of the answers received.

3.2.1. LEA inputs regarding the intelligence analysis process

Respondents indicated that LEAs in their large majority have defined a process for intelligence and analysis work, mostly internally developed and not made public. Strategic analysis processes seem to be less developed. It also appears that while the process exists and is part of the training to come in function, analysis department, especially strategic, may have a significant degree of freedom in the implementation. Some respondents expressed that this freedom is valuable as the analysis process is then more driven by the particular circumstances of the task or case.

Overall, when a process exists it is seen as a very valuable asset. Its absence or inconsistent implementation leads to hindrance in executing tasks due to quality differences and lack of recognisability. Some respondents see room for improvement in the supporting IT system, namely for information exchange. This view is confirmed by the moderate to low degree of satisfaction in communication tools expressed in the answers to the questions regarding generic tools (section 3.2.6).

[4] For instance the UK based College For Policing, and their syllabus for intelligence analysis, https://www.app.college.police.uk/app-content/intelligence-management/analysis/getting-started/ (accessed last on January 11th , 2021)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

3.2.2. LEA inputs regarding (intelligence) analysis techniques, methodologies and tools

Most respondents indicate that their organisations have developed internally a body of techniques, methodologies and tools for analysis work. This body is used as a set of best practices, not as mandatory steps with the possible exception of the construction of certain far-reaching products (national SOCTAs) for which the methodology is more stringent. Overall, the respondents are satisfied with the content. The availability of a central point of information for documentation and evolutions of the practices is mentioned as a significant benefit, to stay up to date. Manuals are mentioned as a point for improvement. Even when such organisation-wide definition is not available (or possibly not spread across all departments), respondents mention that they tend to develop their own body of knowledge at the scale of the department, especially with respect to tools.

3.2.3. LEA inputs regarding standards and practice for training of analysts

Most respondents indicate that their organisations have developed an internal curriculum for the training of analysts. In some cases, this curriculum has been developed by their organisation-wide training department and may be available as a course at their police academy as well. The steps for initial and intermediate trainings are more often well-defined, advanced or specialised trainings tend to be programmed in an ad-hoc manner. Respondents are satisfied with the initial and intermediate training programs. However, several respondents mention that programs for experienced officers and specialised courses (especially regarding the recent developments) are missing or are provided by external partners, sometimes internationally, which is considered as a possible improvement point.

3.2.4. LEA inputs regarding standards and practices for data format, exchange and input / outputs of analysis tools

The received answers show large variations. In general, respondents were able to provide answers for only a few of the categories of data proposed, frequently using the “Non-Applicable” or “Information Not Available”. It is not clear if this is related to actual lack of standards or to the fact that LEA contact points were limited by their own functions (e.g., using this type of data is not part of their activities) and their possibilities to reach-out to other members of their organisation. LEA contact points who had access to technical colleagues or data scientists in their organisation provided more extensive answers, hinting at the second option. GISs constitute a specific case as most organisations have developed a mature internal tool platform due to the sensitivity of geographical information and criticality of having correct information. Some respondents also mention the use of commercial (ArcGIS) or freeware tools (QGIS).

The answers can be divided in two categories:

- Ad-hoc formats are used adapted to a particular situation and need by LEAs. This is the case when an analysis task requires several sources and processing steps. The team will then develop some glue code to transform the data between different tools needing different format to leverage the specific functionalities of the tool. The need for a processing function drives the formats and transformations.
- One format (or in some cases a type of format such as JSON) is “often” or “generally” used (but not enforced) due to the dominance of a popular (third party) platform or a tool in the organisation. The third-party platform or tool can be open-source (storage such as ElasticSearch or Neo4J and Deep Neural Network Framework PyTorch and TensorFlow) or commercial (with IBM i2 Analyst’s Notebook mentioned several times). In the follow-up interviews, it appears that the dominance of a particular tool is not necessarily stable over time. In areas where technology evolves quickly, organisations may change the tools rapidly when new features are introduced.

The satisfaction degree varies significantly. Some respondents indicate moderate satisfaction with this way of working, seemingly tolerating the cost of making ad-hoc transformation as the price to pay for the flexibility and “goal-driven” approach (e.g., functional need). In related comments, the fact that formats and processing goals are strongly related and that unifying may be difficult or counter-productive for some applications was highlighted. Other respondents showed strong dissatisfaction, indicating that, while this was tolerable at the scale of local exchange, it was very serious burden when collaborating with other agencies or at larger (national) scale.

While closed and proprietary formats may be used, some LEA organisations commented during the follow-up interviews on the resulting vendor lock-in and lack of flexibility induced. Different strategies are observed regarding the usage of open-source or proprietary framework and format.

3.2.5. LEA inputs regarding Challenges related to standardisation

The respondents used this question to provide additional insights.

A particularly useful one was that unique standards for data format are difficult to reach as consuming applications may have conflicting needs. The statement was clarified in a follow-up interview with concrete examples, in particular in the domain of graph data: different algorithms require different representations for efficient computation and the corresponding transformations are computationally expensive for large datasets making it inefficient to enforce a particular form. Similarly, formats designed to be practical for visualisation may be unpractical for processing and computation. Such a situation is encountered for more data types in the technical analysis (see section 5 to 11).

Further, the fact that tools should be usable within the legal scope and the need for standardisation in the area of data analytics for predictive policing were mentioned. On the technical side, JSON was mentioned as a frequently used type of data representation.

3.2.6. LEA inputs regarding standards and practices for IT system (optional)

This question was answered by fewer respondents with numerous indications that the information was not available. This was expected as the information is likely to be disseminated across departments and in the IT department in the organisation, reason why the question was labelled as optional.

Nonetheless several aspects are interesting:

- Even inside an organisation, the variety of systems, software and digital tools is significant. LEA organisations do not seem to rely on one vendor or line of products (with the exception of hardware for individual users). This is not surprising given the size and the long IT history requiring the ability to deal with legacy systems, but still worth noting.
- OSs and Databases mentions are in accordance with the expected list of major vendors and open-source systems with a dominance of proprietary systems of the major vendors (Microsoft and Oracle). Most respondents mention the concurring use of many of them. While also not surprising, this is an important point to remember when discussing tools properties with respect to integration and the take-up phase. The (expected) consequence is that, given a certain function to realise, it cannot be predicted in which database system the raw data will be stored and integration will have to be done in an ad-hoc manner.
- In addition to the standard tools, LEAs use additional purpose-made communication systems supporting traceability and integrity for intra- or inter-agencies exchange of (some) critical documents related to the judiciary process.

In general, respondents were neutral regarding the quality or shortcomings of the tools used, although it should be remembered that respondents were mostly analysts. When dissatisfaction was expressed, it was related to the communication tools and, in particular, the logistics difficulties associated with the exchange of large quantities of data (files).

Note that the result could have been different if more members of the IT departments had been among the respondents. Still, the heterogeneity of systems inside organisations is a confirmation of the expected IT situation in LEAs.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

3.2.7. Synthesis of the responses to questionnaires

Overall, we observe that the LEA organisations have a pragmatic approach to the question of standardisation and certification. They have developed internal processes, methodologies and training that suit their needs. Specialised and advanced training for expert officers is identified as a possible gap. This is not surprising as such curricula are also the most difficult to create and formalise.

On the technical side they are confronted with the multiplicity of approaches, formats and third-party tools, often justified by the need to suit specific applications or implement specific functionalities. They are able to cope by developing "best practices". These best practices provide flexibility and support the main goal of the organisation: being able to use the most suitable processing for the case at hand. Another argument for a flexible approach is the ability to introduce third-party tools and platform that are best-of-class at any given moment. Clearly, LEA organisations are aware of the conflicting requirements of different analytical computational processes and expressed the fact that they are wary of a "one size fits all" approach.

In conclusion, the questionnaire did not reveal extremely strong needs for standardisation neither on the process, methodologies and training side nor on the technical and data format side. On the technical side, the questionnaire revealed a potential drawback of standardisation, that of limiting innovation and usage of best-in-class tools. We find particularly important that this potential drawback is kept in mind in further analysis of the needs and benefits of possible standardisation actions.

3.3. Monitoring of technical developments

Next to the enquiries towards LEA partners, the technical areas that could be the source of opportunities for Standardisation and certification were monitored all along the project, starting M7, in parallel with the progress of the technical developments.

As discussed in section 2.2, the potential technical scope is extremely large. A bottom-up approach was therefore followed: each technical partner was asked to identify if and then which relevant standards, state-of-the-art (good) practices and dominant formats (that could constitute a de-facto standard) were relevant **for the tools and functions** that it was developing. Particular attention was given to data formats for the inputs / outputs of their tools to increase opportunities for interoperability. The maturity of the field and of the possible solutions was also taken into account.

The objectives of this monitoring were:

- Raise awareness for existing relevant inputs / outputs formats and other aspects that could be relevant for interoperability.
- Ensure that the project could leverage existing standards and formats for internal use during the project and for future productising phase and take-up by LEAs.
- Identify possible gaps relevant for the objectives of the project, in particular the toolset produced for Objective 2.

The paradigm developed in the COPKIT project proposes to view analytical tools as an eco-system encompassing tools, knowledge (bases) and human expertise. In this view, concrete analytical functions contributing to an analysis task are generally realised by a "chain of tools". In such an eco-system, the flow of information between tools varies for different realised concrete function. The interested reader is invited to refer to Deliverable D2.5 "COPKIT toolkit definition and EW/EA eco-system description". At the beginning of such a chain, one will often find tools in charge of data collection and structuring information (which are also initiating an analytic circle). It is quite common that chains branch-off after the collection step or the step of structuring of information, with the outputs of these tools being used in many other tools. The formats used for the outputs of these tools are therefore more important for interoperability, and extra attention was paid to these outputs.

The monitoring and investigations carried out during the task culminated in the identification of areas relevant for the COPKIT project. This identification was taken into account during the development, as it

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

was an objective of the task. Further, an initial estimation of the opportunities with respect to standardisation and certification was carried out leading to prioritisation for further analysis. The details of the identified areas are provided in section 4 together with rationale regarding areas that were deemed less promising.

4. High level overview of identified relevant areas

4.1. Overview of the status of identified relevant areas

The monitoring of technical development with respect to standardisation and certification opportunities conducted during the project (see section 3.3) enabled the identification of relevant areas with an initial estimate of their potential for impacting actions. This list was complemented (essentially confirmed) by the analysis of the information collected from the LEAs (see section 3.2).

In a second step, the analysis for the potential for impacting actions was refined. Remembering the goals of the task T8.6 (section 2.1) is to:

- ensure that the COPKIT results are in line with current standardisation initiative of future if any can be foreseen,
- if gaps are identified, propose actions. The actions should be realistic in the context of the COPKIT project, e.g., be within the reach of COPKIT resources and influence.

When estimating the potential of an area, the impact of the actions that are within the reach of the project was taken into account. The following aspects were important to determine the potential:

- The maturity of the area. The existence (or smooth emergence) of a consensus or the dominance of particular practices or formats are favourable factors. Conversely, a minimum consensus encompassing only generic aspects or frequently (within a few years) changing dominant actors are a sign of lower maturity indicating that the effort to gather the necessary support for an initiative is unlikely to be within the reach of the project.
- The identification of a need, e.g., sufficient signals that the absence of standards or established best practices is detrimental for the execution of LEA's task, more specifically for the analysts.
- The presence of clear and non-conflicting goals for the standard, e.g., which applications are addressed. If conflicting goals are identified, the existence of research offering a path towards a resolution is critical otherwise the lead time to a standard is likely to make actions from the project futile.
- The chance of being able to influence the standardisation process. In particular the size of the domain impacted by a potential standard (or candidate change) and the field of force are important. The larger the impacted domain is and the stronger the presence of large and well established players is, the lower the chance that the project can achieve influence. As an extreme example, let's imagine that COPKIT would want to propose a change to the query language used for most relational databases (SQL). The impact would not be limited to LEAs databases (which would already be quite ambitious) but would impact any domain using relational database: in the world of today a significant portion of human activities involving a computer. Such actions would have little chance to succeed.

The criteria above led to a list of areas for which the potential for impacting actions justifies a deeper discussion on possible actionable recommendations and are addressed in section 5 to 13. Subsequently a list of areas for which the potential for actions seemed reduced has been established. These areas are addressed in sub-section 4.2. Table 1 provides a list of identified areas and the initial estimation of their potential for Standardisation and Certification, together with references to corresponding questions in the LEA questionnaire and refrenece to sections where the area is further discussed.

Area description	Estimation /Rationale	LEA Question ref.	Section
Web archiving formats	Promising	4.1	5
Textual data, metadata and annotations	Promising	4.2	6
Knowledge representation	Promising	4.4	7
Criminal domain semantics	Promising	4.3	8
Exchangeable representation of graphs and relation networks	Promising	4.8	9
Spatial Temporal data	Promising	4.6	10
(AI) Models	Promising	4.9	11
HMI and visualisations for analytics	Low maturity and conflicting requirements (application adaptation paradigm) but deserving an in-depth analysis	NA	12
Ethical aspects of AI usage for Law Enforcement	Low maturity but deserving an in-depth analysis	NA	13
Intelligence Analysis Process	Current situation with internal processes is satisfactory and a period of consensus building around COPKIT innovations is needed	1	4.2
Intelligence Analysis Techniques Methodologies and tools	Current situation with internal processes is satisfactory and a period of consensus building around COPKIT innovations is needed	2	4.2
Training for analysts	Current situation with internal processes is satisfactory and a period of consensus building around COPKIT innovations is needed	3	4.2

Table 1: Summary of identified area and initial estimation of potential for Standardisation and Certification. The column “LEA Question Ref” refers to the questions in the questionnaire addressed to COPKIT LEAs, with 4.X being the specific field addressed in question 4, row X.

4.2. Insights on the potential for standardisation of the Intelligence Analysis Process, techniques and methodologies and tools and training

The responses to the questionnaires sent to COPKIT partner LEAs shows that the situation is similar for the three areas:

- Intelligence Analysis Process
- Intelligence Analysis Techniques Methodologies and Tools
- Training for analyst

The responses provided a key insight to evaluate the potential for standardisation of the intelligence analysis process: namely, the LEAs who actually have a process are satisfied with it. This indicates that the approach consisting of building a process internally (or a corpus of techniques, methodology and tools, or training curricula) is apparently suitable. A possible explanation is that the elements can be appropriately tuned for the local circumstances, including the nature and state of criminal activities in the corresponding country. Consequently, there is no evidence that actions are required.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Furthermore, the COPKIT project develops a new methodological approach: the EW/EA methodology which may, in the long term, result in modification of the process (such as an increased inclusion of contextual information). Since this methodology is new and has not been applied at large scale, it seems premature to aim for a standardisation of the analysis process. A period of time to obtain more feedback from more agencies and to build consensus is likely to be necessary. In the meantime, the project team will continue its dissemination effort.

It should be noted that the COPKIT project includes activities to develop training material corresponding to its activities. This training material aims at serving immediately LEAs who would decide to expand the application of the methodology and the tools developed in COPKIT. But it could also serve as a base, should it appear in the future (after a period of consensus building) that there is a need of unified training. In addition, the COPKIT training material includes the elements of advanced data analytics that are required to use the tools and can partially answer the expressed concerns of LEAs respondent with respect to the availability of advanced level training for analysts and of data science training curricula specifically targeting LEA analysts.

Also note that the COPKIT team is envisioning the creation of the COPLAB, a living lab aiming at facilitating the co-creation of analytic tools for LEA and the corresponding training as part of the exploitation plan. Discussions are on-going with important European actors among others, Europol and Cepol to define the possible missions and organisation of the COPLAB. Contributions to unification of processes and creation of advanced curricula are possible missions of the COPLAB.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

5. Web archiving standards and format

5.1. Introduction

While there are many of methods to extract and to store data from a website scraping, using standard format for the outputs of a scraper presents advantages for the exploitation of the scraped data in a global pipeline composed of distinct tools. Because of the diversity of potential consuming functions, the scraping output format must limit the data loss as much as possible and, due to AI developments, the integrity of the data is imperative. Integrity is also an important characteristic both to support usage by Machine Learning tools and due to the constraints of data exploitation for LEAs.

Different formats will cater for different trade-offs between integrity, traceability, storage, usability and possibly, preferences towards open source of proprietary.

5.2. Overview of explored / existing standards and accepted formats

Web presentations formats and web site development techniques evolve quickly and possible usage of the data is also increasing. The field is highly volatile and there is no consensus on website storage formats. As a result, there are lots of proprietary formats in response to specific needs or industrial preferences.

Some formats promote usability and human readability. On the contrary, some promote efficiency and integrity with a strict limitation of pre/post processing on the scraped data. The choice of the format will be defined by the type of stored data, the desirability of readability by an extern tool human and the required degree of integrity, particularly strong for application by LEAs.

Still several frequently used approaches are worth mentioning:

- JSON format is dominating the content storage for structured data, due to its flexibility to accommodate content in an ad-hoc and self-describing manner. Another advantage is that it is human readable (although not particularly practical due to its verbosity). The drawbacks of the flexibility and self-describing approach are (i) the verbosity resulting in a relatively poor computational effectivity and (ii) the relative permissiveness resulting in many approximations in implementation.
- HAR format: standard format for storing raw HTTP requests. The content is based on JSON data structure.
- WARC format: standard format for web archiving based on the concatenation of some distinct multimedia contents. The content is based on JSON data structure.
- Proprietary formats with various characteristics generally increasing efficiency for specific aspects (often the storage efficiency).

5.3. Action taken in COPKIT

A web archives creator should create a raw copy of a website to be sure to store all information and all specificities of the website over time. If the website is not stored with a strict integrity policy, the data may be biased or incomplete. The integrity is very desirable when the data is used by for ML processing. But keeping in mind the application for LEAs, it is even more critical to define a “numerical evidence” of an observation at a specific time. Moreover, if the data are modified to be stored, there are some ethical and legal issues about the integrity, fairness and trust. An appropriate format must therefore promote integrity as far as possible but it should also limit the needs for third-party extern tools as well (proprietary tools for example).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

The WARC format is a standard for Web archiving but the format doesn't preserve the complete integrity of the data and some protocols are not (or badly) supported such as web-sockets request. So, this format is not suitable for the envisioned COPKIT applications due to the limited guarantees of integrity.

The HAR format preserves the integrity of the data by storing raw HTTP requests without any processing but the web-socket transactions are not completely interpretable in this format. Web-sockets are essentially used for chat tools. It is a minor task of the scraping goal in COPKIT (essentially based on standard websites), so it is not critical. While the format was originally developed by the W3C consortium it was abandoned before achieving standard status. Still, it can be seen as a de-facto standard and the majority of industrial tools (and most popular browsers) supports it. The format is human readable but not practical and the storage size is not optimized. It may be a problem for huge websites.

One important feature of the data scraping is the capacity to "replay locally" the data. This feature is poorly supported by the HAR format due to its difficulty to interpret AJAX request. Even if the HAR format has strong capabilities, it is not sufficient to respect all conditions of scraping storage policy. To deal with that, another format is necessary to respect all conditions.

Based on the experience of the technical developer, the .REPLAY is used in the COPKIT project. The .REPLAY format is the format used by the open source tool named "mitmproxy" (a HTTP proxy). This format has strong capabilities:

- All web-sockets and HTTP transactions are stored perfectly. The data dump is completely raw and can be considered as a numerical proof.
- Supports visualization of the data (replay locally).
- The storage size is optimized.

Some drawbacks are also worth mentioning:

- It is a proprietary format and it is not a standard.
- The format is not human readable.
- Automatic parsing implies third party dependency to proxy decoder

In absence of solution resolving the conflict between the requirements the COPKIT project is maintaining two formats as output for the developed scraping tool:

- The HAR format is used for the standard COPKIT pipeline because it is efficient enough and reasonably well defined format (due to its origin within the W3C).
- the REPLAY format is used to deal with the visualization task and to guarantee the status of numerical evidence of the data (if needed).

5.4. Plans for further actions

Based on the experience in the COPKIT project and by practitioners using the scrapping tool, the couple HAR / REPLAY formats appears to be satisfying functionally and is operational. It provides a solution to all tasks (related to the persistence of scraped data) realized in the COPKIT project. No more action is necessary.

At this stage, there are no initiatives (such as standard working groups) known to the authors aiming at producing a standard satisfying the conflicting requirements. There is also no proposition of providing a path to reach a technical solution. Nonetheless, the issue should be monitored. In the COPKIT project, the partner developing the scraping tool is also a user and practitioner. The organisation is therefore in an ideal position to monitor the topic. They will therefore follow up the monitoring outside and after the COPKIT project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

6. Textual data, metadata and annotation standards and formats

6.1. Introduction

Textual data are an important type of unstructured data for the COPKIT project. Therefore, a number of Natural Language Processing tools are developed in the project, targeting particularly the recognition of entities (e.g., concepts such as names or digital identities, location etc.) including domain specific entities (for instance firearm models) and relationships between these entities. The reader is invited to refer to Deliverable D4.3 (“Extraction Components Final Release”) for further information regarding these tools. In addition, the COPKIT approach implies the interaction of domain experts (LEA analysts) with the relevant textual data and the annotated version of the texts produced by the tools. This interaction needs to be facilitated by specific visualisations of annotated texts. The reader is invited to refer to the Deliverable D4.4 (“Annotation Tool for the Law Enforcement Domain”) for additional details on annotation tools for LEA analysts.

Finally, the information extracted from textual data (e.g., the entities and the relations) are typical inputs for the analytical tools operating the structured data produced by the NLP tools. The establishment of a convenient and flexible format is an important step towards the capability of tools to interact.

Within the functions targeted in the COPKIT project, the texts being analysed are collected on the web, and mostly on the dark web. The COPKIT project includes development for such a tool (scraper) and considerations about the format of the output of a scraper are discussed in section 5. While such outputs are appropriate for the archiving, further processing will aim at extracting specific portion from these archives that are relevant for the analysis. The realisation of the extraction is dependent on:

- The structure of the web page for which the web site developers have almost complete freedom. It is not possible to assume a certain marking for certain information, certainly not for the type of website targeted by the COPKIT project as dark net developers have no particular reason to follow any web development best practices (for instance regarding code readability).
- The actual targeted functions which drive the relevance of information pieces.

For the purpose of this section, we assume that the output of the scraper has been further parsed to extract the relevant information, e.g., the plain text of relevant content elements and export it into a CSV format.

In addition to the Natural Language processing tools, activities in work package WP4 developed the Data Set Repository to manage input datasets (e.g. text corpora), outputs of NLP tools (annotated text corpora) and model resulting from learning on certain corpora and thus tuned for certain applications (in the context of COPKIT, tuning for different crime types in the COPKIT theme: firearm trafficking and Crime as a Service /Data as a Commodity).

6.2. Overview of explored / existing standards and accepted formats

6.2.1. Input data for text annotation and model creation

The input of information extraction processes is the plain text. The format in which the plain text is encapsulated will depend on the publishing component. For the COPKIT application, the plain text part is generally only a piece of the record, (e.g., one cell in a row of a table) and will be accompanied with other elements for that record (for instance, date of publication, author, provenance etc.). In the absence of domain specific standards, the choice of format should be driven by usability: e.g., flexible and human readable. Therefore, JSON and CSV seem the natural candidates or possibly a combination of both. CSV tend to be more readable for human and can be manipulated by non-technical users using Microsoft Excel or OpenOffice Calc for instance.

In addition to the plain text, the record should accommodate the possibility of storing the result of previous processing of entity and relationship extractions. Modern frameworks for entity representation use JSON representations. The previous results become a JSON formatted field in a column.

The JSON representation of the entities facilitates also the visualisation of input data including entities and relationships for the purpose of inspection or even prior annotations by analysts as many libraries support the visualisation of annotated texts (such as vis.js[5]).

6.2.2. Output data for information extraction results

The output of information extraction tool can be used for two purposes:

- Immediate visualisation, requiring a format compatible with the web-based approach to HMI used in COPKIT.
- Reuse by other analytical tools including COPKIT tools developed in work package WP5 and WP6. The Data Set Repository tool developed in COPKIT aims at providing persistent storage and availability via an API of annotated texts and corpuses. A format suitable for long term storage is required. This format will also cater for post analysis (statistical or other). Variants need to be supported (for instance, anonymised or not).

For the goal of visualisation, the format chosen should be a good balance between practical inputs and internal representations of entities and relationships and web-based visualisation. Open-source components exist that support the transformation of JSON entities into semantic HTML (for instance using the visualisation component displaCy) and are able to create a dynamic HTML visualisation (for instance using the vis.js JavaScript library).

For long-term persistence of annotations in outputs, the W3C consortium developed a web annotation format, called the Web Annotation Data Model[6]. This format can be considered a standard and can ensure interoperability with other tools inside and outside of the COPKIT project.

For long term persistence of relationships in outputs, a graph supporting format is needed. JSON-LD is a W3C/IETF[7] standard developed for this purpose able to represent graph data (see discussion in section 9.2).

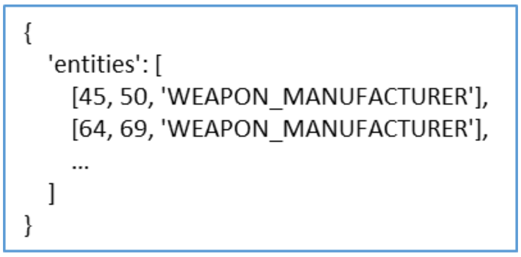
6.3. Action taken in COPKIT

6.3.1. Input data for text annotation and model creation

In order to define unified formats to store or transfer the results of information extraction components (activities carried out in work package WP4), several JSON formats to persist the extracted information have been chosen. For persisting entities in the CSV files it was decided to use the spaCy JSON format for entities and the JSON network graph format of vis.js primarily because of the simplicity and human-readability of the formats. Figure 2 presents the representation for entities used for inputs and visualisation outputs and Figure 3 presents the representation of the relationship graph used for inputs and visualisation.

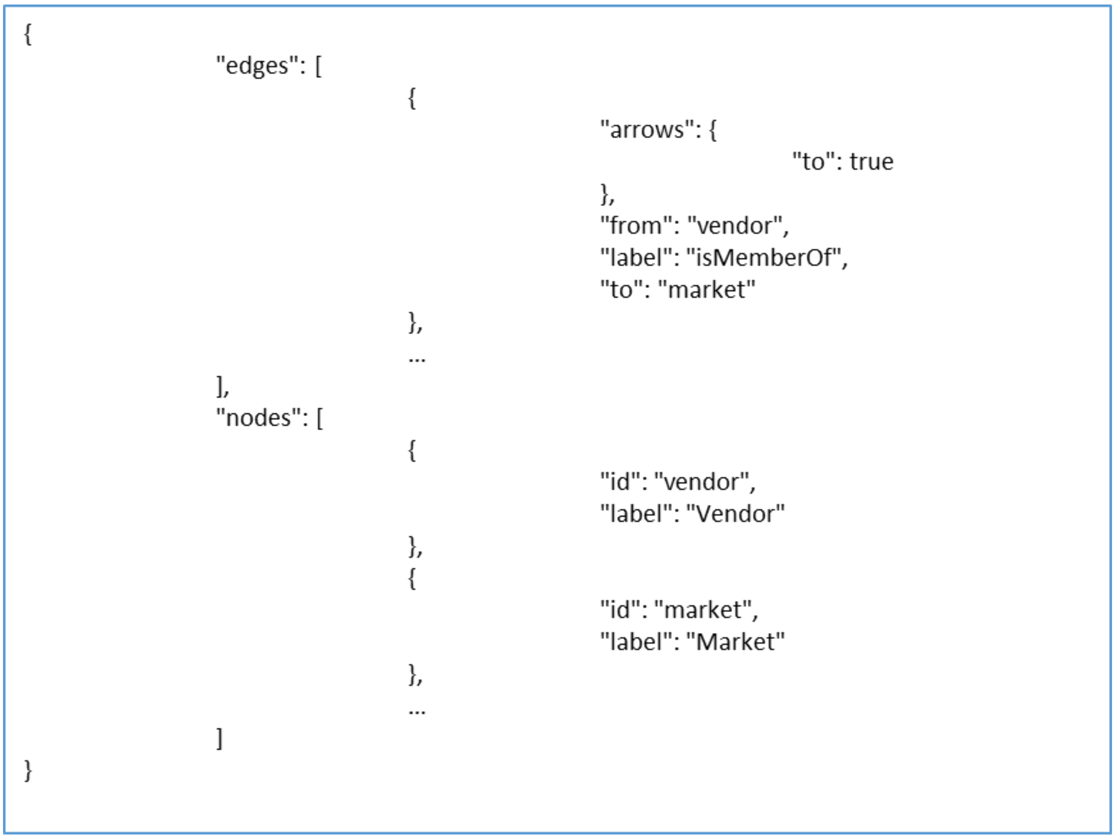
[5] See https://visjs.org/ (last accessed on Jan 15th, 2021)
[6] See https://www.w3.org/TR/annotation-model/ (last accessed on Jan 15th, 2021)
[7] See https://www.w3.org/2018/json-ld-wg/ and the discussion in section 9.2 on standards for graph data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved



```
{
   'entities': [
      [45, 50, 'WEAPON_MANUFACTURER'],
      [64, 69, 'WEAPON_MANUFACTURER'],
      ...
   ]
}
```

Figure 2: Entities representation in JSON for the input and their visualisation if required

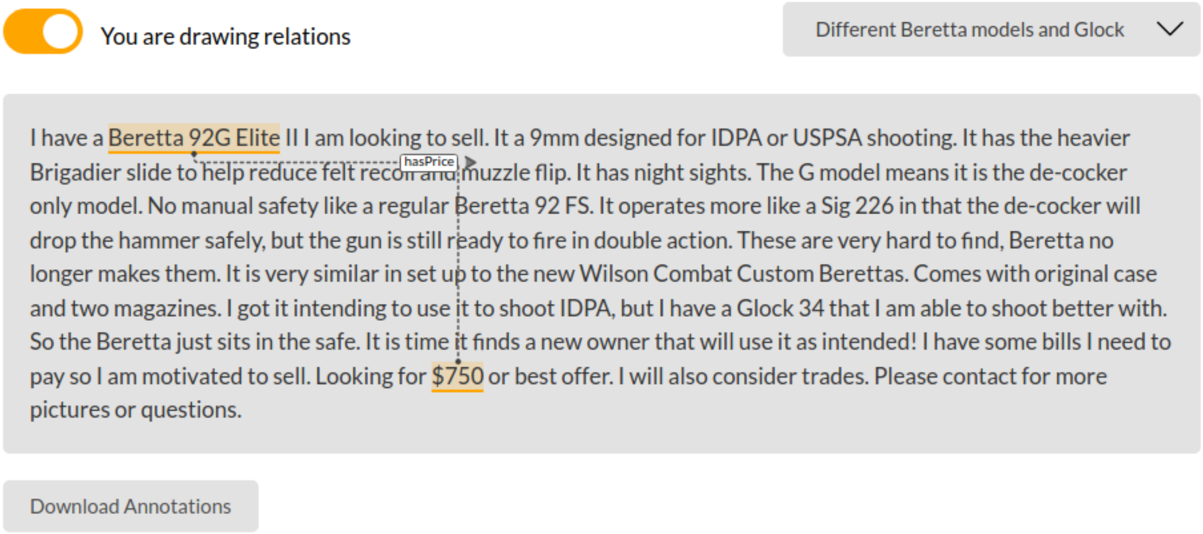


```
{
                "edges": [
                                {
                                                "arrows": {
                                                                "to": true
                                                },
                                                "from": "vendor",
                                                "label": "isMemberOf",
                                                "to": "market"
                                },
                                ...
                ],
                "nodes": [
                                {
                                                "id": "vendor",
                                                "label": "Vendor"
                                },
                                {
                                                "id": "market",
                                                "label": "Market"
                                },
                                ...
                ]
}
```

Figure 3: Representation of the graph of relationships for the inputs of the information extraction and their visualisation when required.

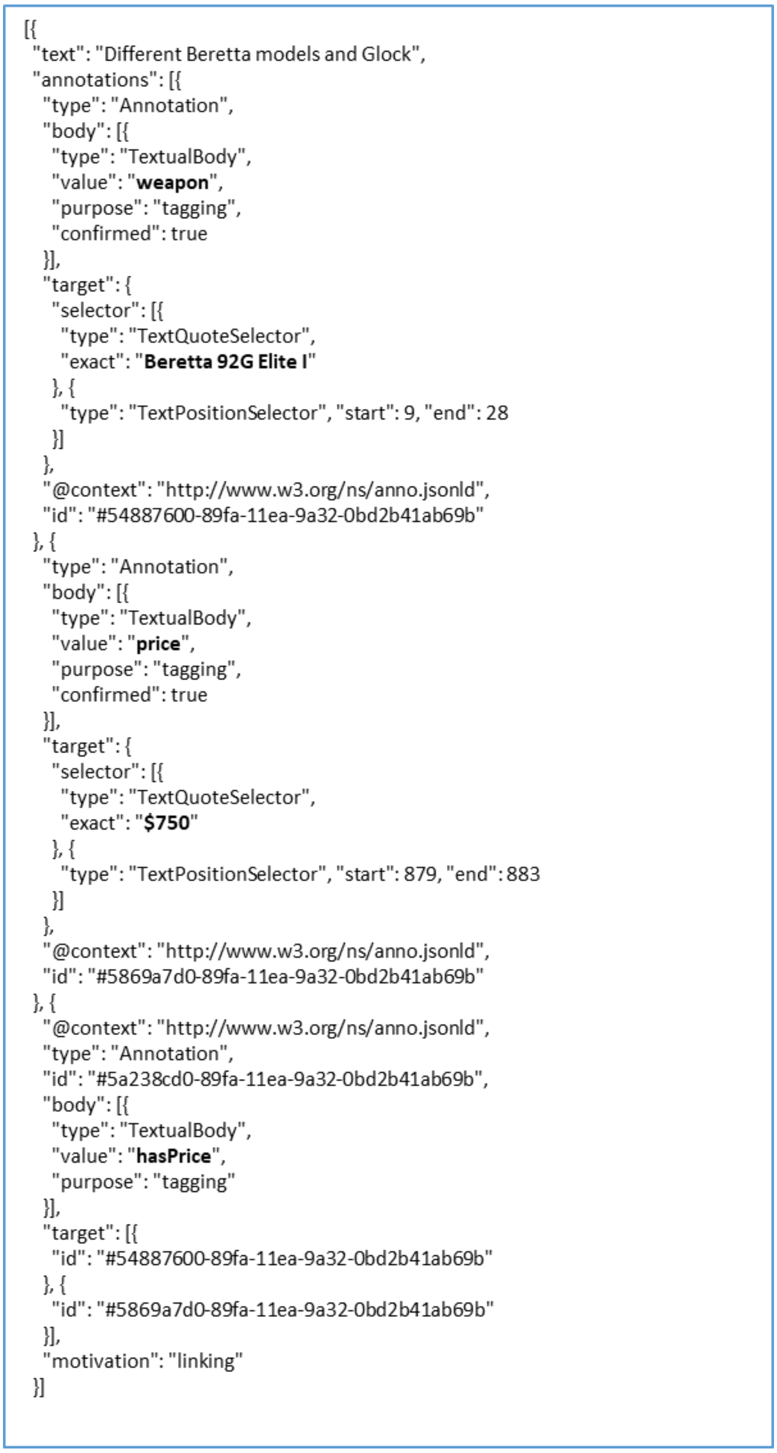
6.3.2. Output data for information extraction results

For the long-term persistence using the DSR, the standard formats “Web Annotation Data Model” and JSON-LD recommended by the W3C are used (see section 6.2.2). The representation of the graph implied some coordination with respect to the semantics of nodes and edges. The ontology used in the DSR was developed in COPKIT as a collaboration of the work-package WP4 and work package WP5 to support the specific domain related semantics. Figure 4 presents a visualisation of an annotated text by the Recogito Tool (see Deliverable D4.4) whereas Figure 5 presents the corresponding W3C formatted output.



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.

Figure 4: Visualisation of the annotation results with the tool RecogitoJS (See Deliverable D4.4)



```
[{
 "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"
 }]
```

Figure 5: The JSON document exported to W3C format corresponding to the annotated text visualised in Figure 4.

6.4. Plans for further actions

At the time of writing, developments are on-going in the COPKIT project so that other tools support the output formats provided by the information extraction tools and can access persisted outputs via the DSR API. As this format is the W3C proposed standard, extra dissemination effort towards LEA partners in COPKIT will be made to encourage adoption. With respect to inputs, the proposed solution offers the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

appropriate flexibility and adaptability without making hypotheses for specific application and seems the best compromise in the current situation. Adaptation can be made at the time of productising, if the acquiring LEA formulates different requirements.

7. Knowledge representation format

7.1. Introduction

A key differencing component of the COPKIT ecosystem is the COPWIK knowledge base. COPWIK serves two main purposes: enrichment of extracted information with a priori and learnt knowledge, and storage of consolidated findings obtained by other components. Technically, COPWIK provides machine-readable knowledge through different mechanisms (direct query, APIs, web-based interface, etc.) to avoid the need to understand the internal details of the representation for other COPKIT tools developers.

Given the complex nature of this type of knowledge, it is not possible to use classic structured representation models, such as the relational model (implemented in relational database systems), and ad hoc unstructured simple models, such as NoSQL documents or key-value stores. At the same time, knowledge in COPWIK is represented in a common and reusable way. For these reasons, we have selected standard graph-based semantic languages, i.e., RDF (Resource Description Framework) and OWL (Ontology Web Language). The first part of Section 7.2 describes in more details these standards.

Here we must distinguish two aspects related to standardization: (1) the low-level format in which the knowledge is encoded, (2) the meta-models created as templates for knowledge instantiation. For (1), we considered the previously mentioned RDF/OWL standards. For (2), we reuse as much as possible widely used models, which are available for general knowledge (e.g., authorship via Dublin Core) but not for domain-specific knowledge (e.g., there is not a formal specification of firearms types at the European or the country levels). The second part of Section 7.2 describes in more details these meta-models.

Based on the decision made in COPKIT (Section 7.3), we have identified several directions and plans for further actions. These proposals are described in Section 7.4.

7.2. Overview of explored / existing standards and accepted formats

In the last two decades, the most influential contributions towards standard knowledge formats have arisen in the context of the Semantic Web. The Semantic Web was an initiative aimed at extending the current web by formally describing and linking “knowledge items” rather than documents. This web of interconnected resources is nowadays known as the Linked Data Cloud. Over the years, the Semantic Web has crystallized in several languages for describing these knowledge items, most notably RDF and OWL. The standardization of these languages is managed by the World Wide Web Consortium, which is also responsible for other web languages, such as HTML, CSS and XML. RDF and OWL are extensively described in Deliverables D5.1 and D5.2. Here we include a short overview of their features.

RDF (Resource Description Framework) is the W3C standard language to describe resources in the Semantic Web. RDF allows data representation in the form of triples, i.e., statements <s, p, o> relating a subject s, a property p, and an object o. This simple mechanism allows representing any kind of information by creating a number of triples that reuse any element of other triples <s, p, o>. This construction is named knowledge graph. RDF graphs can be serialized (i.e., represented as plain text) in several formats. One of them is the standard XML-based syntax, which is often criticized for its verbosity. Other RDF serialization formats are N3, proposed by the W3C itself; Turtle, a subset of N3; and N-Triples, a simplification of N3. Note that most model development and programming tools support reading and writing of RDF documents in any of these formats.

As a side note, it is worth mentioning that there are other languages for knowledge graphs, mostly arisen from industrial practice and associated to specific graph database technologies. For instance, Neo4j, a proprietary graph database uses the Cypher Query Language for query and update. These languages are not standard, and therefore they were not considered in COPKIT.

RDF offers little restrictions regarding the form and the contents of triples and knowledge graphs. That is, RDF can be used to represent factual knowledge, to encode taxonomical relations, and to define a

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

knowledge model. By factual knowledge, we mean explicit asserts about a domain, e.g., the capital of Spain is Madrid. By taxonomical relations, we mean meteorological axioms, e.g., a City is a Location. By knowledge model, we mean template specifying a template for the former asserts, e.g., any city shall define a population value. OWL is a logic-based extension of RDF, in the sense that OWL defines specific RDF tags for triples elements with special semantics enabling inference. For instance, in OWL we can assert that a property r is transitive, which means that <a, r, b> and <b, r, c> entails <a, r, c>. This more expressive set of statements expressed in OWL is known as ontology. As in RDF, OWL conceptual representations can be serialized in different formats, most of them approved by the W3C or in process, e.g., the OWL/XML syntax and the Manchester Syntax.

The query language for RDF is SPARQL. With SPARQL, we can specify a set of triples with variables in any subject, predicate or objects of any triple. A SPARQL engine will resolve this query by finding mappings of variables to actual values inside the knowledge base that realize the query (i.e., the possible ways the KB can satisfy the query). SPARQL offers little support for OWL inferencing. To that aim, there exist specific software components known as reasoning engines, which support resolution of more complex knowledge conditions. In contrast to SPARQL, there is not a formal definition of a query language for OWL, and each knowledge engine may use a different language, either new or derived from SPARQL. The main effort aimed at the unification of OWL query is the OWL API, which is supported by most reasoning engines.

As explained, RDF and OWL does not preclude the specific knowledge to be represented, but only provide a way to express them. They are similar to programming languages: they offer primitive representations that must be combined to create a product usable for a given problem. As mentioned in Section 7.1, in the context of COPKIT we need two types of models: general knowledge and domain-specific knowledge.

For general knowledge, COPWIK relies on several publicly available ontologies. To name some of them, COPWIK uses:

- Dublin Core (DC) for metadata about authorship, resource identification, etc. DC is managed by the Dublin Core Metadata, a section of the ASIS&T professional association in information sciences. [de facto standard]
- PROV-O for data provenance. PROV-O is a W3C standard to represent and exchange provenance information generated in different systems under different contexts. [standard]
- SKOS for additional taxonomical information. The Simple Knowledge Organization System (SKOS) is the W3C standard for defining thesauri and taxonomies in complex organizations. [standard]
- DBPedia for general-purpose models and structured instances extracted from Wikimedia products (e.g., Wikipedia). DBPedia is developed by the Leipzig University and the University of Mannheim, with contributions from many other institutions. [widely used, but not standard]
- Geonames, for geographical identifiers and associated data by the Geonames organization. [widely used, but not standard]
- NUTS (Nomenclature of Territorial Units for Statistics), the RDF version of the classification defined by Eurostat office. [de facto standard, described in EC Regulations]

For domain-specific knowledge, Law Enforcement Agencies lack well-defined standards for most of their tasks. EMPACT and Europol provide several document references for different aspects of organized crime, but these are not widely agreed among the involved stakeholders in different countries. Section 8 elaborates on these limitations.

7.3. Action taken in COPKIT

COPWIK was designed and implemented with the aim of facilitating the reuse and the incorporation of existing resources in the Linked Data Cloud. This motivated the choice of standard RDF and OWL as representation languages. Besides, given our previous experience in the project ePOOLICE, and the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

possibility of reusing part of the components already implemented using the Semantic Web stack, led to this decision. Whenever possible, we used standards for the general knowledge models, e.g., DC, SKOS and PROV-O.

7.4. Plans for further actions

Within the scope of COPWIK, we developed a meta-model for the representation of uncertain and imprecise data in ontologies. This model could be pushed forward to serve as the basis of a future standard, in line (and collaborating with) other initiatives such as the Uncertainty Reasoning group of the W3C (now inactive) and the URREF (Uncertainty Representation and Reasoning Evaluation Framework) initiative for the management of uncertainty in Information Fusion problems.

Furthermore, we have developed a model for the qualification of information quality in terms of reliability and credibility (which actually uses the uncertainty representation ontology). This model is inspired on the JC3IEDM, defined by NATO STANAG 5525 and used for the exchange of data in joint military operations, and NATO STANAG 2511, a standard that defines credibility and reliability. A joint effort to assimilate and leverage these representation models should be considered in the future.

In addition, it would be interesting to agree on an upper-level ontology for framing the concepts used in the security domain at the European level. Currently, there are a few candidates that could serve to that purpose, such as the Basic Formal Ontology (BFO), the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), and the Suggested Upper Merged Ontology (SUMO). The actions in that direction should be coordinated with actions regarding criminal domain semantics in section 8.

Concretely, the actions envisioned are of two types:

- Disseminate COPKIT advances and existing initiatives mentioned above in the LEA community. As problem owner, LEAs are in the position to encourage technical development to build upon the existing blocks when producing tools. This can take place through the COPKIT project meeting with COPKIT LEA partners and dissemination activities (a.o. COPKIT demonstrations event at the end of the project, planned activities with linked projects, and CoU events organised by DG-HOME if possible.) involving other LEAs as well.
- Disseminate COPKIT advances and existing initiatives in the Security Research community. This can take place:
    - via COPKIT dissemination activities, including towards other EC funded research projects. The goal is to build a community of researchers in security involved in the field to coordinate as much as possible and in an opportunistic manner the on-going research actions. This action is already started, see also section 8.4.
    - via contributions aiming for the definition of future EC research topics, for inclusion on the research agenda. This action is already started, see also section 8.4.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

8. Criminal domain semantics and formal models

8.1. Introduction

Criminal domain-specific knowledge in COPKIT allows enriching case information processed during the investigation. In particular, the COPKIT knowledge base (namely, COPWIK) can expand the information extracted from texts with specialized knowledge to leverage the capabilities of other components of the AI toolbox. A typical example of this enrichment is attaching to a firearm found in advertisements its physical characteristics and the regulations applicable in different EU countries.

Knowledge in COPWIK focuses on the “criminal side” of the investigation, rather than the investigation process itself or the exchange of information among LEAs. This makes the COPKIT approach different from other formal models developed in the area, such as the CASE ontology or the UMF format.

The lack of specialized public and private knowledge resources with the potential to be incorporated into COPWIK has required the creation of new models from scratch (see Section 8.2). To do so, several knowledge acquisition sessions were held with LEAs in order to identify sources and elaborate materials to be formalized (Section 8.3). In line with the considerations made in Section 7, domain knowledge of COPWIK has been encoded by using the semantic languages RDF and OWL. Based on these experiences, we have identified several directions for future research (Section 8.4).

8.2. Overview of explored / existing standards and accepted formats

The main effort on formalization of (cyber-)crime investigation at the European level is the Cyber-investigation Analysis Standard Expression (CASE) ontology (https://caseontology.org/). This ontology is aimed at representing the attributes of “traces”, i.e., the primary observable objects in the investigation of digital crime including data sources (storage units, devices, etc.) and digital objects (files, messages, etc.). Among the use cases of CASE, we can find exchanging information in a standardized form, maintaining provenance at all phases of cyber-investigation, and marking data for controlled access to privileged, proprietary and personal information. CASE is developed in OWL, and the first stable version was issued in Aug 2020, following an initial prototype on Jan 2017. While there are several ongoing initiatives and projects that are using or considering CASE (e.g., INSPECTr, ASGARD), there is not yet a consensus on its use on larger applications.

CASE is built on the Unified Cyber Ontology (UCO) (https://github.com/ucoProject/UCO), which is defined as a model of “classes of cyber objects (e.g., items, tools, people, places), the relations to other cyber objects, provenance of items and actions taken in an action life-cycle”. In fact, UCO is not a precursor of CASE, but a complimentary asset originated from the identification of needs for a more general knowledge framework by the CASE community. In this regard, UCO does not cover but favour the development of specific information representations focused on individual cyber security domains.

Another ongoing initiative is UMF (Universal Message Format), a standard defined by Europol for the exchange of LEA data across European borders. (https://op.europa.eu/s/ozu8). The purpose of UMF is acting as an intermediate representation (inter-lingua) between formats used by different and independent law enforcement systems, and therefore it does not require rebuilding national systems, legislations or processes. Accordingly, UMF is based on XML schemas, that is, it establishes the format of the UMF XML documents, but it does not require a strong commitment on the semantics of the represented data. UMF was first proposed in 2014, but it has not been until recently that it has gained traction due to the regulations issued by the EC, such as the ‘Regulation (EU) 2019/818 of the European Parliament and of the Council of 20 May 2019 on establishing a framework for interoperability between EU information systems in the field of police and judicial cooperation, asylum and migration and amending Regulations (EU) 2018/1726, (EU) 2018/1862 and (EU) 2019/816’. This regulation establishes that UMF “should serve as a standard for structured, cross-border information exchange between information systems, authorities or organisations in the field of Justice and Home Affairs”. This regulation explicitly acknowledges UMF as the reference

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

format for "cross-border information exchange between information systems, authorities or organisations in the field of Justice and Home Affairs". Furthermore, as far as the information available to the authors, the UMF focuses on persons and identities. It does include means of representing rich elements of criminal activities and modus operandi.

8.3. Action taken in COPKIT

The knowledge representation models mentioned in 8.2 (CASE, UCO, UMF) have not been used in the creation of the COPKIT knowledge base. The main reason is that they focus on investigative actions and do not provide the kind of domain-specific knowledge that is required to materialize the COPWIK vision. There are still some primitives in CASE and UCO that could be useful to frame the knowledge in COPWIK. For example, the firearms taxonomy (see below) could be used to specialize the Product category of CASE and as a source of instances for a given use case. A second reason is that the level of maturity of CASE and UCO at the time of the design of the conceptual structure of COPWIK (task T5.2) was still preliminary.

In line with the COPKIT requirements, we have developed two specialized ontologies based on the contributions of the LEAs involved in the project: (1) the firearms ontology, and (2) the CaaS ontology. Both ontologies are represented in OWL.

The firearms ontology contains concepts and inference rules about firearms categories according to EMPACT and the Spanish Regulation on Firearms, as well as about the main actors involved in firearms trafficking. In addition, the firearms ontology also gathers data on firearms models from sources recommended by LEAs, specifically, DBPedia[8] and the Internet Movie Firearms Database (IMDB[9]. The firearms ontology is not currently publicly available.

The CaaS ontology contains concepts about categories of 'crime as a service' events, e.g., selling of data as a commodity, translation services, phishing kits, etc. The CaaS ontology incorporates knowledge from a few external sources. Most notably, following LEAs recommendations, we have brought into the ontology the contents of the Mitre Common Vulnerabilities & Exposures List (CVE), the most widely used database of software vulnerabilities. Along with the core CVE, we have also incorporated a vocabulary of terms frequently used in the cyber-criminality domain. Besides, COPWIK is expected to provide a transparent gateway to other external databases of known security and data leaks, e.g., the "have I been pwned" web page (https://haveibeenpwned.com/). The CaaS ontology is not currently publicly available.

8.4. Plans for further actions

As an immediate exploitation of the results, the COPKIT team is discussing with Guardia Civil to assess the opportunities to disseminate the results of taxonomy of firearms to the LEA community. One option would be to disseminate it via the EMPACT working group with the goal of improving it (if necessary) and hopefully increase adoption.

It can be envisioned to link the domain knowledge ontologies developed in COPWIK to UMF and CASE representations. However, particularly with respect to UMF this would have value in the COPWIK (or more generally the COPKIT eco-system) if the COPWIK is directly interfaced with the systems for cross-border Exchange of information. At the time of writing, there is no clear request for this function and thus the implementation can be postponed to the productising phase.

Further, during the project activities, it appeared clearly that several other H2020 projects were carrying out activities pertaining to domain knowledge ontologies. Using the network of COPKIT and linked Project, the team engaged in discussion with other H2020 projects (MAGNETO / PROVISION, CC-DRIVER and

[8] See https://en.wikipedia.org/wiki/DBpedia for a definition and https://www.dbpedia.org/ for access. (last accessed, Jan 15th, 2021)
[9] See http://www.imfdb.org/ (last accessed, Jan 15th, 2021)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

more projects have been contacted) to investigate the different domains covered. The following conclusions were reached:

- Two approaches co-exist: models taking the point of view of the investigator (e.g., modelling evidence pieces, there type etc.) or models taking the point of view of the criminal activities and modus operandi. The latter is the COPKIT approach. While both approaches are likely to reach a common area at some point, so far, the knowledge modelled is very disjoint of far, not easily allowing re-use.
- Among the Project developing models from the point of view of the criminal activities, projects tend to tackle specific type of crime. For instance, firearms and CaaS for COPKIT. For these projects, collaborations are easier as it only involves a common translation of high-level concepts.
- At this stage, an overview of which crime type has been studied and to which point does not seem to exist. Such an overview would help future projects and LEAs to identify the gaps and develop a more comprehensive knowledge model.

Based on these conclusions, the COPKIT project has launched actions with other EC research projects (MAGNETO / PROVISION and CC-DRIVER) with the following goals:

- Identifying means to have LEAs take ownership of these knowledge models and making them available for future development. This action will involve contacts with DG-HOME during the project and possibly support from DG-HOME,
- Disseminate the analysis and existence of gaps in the Security Research community as a topic to be put on the research agenda for EC funded project. This action will involve contacting and possibly obtaining support from DG-HOME.

9. Data exchange formats for graphs

9.1. Introduction

Graphs dataset are occurring very frequently in the context of LEA analytical tasks. It is frequently used to represent relations between people, identities or other entities. A graph is represented by a set of objects, which can be a node or an edge. Typically, nodes represent entities or instances such as people, businesses, accounts, or any other item to be tracked. Edges, also called termed relationships, connect nodes to other nodes; representing the relationship between them. Both nodes and edges can have properties associated to them. For nodes, this will be the identifier of the entity, its nature and possibly other data associated to it. For edges, this can be the semantic of the relationship and its direction. In COPKIT a number of tools aim at analysing graph datasets (Connection Finder and Graph Partitioning, see Deliverable D6.5), producing datasets that can be represented as a graph (CKRELEXT see Deliverable D4.3, FIS/ARD, see Deliverable D6.5) or are using underlying models consisting of graphs (COPWIK, see Deliverable D5.2 or CTSAE see Deliverable D6.6). Agreement on representing this type of data is therefore relevant.

When considering exchanging graph data, the nature of the publishing and consuming sides should be taken into account. Graph data can be exchanged:

- Between automatic analysis tools using specific algorithms aiming at determining properties of the graphs (connectivity) or relationships between individuals: for instance, the existence of relationships and paths such realised with the COPKIT tool Connection Finder (see Deliverable D6.5).
- Between an automatic analysis tool and a visualisation tool, with the goal of performing visual analytics in which the expert performs the analysis based on a presentation that aims at facilitating the recognition of patterns.
- Between a visualisation tool and an automatic analysis tool. This can occur when an expert assembles data in the purpose of further analysis. For instance, one can imagine a situation in which an expert selects parts of an existing graphs and wishes to export it towards a automatic analysis tool.

For efficiency reasons, automatic analysis tools and visualisation tools favour different internal representation of graphs. When the input exchange format is not appropriate, the tool will need to perform a conversion to an adapted representation. This conversion can be computationally costly and can lead to performance degradation. For instance, let's assume that two automatic analysis tools are used successively and exchange their data using a format that is appropriate for visualisation but not appropriate for computation. A double conversion is then performed, one by the publisher and one by the consumer, leading on inefficiency. This challenge should be taken into consideration if a unified exchange format is desired in an eco-system of analysis tools and visualisation tools. Further details on the different needs of consumer and publishers are described in the remaining of this section.

Automatic analysis algorithms for graph tend to be very computationally expensive and efficient implementations generally rely partially on a representation that is efficient for this algorithm. Very often the entire graph has to be loaded in memory which creates a need for an efficient representation in size. Often properties are not useful during the computation (with the exception of direction and weights of the edges) and are therefore dropped. Therefore, self-describing representations (such as JSON and XML) are rarely favoured for computation purpose also due to their verbosity. The main representations used for computations are:

- A list representation. Typically, a list of edges, with, separately an optional list of nodes (to avoid repeated computation) in particular when nodes have properties. This representation is very efficient in term of size and for some algorithms. Sparse matrices can be represented in this way.

- A matrix approach with each column and row being a node and the cell being the weight of the edge between them. This representation is not very efficient in term of size especially if the graph is sparse (it represents non existing links) unless specific schemes are used. Operations on sparse matrices are an important field of numerical analysis, and advanced schemes have been developed and stabilised since the 1970s.
- Adjacency lists. A list of nodes in which each row describes all the relationships of a given node.

On the other side, visualisations have different requirements. For visualisations, the favoured representations keep close together objects and their properties (likely to be displayed at the same time) and tend to use self-describing schemes (such as JSON) to provide flexibility for the rendering of properties. For large graphs, the performance (speed) of visualisation is a serious challenge and representations are generally tailored for that purpose.

The conflicting requirements were noticed during the development of the COPKIT tools. Further, graph data was also noted by LEA representatives in their responses to the questionnaire as a typical example of the cases in which formats should be adapted to the goal (see section 3.2.5) and flexibility is required. To the knowledge of the authors, there is no clear path for resolution of this conflict at the time being.

9.2. Overview of explored / existing standards and accepted formats

For computation-oriented formats, the analysis could not show that a particular format is dominating. This may be due to the fact that graph analysis is a very active research field with dominant computation library changing quite rapidly.

At the end of the 1990s, the National Institute of Standards and Technology (NIST) supported the creation of the Matrix Market format[10]. This is a very simple text-based representation of graph, which takes as a starting point the graph as a matrix concept. It results in a list of edges type of inputs. Many tools including those proposed by the NIST read this format. More recently the SNAP project[11] developed by Stanford University takes edges list as input format. With this project, a many implementations of analytic approaches are available as well as a significant number of datasets for benchmarking purposes.

A number of initiatives are proposing richer formats. The most popular at this stage is probably GraphML[12], a format that is supported by a large number of tools. GraphML is formulated as a standard and thus well defined. GraphML is XML based. XML often considered very verbose and not very flexible with JSON often being preferred in modern implementation. Furthermore, two initiatives are attempting to define JSON formats for graphs data:

- The JSON Graph Format (JGF) (http://jsongraphformat.info/), since it already contains a number of extensions that contribute to capture the basic graph structure. Some of the abovementioned extensions could be proposed as enhancements to JGF for a more powerful standard for graph data exchange formats.
- The JSON-LD (JavaScript Object Notation for Linked Data) is developed by a specially created working group[13] of the W3C consortium launched in 2018 with the goal of producing a JSON format for graph aiming at capturing semantics. This will add to the broader interoperability of graph resources towards a universal data exchange format for graphs, similar to the idea of the semantic

[10] R. Boisvert, R. Pozo and K. Remington, The Matrix Market Exchange Formats : Initial Design, National Institute of Standards and Technology Internal Report, NISTIR 5935, December 1996. See also https://math.nist.gov/MatrixMarket/info.html

[11] See https://snap.stanford.edu/index.html (last accessed Jan 14th, 2021)

[12] See http://graphml.graphdrawing.org/specification.html (last accessed Jan 14th, 2021)

[13] See https://www.w3.org/2018/json-ld-wg/ (last accessed, Jan 15th, 2021)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

web and the RDF (Resource Description Framework), JSON-LD is a W3C/IETF[14] standard and a World Wide Web Consortium Recommendation. Despite this effort, the analysis and experience in COPKIT did not come up with a large community of users implementing it which limits its value at this stage for interoperability.

It should be noted that JSON-LD further provides an interesting data exchange format for semantic graphs like COPWIK's RDF knowledge graph described in Section 7.2. Since data graphs also can be represented in JSON-LD, application of this as a common format will broaden the interoperability and joined exploitation of knowledge and data graphs.

Among the commercial tools, it is important to mention IBM Analyst's Notebook as it is a well spread tool among criminal analysts. This tool uses a proprietary format (.anb) but also includes an XML format for exchange (.anx) and import / export functions are provided. Similarly, KNIME (which has both a commercial and an open-source version) provides built-in import and export for a very limited number of formats but the transformation functions allow the user to implement readers for various list formats.

9.3. Action taken in COPKIT

While the COPKIT project produces a number of tools analysing or producing graphs, it is only a fraction of the range of graph analysis approach. An important concern when choosing the format within COPKIT project was therefore the interoperability. Tools opted for a flat text approach with list of nodes and edges accompanied with their properties, typically in CSV (Comma Separated Value form) or equivalent. Such format is also easy for a human to read, easy for software to manipulate (no schema needed for parsing) but has the drawback that it cannot capture graph level properties. When a richer format was needed, JSON was favoured over XML. Among the advantages of JSON over XML, one can name[15]:

- JSON is not a document markup language, so it is not necessary to define new tags or attributes to represent data in it, being therefore more dynamic to use.
- JSON is processed more easily because its structure is simpler. There is a wide range of reusable software available to programmers that accept JSON format for graph visualization.
- JSON is easier to read by both humans and machines.

The Graph Partitioning tool opted for the Matrix Market format, in particular because it does not make direct use of properties of nodes and edges (with the exception of the weight).

The Connection Finder tool makes use of such information and implemented a dual approach with both a CSV representation and a JSON representation for graph data exchange.

Regarding the FIS/ARD tool, the output (a set associations rules) is approached as being a model of the data (e.g., the discovered structure of the data). This case is developed in section 11. The tool also opted for a dual approach with an output for processing and an intermediate format for visualisation using a JSON representation (described in Deliverable D6.5).

Regarding the CTSAE tool, the graphs used are considered models (in fact Probabilistic Graphical Models) and this case is developed in section 11.

The experience from the Connection Finder gave rise to proposing the following extensions of the JSON format for graph definition:

- **Graph type**: directed *or* undirected
- **Weighting**: weighted *or* unweighted
  In case of weighted:

[14] The Internet Engineering Task Force (IETF) is an open standards organization, which develops and promotes voluntary Internet standards, in particular the standards that comprise the Internet protocol suite (TCP/IP). (https://en.wikipedia.org/wiki/Internet_Engineering_Task_Force)
[15] See http://www.json.org/xml.html for instance for a comparison between JSON and XML.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

  - type: cost (distance) *or* strength (certainty)
  - normalization: unnormalized *or* normalized
  - scale: from (value) to (value)
- **Item types**:
  - Node type ID: (example) 1: person, 2: company, …
  - Edge type ID: (example) 1: works for, 2: located in, …
- **Node ID scheme**: serial number from 1 *or* unique string
- **Items**:
  - Node: Node ID, Node type ID
  - Edge: Start node ID, End node ID, Edge type ID, Weight

The format includes the possibility of extending nodes and edges with domain specific attributes.

The relevant distinctions were encoded in the software. For instance, the investigative graph applied in the Connection Finder is undirected and weighted (strength weighting, normalized in the interval [0, 1]), while the social media graph applied in Situation Assessment was directed and unweighted.

In the representation of the graph data in the Connection Finder, we used type ID numbers for node and edge types, which has the advantage that the exchange file takes less space and it ensures that each item's type is one of the standard types for the graph. On the other hand, showing the type names (instead of the type ID) for nodes and edges will make it more readable for humans, which is one of the advantages of JSON formats..

Please refer to section 9.4 for further actions.

9.4. Plans for further actions

While a standard exists, the analysis shows that it is very recent, not widely spread and is more adapted to visualisation tools that to automatic analysis tools. The value of its usage and implementation during the project are therefore likely to provide limited contribution to the interoperability of COPKIT tools with other tools already owned by LEAs and to the take-up of the tools.

Furthermore, implementation of the standard would not be an internal representation (as it is not efficient for most computations) but an import/export function not modifying the core of the tool. Such implementation can be efficiently and without risk carried out in the productising phase (TRL 7-8) and be tailored to the needs of the specific LEA acquiring the tool, taking into account their legacy tools performing graph computation.

However, the situation may change rapidly if the standard is adopted. The technical partners will follow-up on the activity of the corresponding W3C work group. It is also strongly recommended that the situation is re-assessed at the time of productising and actual take-up. In particular, once the standard is broadly adopted, it is possible that software libraries will be developed offering highly optimised conversions in various programming language. If such libraries exist, their incorporation in the automatic analysis tools could be a solution for the efficient usage of a visualisation for both automatic analysis tools and visualisation tools.

Finally, please note that section 11 provides details for the situation in which the graph is considered to be a model of the data.

10. Spatial temporal data standards and format

10.1. Introduction

The COPKIT project develops a number of tools that are targeting spatial-temporal analysis. While spatial temporal analysis is a long-standing subject of research, the representation of spatial temporal data remains not trivial. To a large extent, this is due to the complexity and variability of the objects to represent. For instance, among the many different types of spatial data, one can name (following the MADS model[16]):

- Simple such as points, convex areas, and (oriented) line.
- More complex such as point sets, complex (disjoint) areas, (oriented) line sets.

For time as well, different types of objects can be identified, such instant and period. Both time and space involve the notion of referential system of coordinates (calendar system for time, projection system in space). This leads to a large number of possible objects to represent: for instance[17], Moving features, discrete events, areal features etc., some of them requiring many parameters for definition. A generic definition that is also suitable to accommodate the most complex object is likely to become extremely verbose when processing a simple object. Further, it should be noted that the representation of a spatial object can very easily get heavy: the representation of a polygon such as a country a decent resolution requires thousands of latitude-longitude pairs to represent the borders. Storage and manipulation efficiency are therefore a challenge. In many situations, geographical objects that are static at the scale of the analysis can better be represented separately and the corresponding attribute value at a certain time will refer to an idea, avoiding the repetition of the definition of the object.

There are also conflicting requirements between representations suitable for computation and those suitable for visualisation. Regarding computation-oriented representations, several approaches are frequently used depending on which dimension is used for the table:

- The time is used as the outer dimension. In a database, each table would have spatial indication as row, column would be values of attributes and each table would be a different time snapshot. This approach is frequently used by systems that give priority to the representation of such as GIS's.
- The attribute is used as the outer dimension. In a database, each row is a different area and time and the value of specific attribute, different tables are used for different attributes.
- A flat form location, time, set of pairs attribute identifier - attribute value. If attributes are “synchronised”, e.g., time and location of measurement are identical, this can be reduced to a singletable structure with column name being attributes.

The dimensionality of spatial temporal data (3 dimensions plus values) results in computational costs even with relatively simple computation. For spatial temporal data, intermediary computations are therefore frequently stored for re-use (for instance coordinate to area aggregations or neighbours matrices).

Regarding visualisation, GIS systems are well developed systems aiming at representing spatial (in the sense of geographical) information. They have been part of the IT tools of LEA for a long time, due to the importance of geographical information in LEA tasks (addresses etc.). Generally, LEAs have their own system developed relying on the state's resources (such as the national cartography institute), as was confirmed by LEA inputs (see section 3.2.4). GISs may be used as a visualisation for outputs of COPKIT

[16] See Parent, C., Spaccapietra, S., and Zimányi, E.. **1999**. Spatio-temporal conceptual models: data structures + space + time. In Proceedings of the 7th ACM international symposium on Advances in geographic information systems (GIS '99). Association for Computing Machinery, New York, NY, USA, 26–33. DOI:https://doi.org/10.1145/320134.320142

[17] The presented taxonomy is inspired by https://www.esri.com/about/newsroom/arcuser/working-with-temporal-data-in-arcgis/ (last accessed on Jan 15th, 2021) but is only provided as an example.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

tools. In general GIS systems work with layers starting with a cartography layer. GISs may provide several ways to represent spatial temporal data (depending on their nature):

- Sequential snapshots. Time is seen as the layering element.
- Space time composite. Different times are projected on the map as special markers.
- Space time object. Each spatial object (and attached attribute value) is evolving over time and its representations over time are linked.

Modern GISs offer a number of input formats for spatial temporal data, suitable for different purpose, including feature maps (in which the layer is two-dimensional value – time) the most common in GISs views, mosaic (rasters in which time evolution is projected) and NetCDF (4 dimensions matrix of one or several attributes, mostly used for environmental data. It should be noted that GISs are complex and heavy tools. Commercial GISs licences are expensive and require hardware often above the standard individual workstation.

10.2. Overview of explored / existing standards and accepted formats

The tools produced in the COPKIT project use two types of spatial temporal data:

- Spatial temporal discrete events (possibly having properties) which typically can be represented by a triplet latitude, longitude and time of occurrence, without a “lifespan”.
- Small area – period features in which the value represents a synthesis (aggregation, average etc.) of a certain feature over a certain time period for each area defined in a set of static areas. This type of data can be an aggregation of discrete events for an area and a time period.

In this section, we will limit ourselves to these types of spatial temporal objects.

The data used as inputs are expected to come from LEA databases, mostly from relational databases and exports (and query responses) are likely to be tables with columns for spatial info (an area id or a latitude – longitude) a time indication (time stamp also representing a time period) and columns for values. This is also a format that is computationally suitable and that humans can examine (although the spatial proximity is likely to be lost for human examination).

In the specific case of small area data, the geometry of areas (precinct, district, countries etc.) is needed in order to compute different forms of distances (or neighbours). A number of formats can represent static geographical data. The following three are worth mentioning:

- Shapefile[18]. The shapefile format is developed and maintained by ESRI, a leading supplier of GIS system. It is an open specification for a large part. However, it is relatively complex, being actually a set of files and not human readable. Many GISs can read shapefile including, of course ESRI’s ArcGIS. It is possibly the dominating format at the time being. Eurostat (and many national statistical institutes) publishes administrative boundaries (NUTS and LAUs) in shapefile format.
- GeoJSON. More recent, GeoJSON, using JSON syntax to define objects, seem to be picking up support rapidly. GeoJSON is an open “proposed” standard created and maintained by IETF under RFC 7946[19]. A recent extension TopoJSON[20] (not standardised yet) adds topological features (notion of common boundaries facilitating the identification of neighbours) is supported by a number

[18] See the specification at https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[19] Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., and T. Schaub, "The GeoJSON Format", RFC 7946, DOI 10.17487/RFC7946, August 2016, <https://www.rfc-editor.org/info/rfc7946>. Accesible here https://tools.ietf.org/pdf/rfc7946.pdf (last accessed January 15th, 2021)

[20] See https://github.com/topojson/topojson-specification/blob/master/README.md (last accessed January 15th, 2021)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

of open-source GISs (such as PostGIS[21]). Eurostat publishes recognised administrative boundaries (NUTSs and LAUs) in GeoJSON and TopoJSON formats as well.

- KML is sometimes found. KML is recognised as a standard by the Open Geospatial Consortium since 2008 (latest version 2.3 in 2015). Google is using KML for Google Earth. It is using XML syntax. However, the authors did not find many other tools actively using it. Still, it can be a practical way to overlay objects on terrain information. A number of tools can import and transform KML files. It seems less used as an exchange format. The authors expect that it will be slowly superseded by GeoJSON.

In addition to these existing formats there are also projects which aim to provide universal conversion abilities such as GDAL[22] by the Open Source Geospatial Foundation. This software library allows for the ingestion of a wide range of raster and vector formats, including all formats mentioned above. The library represents the data in a unified internal data model where the data can be processed or saved as a different data format. This library is used in a large number of GIS software[23] such as Google Earth, allowing these applications to avoid choosing a particular data format.

Further, propositions have been made by the research community to incorporate semantic into the spatial temporal data for GIS as a manner to resolve the different formats and nature of data observed[24]. To the knowledge of the authors, these approaches have not yet been implemented in tools.

Finally, the COPKIT tools make use of spatial temporal statistical data such as economic or social indicators. Technically, this type of data is generally a form of small areas data. However, the value represents often complex concepts resulting from counting choices, evolving definitions (think of the definition of unemployment or economic activity) and statistical transformation (rectifications for seasonality etc.). Such data typically requires significant additional information and expertise to be interpreted correctly. Eurostat contributed to the definition of the Statistical Data and Metadata Exchange format (SDMX), an XML based format for statistical data which was approved as a standard ISO 17369:2013. It is acknowledged by major international organisations (ECB, OECD, World Bank). It presents clear advantages for automatic retrieval and processing as it caters for non-ambiguous definitions of the indicators and the meaning and handling of exceptions. A previous EC funded research project (FP7 EPOOLICE) demonstrated how the indicators can be represented in a knowledge base and client services can use the knowledge base to automatically retrieve Eurostat data and format it to their convenience. However, SDMX is very heavy to implement and it did not spread much in the community of open data with national institutes even implementing it sparsely. While technically human readable, it is practically impossible to use it for data exploration, certainly compared with a traditional table representation.

10.3. Action taken in COPKIT

With respect to static geographical objects, the COPKIT project is using GeoJSON files as inputs for static objects such as area boundaries due to its status of “proposed standard” and practicality, although it is possibly less efficient (storage wise) than shapefiles. In addition, it is now proposed by Eurostat and several national statistical institutes. COPKIT tools incorporated some pre-processing functions and pre-computed intermediate input data (neighbouring matrices for specific regions and countries tuned to the partner LEAs) to facilitate the testing and evaluation.

[21] See https://postgis.net/, (last accessed January 15th, 2021).
[22] https://gdal.org/
[23] https://gdal.org/software_using_gdal.html#software-using-gdal
[24] Ferreira, K. R., De Oliveira, A. G., Monteiro, A. M. V., De Almeida, D. B. F. C. Temporal GIS and Spatio Temporal Data Sources. Revista Brasileira de Cartografia, v. 68, n. 6, 11.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Some components (ESTF for one) make use of open source libraries (geopandas[25], ultimately using the GDAL library for geo spatial representation) which handle a number of popular file formats (shapefile, GeoJSON, etc.) and make them available as python data structures for use in data processing.

For the input data themselves, the COPKIT tools rely on flat format as it seems the best strategy to facilitate testing and evaluation during the projects and integration with RDMS of cases or statistics during the productising phase. Specific tuning can be done for the different representation of latitude, longitude and timestamp at productising phase (TRL7-8).

As discussed earlier, a generic visualisation of spatial temporal data including cartography requires a full-blown GIS which are complex and heavy tools. LEAs often have their own GIS which they may want to use for visualisation of the results. Alternatively, the COPKIT project proposes an “HMI for multi-level intelligence analysis” which includes means to represent spatial temporal data as features on a map and time-value graphs (see Deliverable D3.6 for more details). Further, the COPKIT ESTF tool provides means to visualise its results. Regarding the CASTF tool, the possibility of implementing a partial integration with the above mentioned COPKIT HMI is being investigated at the time of writing.

With these measures, the technical team expects that interoperability is sufficient for testing and evaluation within the project. Due to the lack of uniformity, no generic measure can be taken to facilitate further the potential integration with GISs, should it be the wish of the LEA acquiring the tools. It is also possible that LEAs acquiring one of the spatial temporal analysis tools would opt for a strategy of not integrating with their GIS (for instance due to cost of licences). The simple flat file output format is the most flexible option and most cost-efficient.

10.4. Plans for further actions

The tools are already following the existing standards as much as practical. No gaps have been identified. The COPKIT team recommends monitoring the evolution of the market of GIS systems and act accordingly. In particular, possible evolutions aiming at introducing semantic to support heterogeneous sources should be monitored as they are particularly well suited for the COPKIT paradigm and could be supported using the COPKIT Knowledge Base (COPWIK).

The main challenge is the lack of harmonisation of GIS systems and formats across LEAs. This has been noted by respondents to the Task T8.6 questionnaires who expressed their concerns regarding the challenges of identifying and localising foreign addresses (with respect to their area of responsibility). This situation is unlikely to change in the near future, especially as the costs of changes for LEAs are likely to be significant due to the need to update legacy systems consuming the GIS information. While the COPKIT team will attempt to raise awareness of partner LEAs regarding the resulting challenges, the COPKIT does not specialise in GISs and is not in a position to usefully advocate the harmonisation.

[25] https://geopandas.org/io.html

11. (AI) Models

11.1. Introduction

Many of the tools developed in COPKIT use or produce models. These models can be seen as a formal representation of phenomena which, when used in an appropriate execution engine (sometimes called reasoning engine) output new information about a particular instance of the phenomenon represented. The models themselves are strongly related to the specific techniques used by the tool and often the specific function or goal of the model. The construction of models and their use are therefore two different tasks that do not necessarily need to be realised by the same tool.

While this is not a very common approach currently, one can imagine that a tool suitable for model construction would be less efficient for usage (or that it can be cost-efficient to use another tool for the usage phase). Therefore, researching the possibility of standardising model representations can be useful, although probably for long term future. Note that the challenge is not limited to "reading" a specific format. The computations required by the specific model should also be available in the tool chosen for the usage phase.

Models tend to be extremely dependent on the analytical task carried out and the technical approach used in the tool. A model representation standard is therefore extremely challenging as the object to represent are very heterogeneous. In addition, even for the same techniques, different algorithms can have subtle differences. A specific model may rely on a subtle feature of the execution engine algorithm that possibly is not implemented (or slightly differently implemented) in another tool or framework. With sometimes very subtle differences in features of algorithms, the portability to another implementation seems difficult to guarantee.

Furthermore, the AI and ML fields are evolving extremely quickly and new approaches and model types are appearing and fading out at a fast pace. The same is true for tooling, platforms and engines: the dominating tool or platform may switch within a couple of years. The search for a unified format supporting the representation of ***any*** type of model is therefore illusory for the time being and may even be not desirable. However, for a given technique, it may become relevant. We therefore investigate this section means to represent model corresponding to a number of techniques used in COPKIT:

- Language models
- (Associations) rules models
- Bayesian Models
- Deep Neural Network models

11.2. Overview of explored / existing standards and accepted formats

11.2.1. Agnostic model representation

11.2.1.1. PMML (and PFA, Portable Format for Analytics)

The Predictive Model Mark-up Language is an open standard developed since 1997 by the Data Mining Group for XML representation of models. Latest Release is 4.4 (2019). PMML proposes a structure for a number of frequently used techniques and associated models. It can also include transformation functions (support for pre-processing for instance) as well as an input data dictionary to map the inputs. The format is supported by many of the common frameworks both open source (KNIME, Weka, sk-learn…) and commercial (SPSS, SAS…). The main limitation is the set of models that can be represented: for instance, while Naive Bayesian Network can be represented, generic structure Bayesian Networks (as used in COPKIT) cannot be represented.

PFA (Portable Format for Analytics) aims at overcoming the absence of certain models in the PMML specification by offering a functional-programming-like representation of computations required for unsupported models. PFA is fairly new and embraced by a limited set of platforms. A deeper study would be required to evaluate the practicality and extent of the primitive offered to define models. It is indeed not trivial to ensure that any computation can be realised.

During the analysis, it also appeared that PMML was lacking support for traceability and referencing of models with respect to the datasets that it is based on. This is a drawback for usage with models that are strongly related to a particular corpus (for instance NLP models for a certain domain) and for Knowledge Discovery results.

11.2.1.2. ONNX

Originally designed to represent Deep Learning Model, the Open Neural Network Exchange (ONNX)[26] attempts to incorporate what the designers of ONNX call “traditional” Machine Learning (in essence any AI method that is not Deep Neural Network). In principle ***any*** model can be represented. This project was established in 2017 from a partnership with Facebook and Microsoft[27]. Details on its importance for Deep Neural Network are discussed in section 11.2.5.

With ONNX, the models are defined by the operations to perform to compute an output given inputs. The operations implicitly include the parameters of models. This definition includes a computation graph model expressing the computation using built-in operators and standard data types. It is clear that such definition is complex and lengthy to construct. Eventually it can be seen as a form of programming. In practice, its usability will depend in the extent of the set of operators and the complexity of modelling the execution engine algorithm using the set of operators. A deeper study would be required to evaluate the practicality and extent of the set of primitives offered to define models.

In practice, while ONNX has spread well and maybe promising for DNNs (see section 11.2.5), it does not seem to have a large community of users for other ML models. Sk-learn[28] is one of the notable platforms that has implemented ONNX export, although only for a limited subset of the algorithms that it offers[29]. It should be noted that some converter ONNX-PMML have been implemented.

11.2.2. Language Processing Models

The possibilities to store languages processing models are generally linked to the framework used. When DNN frameworks are used (such as Keras, TensorFlow, PyTorch), export functionalities to ONNX are likely to be available (see the discussion in section 11.2.5). Further, spaCy[30] a widely used framework provides means to save the model computed (structure and parameters). spaCy interoperates seamlessly with PyTorch and TensorFlow among others.

Additionally, the NLP Interchange Format (NIF)[31] is worth mentioning. The NLP Interchange Format was funded in 2013 and is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. NIF consists of specifications, ontologies and software. However, there does not seem to be significant community activities in the last five years and it has not been widely used to the knowledge of the authors.

[26] https://github.com/onnx/onnx
[27]https://research.fb.com/facebook-and-microsoft-introduce-new-open-ecosystem-for-interchangeable-ai-frameworks/ (last accessed on Jan 15th, 2021)
[28] See https://scikit-learn.org/stable/
[29] See http://onnx.ai/sklearn-onnx/supported.html (last accessed on Jan 15th,2021)
[30] See https://spacy.io/
[31] See https://persistence.uni-leipzig.org/nlp2rdf/

11.2.3. Association Rules Models

The PMML model supports the representation of frequent item sets and association rules[32] that can be useful for FIS/ARD tool developed in COPKIT (see Deliverable D6.5).

The PMML provides a representation model comprised of four major parts:

- Model attributes.
- Items.
- Item sets.
- Association Rules.

An *AssociationModel* (see Figure 6) can contain any number of *Itemsets* and *AssociationRules* (see Figure 7), taking into account that all *Itemsets,* comprised of *Items* must be listed before any of the rules.

[32] See http://dmg.org/pmml/v4-4-1/AssociationRules.html (last accessed on Jan 15th, 2021)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

```
<xs:element name="AssociationModel">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="MiningSchema"/>
      <xs:element ref="Output" minOccurs="0"/>
      <xs:element ref="ModelStats" minOccurs="0"/>
      <xs:element ref="LocalTransformations" minOccurs="0"/>
      <xs:element ref="Item" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="Itemset" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="AssociationRule" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="ModelVerification" minOccurs="0"/>
      <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="modelName" type="xs:string"/>
    <xs:attribute name="functionName" type="MINING-FUNCTION" use="required"/>
    <xs:attribute name="algorithmName" type="xs:string"/>
    <xs:attribute name="numberOfTransactions" type="xs:nonNegativeInteger" use="required"/>
    <xs:attribute name="maxNumberOfItemsPerTA" type="xs:nonNegativeInteger"/>
    <xs:attribute name="avgNumberOfItemsPerTA" type="REAL-NUMBER"/>
    <xs:attribute name="minimumSupport" type="PROB-NUMBER" use="required"/>
    <xs:attribute name="minimumConfidence" type="PROB-NUMBER" use="required"/>
    <xs:attribute name="lengthLimit" type="INT-NUMBER"/>
    <xs:attribute name="numberOfItems" type="xs:nonNegativeInteger" use="required"/>
    <xs:attribute name="numberOfItemsets" type="xs:nonNegativeInteger" use="required"/>
    <xs:attribute name="numberOfRules" type="xs:nonNegativeInteger" use="required"/>
    <xs:attribute name="isScorable" type="xs:boolean" default="true"/>
  </xs:complexType>
</xs:element>
```

Figure 6: PMML format for the “*AssociationModel*” used to represent Association Rules models (taken from http://dmg.org/pmml/v4-4-1/AssociationRules.html)

```
<xs:element name="AssociationRule">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="antecedent" type="xs:string" use="required"/>
    <xs:attribute name="consequent" type="xs:string" use="required"/>
    <xs:attribute name="support" type="PROB-NUMBER" use="required"/>
    <xs:attribute name="confidence" type="PROB-NUMBER" use="required"/>
    <xs:attribute name="lift" type="xs:double" use="optional"/>
    <xs:attribute name="leverage" type="xs:double" use="optional"/>
    <xs:attribute name="affinity" type="PROB-NUMBER" use="optional"/>
    <xs:attribute name="id" type="xs:string" use="optional"/>
  </xs:complexType>
</xs:element>
```

Figure 7: PMML format for an “*AssociationRule*” element used to represent Association Rules models (taken from http://dmg.org/pmml/v4-4-1/AssociationRules.html)

Still there are several limitations for the representation:

- It does not contain all the meta-data about the experimentation done. It would be useful to have some meta-data about the experimentation like the date and time stamp, name of the database employed, etc.
- It is not flexible / dynamic enough to include other types of association rules like fuzzy rules, or other assessment measures different to those taken into account in the standard.
- The file generated by using the PMML format can be very heavy when many association rules are managed.

Another limitation is related to the fact that Association Rules are both a model and the output of a Knowledge Discovery mechanism. Therefore, the chosen format should be:

- Compatible with visualisation library (in general this will be graph visualisation).
- Suitable for exploration by a human. A human expert should be able to quickly scan the set of association rules discovered.

11.2.4. Bayesian Network Models

The COPKIT tool CTSAE uses Bayesian Network model of **any** structure, both in learning and in usage (classification) phase. In fact, the models developed in COPKIT have a complex structure, not naïve or Tree Augmented naïve (TAN). PMML supports Naïve Bayesian Models only. Sk-learn also supports naïve Bayesian network only and supports their export to ONNX. Both tools are in their current state not suitable for the work in COPKIT.

Separately there exist a few formats specifically addressing Bayesian Network. The two most frequently used seem to be:

- Hugin[33] flat net file format (.net), developed by the Hugin company, one of the two major Bayesian Network software provider. It is a text-based format, key-value structured (except for the tables). This format is very well spread both with open-source platform as with other commercial providers.
- The XMLBIFF format[34] originating from the research community with Carnegie Mellon University as a leader. This format seems less spread and only used in the open-source community. It is more verbose than the Hugin flat net file format but for Bayesian Network models this is unlikely to be an issue.

11.2.5. Deep neural Network Models

Due to the potentially large size and complexity of Deep Neural Network models (Models can be several Gigabytes) Models are generally stored in a binary, non-human readable format.

The Field of DNN is dominated by a small number of popular open-source frameworks (Tensorflow, Pytorch, Cafe, Keras etc), each with its own standards and formats for defining and storing models.

The Keras project represents an early attempt to standardize model definitions under one interface, providing a simple interface to implement models in some of the most popular models at the time (tensorflow, Theano, Deeplearning4j). However, this project was undercut by the fast pace of change in DNN framework development and popularity. New frameworks were developed which Keras did not support, such as Pytorch, leaving Keras as just one more framework to consider.

Most of the major Deep learning organisations have developed their own frameworks such as Google (Tensorflow), Facebook (Pytorch) and Microsoft (Microsoft Cognitive Toolkit). This factor, in conjunction with the rapid pace of development, increases significantly the challenges associated to the task of standardising model definition and formats.

The most recent and active attempt at reaching a standard is the Open Neural Network Exchange (ONNX)[35], already mentioned in section 11.2.1.2. At present the project includes automatic conversion methods for all the popular frameworks and as such, could provide a method to exchange models between frameworks.

In general ONNXs interoperability is made possible from the fact that there is significant similarity in how DNN frameworks represent models at low levels, generally being represented as a computational graph, making use of broadly similar operations (LSTMs, FC, Convolutions). Such similarities likely stem from the need for such frameworks to interact with Hardware accelerators such as GPU libraries like CUDA[36].

ONNX promises the ability for a model to be developed, trained and finally run on different frameworks, this ability may prove to be of some value in the DNN ecosystem as different frameworks offer different abilities and performances when it comes to deployment options and distributed training (federated learning, deployment to mobile environments, model serving as a service, etc).

Time will tell if ONNX will succeed in defining an accepted standard for DNN models, or if it will be made irrelevant by a new framework which rapidly gains popularity. Pytorch, which is now the most popular framework among research[37] and gaining fast in enterprise[38], was only released in 2016. Given this pace of change, it is difficult to justify significant effort into standardising work. Still, the ONNX project shows real promise and should be monitored.

[33] http://download.hugin.com/webdocs/manuals/api-manual.pdf (last accessed Jan 15th, 2021)

[34] See http://www.cs.cmu.edu/afs/cs/user/fgcozman/www/Research/InterchangeFormat/ (last accessed Jan 15, 2021).

[35] https://github.com/onnx/onnx

[36] https://developer.nvidia.com/CUDA-zone

[37] https://thegradient.pub/state-of-ml-frameworks-2019-pytorch-dominates-research-tensorflow-dominates-industry/

[38] https://learning.oreilly.com/library/view/ai-adoption-in/9781492051800/ch01.html#tools_for_building_ai_applications

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

11.3. Action taken in COPKIT

In COPKIT, the favoured framework used for NLP tools was spaCy. The favoured format for the persistence of the models is therefore the one provided by spaCy due to the richness of its community and available models and library for NLP applications. The possibility of conversion to ONNX is still under investigation.

As discussed in section, 11.2.3 the PMML representation has limitation. Therefore, the FIS/ARD tool has implemented an intermediate form, which can be derived from the PMML standard, to efficiently manage and represent association rules for their later visualization. Figure 8 presents this new format.

```
<PMML xmlns="http://www.dmg.org/PMML-4_4" version="4.4">
  <Header copyright="www.dmg.org" description="example model for association rules"/>
  <DataDictionary numberOfFields="2">
    <DataField name="transaction" optype="categorical" dataType="string"/>
    <DataField name="item" optype="categorical" dataType="string"/>
  </DataDictionary>
  <AssociationModel    functionName="associationRules"    numberOfTransactions="4"    numberOfItems="3"
minimumSupport="0.6" minimumConfidence="0.5" numberOfItemsets="3" numberOfRules="2">
    <MiningSchema>
      <MiningField name="transaction" usageType="group"/>
      <MiningField name="item" usageType="active"/>
    </MiningSchema>

    <Output>
      <OutputField    name="Rule    (Highest    Confidence)"    rankBasis="confidence"    rank="1"
algorithm="exclusiveRecommendation" feature="rule" dataType="string" optype="categorical"/>
      <OutputField   name="Recommendation   (Highest   Confidence)"   rankBasis="confidence"   rank="1"
algorithm="exclusiveRecommendation" feature="consequent" dataType="string" optype="categorical"/>
      <OutputField   name="Rule   Id   (Highest   Confidence)"   rankBasis="confidence"   rank="1"
algorithm="exclusiveRecommendation" feature="entityId" dataType="double" optype="continuous"/>
      <OutputField   name="Rule   (2nd   Highest   Confidence)"   rankBasis="confidence"   rank="2"
algorithm="exclusiveRecommendation" feature="rule" dataType="string" optype="categorical"/>
      <OutputField  name="Recommendation  (2nd  Highest  Confidence)"  rankBasis="confidence"  rank="2"
algorithm="exclusiveRecommendation" feature="consequent" dataType="string" optype="categorical"/>
      <OutputField   name="Rule   Id   (2nd   Highest   Confidence)"   rankBasis="confidence"   rank="2"
algorithm="exclusiveRecommendation" feature="entityId" dataType="double" optype="continuous"/>
      <OutputField   name="Rule   (3rd   Highest   Confidence)"   rankBasis="confidence"   rank="3"
algorithm="exclusiveRecommendation" feature="rule" dataType="string" optype="categorical"/>
      <OutputField  name="Recommendation  (3rd  Highest  Confidence)"  rankBasis="confidence"  rank="3"
algorithm="exclusiveRecommendation" feature="consequent" dataType="string" optype="categorical"/>
      <OutputField   name="Rule   Id   (3rd   Highest   Confidence)"   rankBasis="confidence"   rank="3"
algorithm="exclusiveRecommendation" feature="entityId" dataType="double" optype="continuous"/>
    </Output>

    <!-- We have three items in our input data -->
    <Item id="1" value="Cracker"/>
    <Item id="2" value="Coke"/>
    <Item id="3" value="Water"/>

    <!-- and two frequent itemsets with a single item -->
    <Itemset id="1" support="1.0" numberOfItems="1">
      <ItemRef itemRef="1"/>
    </Itemset>
    <Itemset id="2" support="1.0" numberOfItems="1">
      <ItemRef itemRef="3"/>
    </Itemset>

    <!-- and one frequent itemset with two items. -->
    <Itemset id="3" support="1.0" numberOfItems="2">
      <ItemRef itemRef="1"/>
      <ItemRef itemRef="3"/>
    </Itemset>

    <!-- Two rules satisfy the requirements -->
    <AssociationRule support="1.0" confidence="1.0" antecedent="1" consequent="2"/>

    <AssociationRule support="1.0" confidence="1.0" antecedent="2" consequent="1"/>

  </AssociationModel>

</PMML>
```

Figure 8: Example of an association model in ***modified*** PMML format (derived from http://dmg.org/pmml/v4-4-1/AssociationRules.html)

Since the generated files are very heavy and inefficient to manage by graphical visualization libraries, the FIS/ARD tool has established the following procedure to be implemented in the future during productising:

- Results will be formatted using the PMML format, where some meta-information about the experiment should be added.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

- This file will be converted into a simpler and more dynamic format accepted by most common visualization libraries. For that, we have provided an intermediate structure using JSON format (described in D6.5) that enables a more efficient management of information.
- Using the intermediate structure, multiple visualizations for association rules can be made (see examples of visualizations in D6.5).

To appropriately support the model resulting from the FIS/ARD tool, it would be necessary to improve the PMML standard for association rules by including information about the employed dataset, the date and time stamp for the experimentation done, as well as new assessment measures that could be employed to measure the strength of the association (e.g., the certainty factor). The justification of using JSON instead of XML is that the former is much simpler, less verbose, while keeping the flexibility and self-describing properties. More details on this topic are provided in section 9.3

The CTSAE tool currently uses the Hugin flat net file format due to its larger base of software capable of reading it. Note that implementation of a reader and an export function for XMLBIF is trivial and can be realised during the productising phase. PMML does not support the generic structure Bayesian model representation. In the current situation of PMML and ONNX, attempting to represent Bayesian Models would mean using the functional modelling approach of PFA or ONNX. Implementing the required execution engine seems likely to be extremely expensive, if possible at all. Current implementations of the Bayesian reasoning using the Junction Tree algorithm (as supported in the CTSAE tool) are highly optimised and it is not clear if such level of optimisation is reachable within the framework of operators of PFA or ONNX. The recommendation is therefore to monitor evolutions and wait for stabilisation of the options before engaging in a costly implementation.

For DNN: The Keras (tensorflow) and PyTorch frameworks were chosen for different applications, due to suitability with the task and existing expertise. ONNX provide automatic conversion from both frameworks, making future standardisation simple.

11.4. Plans for further actions

Overall, it is clear from the discussion in section 11.2 and 11.3 that the maturity of existing formats is relatively low and far from reaching the status of a standard. The main challenges are:

- The rapid progresses and evolution in the field of AI resulting frequently in new models that dominates a specific field for a short period of time.
- The heterogeneity of the AI models, with different models suitable for different applications and very subtle implementation differences having significant influence on the performance.

In fact, at this stage, it is not even certain that initiative aiming at format supporting ***any*** model can be achieved. The added value of using current initiatives such as PMML and ONNX seems low for most tools developed in COPKIT, in particular since, even if the formatted representation of the model could be achieved, there may not be other tools implementing the needed computation features rendering the potential interoperability useless. Still components have taken preliminary measures to ensure that the implementation of standards is facilitated in the productising phase.

Regarding the approach used in PFA and ONNX for “traditional” is equivalent to a re-implementation and therefore is (i) very costly and (ii) very uncertain if it can be achieved. Given the high risk at this level of maturity, the COPKIT team does **not** recommend engaging in such activity for the time being.

The use of ONNX for DNN based tools (ESTF and MoRec mostly) is one significant exception, as DNN frameworks cover relatively similar implementation, given potential value to the unified format with the caveat that it may be superseded relatively fast. The implementation in COPKIT supports conversion of DNN models to ONNX and is therefore ready to implement this step in the productising phase if judged suitable.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

In summary, the recommendation of the COPKIT team is to carefully check the added value before attempting to implement one of these standards. Still a few actions are recommended:

- A few gaps are identified in PMML. In particular, support for the traceability of the model (by providing metadata) is missing. The COPKIT team will reach out the PMML development body to investigate the possibility of implementing some changes.
- Investigation will be carried out regarding possible inclusion of generic Bayesian Network in PMML.
- Monitor the evolutions in PMML and ONNX during the execution of the COPKIT project and, more importantly, during the productising step and act accordingly if a format emerges as stable and with added value. The action would then be to implement export functions in the component (FIS/ARD, CTSAE and ESTF in particular).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

12. HMI for LEA analysts and visual analytics

12.1. Introduction

The COPKIT project proposes as part of its activities a "Prototype HMI for analysts for usage of multi-level intelligence" (Task T3.1, Deliverable D3.6). The goal of such HMI is to propose concepts and implementations supporting the new workflows resulting from the application of the EW/EA methodology, in particular the incorporation of strategic knowledge in a case / operational analysis and vice versa. It also proposes a set of visualisations for the data types relevant for the COPKIT development.

The potential advantages of unification of HMIs have been recognised for since the massive deployment of IT tools (late eighties, early nineties): it can facilitate the in-boarding of users and diminish the ramp-up and learning time when using new software and its HMI. The approaches are pragmatic and heuristic-based, trying to balance ergonomics and recognition. The approaches resulted in the notion of "look and feel" and recognised practices. However, in the last 10 years, the focus has shifted towards adaptation: the development of HMI should be tuned to the actual application, resulting in very different approaches (for instance for e-commerce or industrial system). The following sections attempt to provide an overview of the situation, and provide details on the COPKIT vision on the topic.

12.2. Overview of existing work in the area of HMI standardisation

The design of HMI encompasses different aspects that need to be distinguished to analyse the existing standardisation initiatives.

As first aspect, one observes that a number of very low-level aspects of ergonomics are defined in the standard ISO9241 (for instance font size, or touch interfaces). Such aspects are nowadays part of the standard package for industrial designers.

The second aspect is the notion of "look and feel". The look and feel aims at facilitating recognition for users and helps users finding similar functions across different products. It should be noted that the "look and feel" is largely governed by the desire of developing a "brand" of the vendors: users recognise a certain look, get accustomed to certain ways of presenting functions and, ultimately tend to favour products with the same "look and feel" providing an advantage for the vendor of said products. This behaviour is not necessarily traceable to an increased efficiency of the look-and-feel and design used. The field is seen as critical in the competition between vendors blocking the way toward standardisation. Also, efficiency benefits are limited to the use of "common functions", in general support functions such as file management, setting management etc. (core functions of a given product will be unique). New "look-and-feel" will emerge rapidly when new vendors of tools suites gain market shares.

For some very specific applications, the design of the user interface has been standardised. One can name for instance ISO 11064 "Ergonomic design of control centre"[39] which provide guidelines to build an HMI for industrial control rooms. The type of application is quite different than the one targeted in COPKIT: the goal is to ensure minimal errors rate for a critical process, in application in which the user has limited autonomy of action (strong processes are in place). In COPKIT, the analysts are expected to have a lot of freedom and to resolve unforeseen challenges. Similarly, proposals of unified design have been made for specific operative tasks that have to be executed relatively quickly (e.g., have low value and require low autonomy), for instance for the configuration of similar components provided by different vendors such as network routers or computer BIOSes.

For consumer applications, designs tend to converge due to platform dominance. With Android, Google proposed "Material Design"[40] a design language aiming at unifying the user experience of services across

[39] See https://www.iso.org/standard/19042.html. It is developed by the working group ISO/TC 159/SC 4 "Ergonomics of human system interactions" as is also the ISO 9241-X serie.
[40] See https://material.io/design/introduction, first introduction in 2014.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

platforms. Due to the market share of Google, the adoption by designers and developers (due to the facilitating toolkit) resulted in a convergence of the design for web pages and applications. However, in the last few years, some other important players are attempting to build their own unique user experience resulting in more options[41]. Currently, websites are also optimised for usage on touch displays. Overall, while some innovations provided are interesting, the type of applications is also far from the COPKIT application with goals such as "attention retention", "1 click shopping" etc. Following the approach of Google, large software vendors who provide HMIs for their line of product started implementing internally standard widgets and usage[42]. The goal is to streamline the development process and such initiative will not particularly contribute to increase efficiency from the user perspective.

In the last decennia, the paradigm for the design of professional software tools is also evolving towards user-centred design and the idea that the specific context of the task is the starting point of the design. When such approach is suitable, then standardisation is more likely to be undesirable.

Finally, it should be noted that the study of human cognition aspects of visual analytics is relatively under-developed. While much work has been done regarding control tasks, reaction tasks and monitoring tasks, few studies have tried to analyse the way visualisations are interpreted and patterns are recognised in complex visualisations. This may be related to the fact that tasks such as "recognising visual patterns" or "finding anomalies" are difficult to define in essence and therefore building benchmarking tests seems quite challenging. It is also possible that strong inter-personal differences exist in the cognitive task of "seeing pattern" which would reduce the applicability of standardised visualisations for recognised problems.

12.3. Action taken in COPKIT

The implementation of the "Prototype HMI for analysts for usage of multi-level intelligence" (Deliverable D3.6) in the COPKIT project followed the methodology of user centred design as the complexity of the analysts' tasks called for it. Considering the context of the tasks and the associated workflows was part of the research activity. These aspects were therefore not standardised. However, industrial designers involved in the task applied the appropriate ergonomic recommendations (as in ISO9241) and generally accepted guidelines regarding the use widgets. Adaptations to specific look-and-feel are left for the productising phase.

The above mentioned HMI also proposed four types of innovative visualisations (for annotated texts, graph or relation network data, spatial temporal data and correlation between numerical indicators) for data types relevant for COPKIT and correlations. These visualisations were validated during the development by LEA analysts in the project.

12.4. Plans for further actions

Regarding the look-and-feel, it is expected that the COPKIT HMI would be one of the many tools used by an analyst in his/ her daily activities (for instance, different tools will be used for administrative task, reporting office tasks and communication tasks). Given the legacy systems in place in LEAs, the strategy is that the HMI should adapt to the dominating "look-and-feel" in the acquiring LEA, if necessary. Actions can better be carried out during the productising phase.

Regarding visual analytics, the research in human cognition for tasks such as recognising patterns seem not mature enough to envision the development towards a standard even in the long term. In addition, if

[41] See the discussion https://www.textmaster.com/blog/user-experience-standardisation-internationally/ (last accessed on Jan 15th, 2021)

[42] See the discussion https://uxdesign.cc/you-should-standardize-the-ui-patterns-and-components-in-your-pattern-library-97b9da87722c (last accessed on Jan 15th, 2021)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

one adopts the hypothesis that good design must be tailored to the context, standardisation may even be undesirable. No further action is expected.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

13. Ethical Aspects of the use of AI for Law Enforcement

13.1. Introduction

The development of AI technologies in recent years has been a major step forward in the IT sphere and it is expected that its impact will be just increasing in fast pace from now on. Its societal impact is already significant and the concerns in civil society are growing fast.

To this end, a regulation in terms of how this development is to continue in the future is necessary. This regulation should go beyond strictly legal requirements and cast a look towards ethics as an underlying foundation to technology emerging of the EU.

It is undoubtful that AI can and should contribute in a great manner to the work of the law enforcement agencies and the judicial authorities by implementing its abilities such as facial recognition technologies, automated number plate recognition, speaker identification, speech identification, lip-reading technologies, aural surveillance (i.e. gunshot detection algorithms), autonomous research and analysis of identified databases, forecasting (predictive policing and crime hotspot analytics), behaviour detection tools, autonomous tools to identify financial fraud and terrorist financing, social media monitoring (scraping and data harvesting for mining connections), international mobile subscriber identity (IMSI) catchers, and automated surveillance systems incorporating different detection capabilities (such as heartbeat detection and thermal cameras). However, the sensitive aspects of law enforcement work, the ethical challenges associated to it and its potential impacts call for even greater control and responsible usage than for other applications. The concerns in the civil society for application of AI to the field of Fighting Crime and terrorism are also significant.

The COPKIT project dedicates specific tasks (Task T2.4 focusing on the overview of applicable requirements in the use of AI for law enforcement and Task T3.4 for the evaluation process developed in COPKIT and its application to the COPKIT tools) to the Ethical, Legal and privacy challenges associated to the use of AI for Law Enforcement. The following sections provide an outline of existing recommendations and an overview of the actions taken in COPKIT. The interested reader should refer to Deliverable D2.4 (for an overview or relevant recommendations and approaches) and to Deliverable D3.4 (for a detailed view of evaluation actions carried out with respect to COPKIT developments).

13.2. Overview of explored / existing recommendations

An important step was taken recently at EU level to foster AI developments that are carried out in agreement with the norms and values supported by the EC. **The guidelines of the European Commission on a trustworthy AI** state that this type of technologies should be lawful, ethical, and robust.[43] Furthermore, 7 key requirements are outlined in the said guidelines that an AI based tool to be seen as trustworthy. These 7 conditions state that the AI has to be[44]:

- Overpowered by humans and be overseen by them as well,
- resilient and secure,
- ensuring privacy and data protection,
- transparent,
- non-discriminatory and fair,

[43] 'Ethics guidelines for trustworthy AI' (2019), European Commission <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> accessed 14 December 2020.
[44] Ibid.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

- sustainable and environmentally friendly, and
- responsible and accountable for their outcomes.

Taking into account the sensitive nature of combatting crime, the work on a dedicated report to outline the use of AI in the context of combatting crime has been initiated. In this section, the version of the said report as per 22.12.2020 is considered. **The draft report on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters** states that the ultimate purpose of the developing AI should be to increase the human well-being. However, this should be done in respect with the Union values: human dignity, freedom, democracy, equality, the rule of law, and human and fundamental rights.[45] Furthermore, this draft report confirms the 7 requirements mentioned above should be respected. Another point that was precisely outlined in this draft version of the official document is the fact that the Union and the Member states have a great responsibility to enact such policies for the development and usage of AI that the security and the fundamental rights are not put at risk.[46]

While the (potential) great contribution of AI to the work of LEAs and judicial authorities is acknowledged, **the Draft report on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters** underlines that the accuracy of such tools can be quite variable, hence, the legal responsibility and framework is essential in this field, especially in cases with potentially harmful effects of AI systems.[47] AI presents numerous possibilities for the law enforcement agencies and the judicial authorities that eases their work and strongly strengthens their effectiveness. However, these systems come up with significant risks to the protection of privacy and personal data as well as the right to a fair trial and the correlating presumption of innocence, and the risk of discrimination among other fundamental rights. Therefore, since the processing of a large amount of data is the corner stone of the AI's success, a full compliance with the Union legal framework for fundamental rights is crucial. Taking this into consideration, **the draft report on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters** stated that the minimum requirements for an AI used by law enforcement agencies or judicial authorities are that it should be safe, secure, fit for the planned purposes and respect the principles of fairness, accountability, transparency and explainability, with their deployment subject to a strict necessity and proportionality test. Furthermore, it is required that the final decision in all cases must be made by a human who can be hold responsible for the outcome.[48]

Another vital aspect discussed in the mentioned draft report is the fact that mass surveillance by AI should be prevented and any such application to be banned. Furthermore, the security and safeguards need to be strong enough to not allow "*potentially catastrophic consequences of malicious attacks on the AI systems*". This includes regular testing and evaluation of the systems in order to be diagnosed whether there are any errors or potential risks present.[49]

13.3. Action taken in COPKIT

A full description of the COPKIT actions with respect to supporting the guidelines and recommendation is described in Deliverable D3.4; this section only provides a brief overview.

As part of the activities carried out in Task T3.4, a process of evaluation of tools was designed, applying both to the development phase (e.g., with technical partners during the project) and the usage phase (e.g.,

[45] 'Draft report on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters' (2020), 2020/2016(INI). European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Rapporteur: Tudor Ciuhodaru.
[46] Ibid.
[47] 'Ibid.
[48] 'Ibid.
[49] Ibid.

with LEA representative during acquisition and usage). This process was built using the existing laws, guidelines and recommendations mentioned in section 13.2 as inputs to design the evaluation points. The process was further applied by the COPKIT ELP Team to evaluate the tools developed in COPKIT during the development phase and further with LEA representative as a try-out before potential acquisition. In this way, the laws and accepted recommendation were applied to the COPKIT project mediated by an evaluation process.

13.4. Plans for further actions

AI based technologies supporting the structuring of unstructured information are particularly useful for LEAs. Among these, analysis of audio and image / video material is expanding quickly due to the cost of manual analysis but are also particularly challenging due to accuracy challenges. The **second Interpol-UNICRI report on artificial intelligence for law enforcement - 'Toward Responsible AI Innovation'** addresses these issues.

The AI technologies used for these applications still suffer from practical challenges:

- Audio quality degradation, overlapping human voices, voice manipulation, deep-fake voice applications and many others.
- For image or video overlapping objects (usage of subjects that cover a person's face such as glasses, accessories and masks) are also challenging. More recently, deep-fake videos have reached a quality sufficient to make them an issue as well.

All these factors should be considered, and a strict evaluation of performance and bias should be performed. Further, processes should be made for the mitigation of potential errors, especially when in use by law enforcement agencies and judicial authorities.[50]

Despite the fact that the implementation of AI technologies presents some risks to the fundamental rights and freedoms of the citizens, its development has provided the law enforcement agencies with numerous instruments and tools to fight crime in a much more effective manner. A proof of such innovation, outlined in the Interpol-UNICRI recent report, is the 'improved strategies of dynamic matching in resource supply and demand' that have significantly improved the timely arrival of help asked by emergency calls. Crime mapping has been an objective for the law enforcement agencies since the 1990s, however, machine learning algorithms have 'strengthen the connection between alert, response and reaction'.[51] As the Interpol-UNICRI report stated, it is crucial that the law enforcement agencies take part in the development and improvement of such AI systems as they are able to provide valuable feedback that will improve the effectiveness of the designed tools and the implemented approaches.[52]

Overall, it is expected that, the Ethical, Legal and Privacy community will continue to develop their framework of guidelines and recommendations. The current challenge is to develop practical implementation of the evaluation process. The COPKIT project developed and implemented such a state-of-art evaluation process. The ELP Team of the COPKIT project is very active in the community (among others through the Community of User event organised by DG-HOME). The team is disseminated the COPKIT experience within the community and will continue to do so after the project. The COPKIT Team estimates that the community built around EC funded project in the domain of FCT is appropriate for such dissemination as it is multi-disciplinary in nature (representative of LEAs, of technology providers and of

[50] 'Toward Responsible AI Innovation' (2020), Second Interpol-UNICRI Report on Artificial Intelligence for Law Enforcement.
[51] Ibid.
[52] Ibid.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

the ELP research community). It is expected that the initiative will allow the community to build consensual recommendations including implementable and actionable of evaluation plans on the medium long term.

14. Conclusion and next steps

This document presents the analysis of Standardisation and Certification opportunities for the product of the COPKIT Project. The analysis is not limited to technical aspects but encompasses processes, methodology training etc. Furthermore, the analysis goes beyond official standards and certification and considers de-facto standards best practices and formats with a significant level of adoption. The analysis is based on the responses of COPKIT LEA partners to a questionnaire and on the monitoring of relevant standards and initiative during the development of the technical components.

The analysis of the responses of LEA showed several interesting insights:

- LEAs have developed internally best practices over which they are globally satisfied.
- A certain amount of flexibility is seen as necessary for data formats to adapt to different usage.
- The exchange of data is still sometimes challenging, partly due to communication tools inadequate for (large) data exchange (not in the scope of COPKIT) but also due to different data format requiring the development of glue code.
- Organising advanced or specialty training is challenging (although this is not a challenge related to the lack of standards).

The monitoring of technical development led to the identification of several relevant areas for which a deeper study of existing standards and potential gaps was carried out. Not surprisingly, since many of the tools developed in COPKIT make use of Artificial Intelligence and Machine learning, the analysis showed that a number of areas were relatively immature technically with rapidly changing dominant approaches. In this situation, it is often premature to envision actions leading to standards (or recommendation for best practices) within the timeframe of the project. Still, several standards (or dominating formats) could be identified and have been implemented by the COPKIT tools. For most types of data, several formats are found to co-exist. The COPKIT strategy has been to choose one that balances popularity and practicality, and provide a flexible implementation to be able to easily adapt during the productising phase.

Table 2 provides an overview of the action envisioned for the “promising” areas as defined in section 4.1.

Area description	Conclusion and proposed actions	Timeframe
Web archiving formats	Monitor standardisation initiatives (no short term expectations as standard would need to overcome contradicting requirements)	Post project: Productising phase
Textual data, metadata and annotations	Encourage adoption of used standards with project partners (in particular usage by other tools)	Within project. Action already started.
Knowledge representation	Disseminate COPKIT advances in the LEA community with the goal of LEA taking ownerships of the standardisation issue	During project dissemination efforts
	Partner with other EC funded projects to build a community of researchers active in Security Research community to coordinate research actions	During project. Action already started.
	Partner with other EC Funded projects and actors of the Community of Users to raise awareness aiming at inclusion in the research agenda	During project. Action already started.
Criminal domain semantics	Assess the opportunity for special dissemination actions of firearm taxonomy towards LEAs	During project. Handled by responsible partners (UGR, Guardia Civil) and project coordination
	Partner with other EC funded projects to increase LEA awareness of the fragmented developments and the need for LEA to take ownership of the issue	During project. Handled with the two previous actions
	Disseminate the analysis and found gaps in the research community with the possible goal of inclusion in the research agenda.	During project. Handled with the two previous actions
Data Exchange Formats for Graphs	Monitor the take-up of (recent and not widely used) standard (W3C JSON-LD) and reassess value of implementation.	Post project: Productising phase
Spatial Temporal data	Monitor evolution of GIS with respect to introduction of semantics Spatial Temporal data that could make some formats more attractive	Post project: Productising phase
Spatial Temporal data (AI) Models	Unification w.r.t to GIS for LEA maybe desirable but is outside of the project reach. Raise awareness about the challenge in the LEA community	During project
	Monitor the take-up of PMML and ONNX initiative for possible implementation	Post project: Productising phase
(AI) Models HMI for LEA analysts	Gaps identified within PMML. Finalise analysis and launch a request for change if appropriate	During project (initiation), will likely continue after.
	Further analysis for proposal of introduction of generic BN in PMML. Finalise analysis and launch a request for change if appropriate	During project (initiation), will likely continue after.
	None. Maturity is too low.	NA
Ethical aspects of AI usage for Law Enforcement	Disseminate COPKIT results and contribute to advances of the community although a standard is not within reach	During and after project

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Table 2: Identified areas with a summary of proposed actions with regard to Standardisation and Certification

Beyond the format, agreed-upon semantics are required to be able to exchange data, especially for data such as (criminal) intelligence and knowledge. The analysis carried out in section 7 and 8 shows that a gap exists on this matter, both for general knowledge and for subject-matter knowledge. The COPKIT project team estimates that, while the challenge cannot be solve in the short term, actions can be undertaken to improve the situation on the long term. The COPKIT project team considers that appropriate actions involving collaboration between H2020 projects are a good way forward and is carrying out (and will continue to do so) actions aiming at establishing a community sharing the specific interest in this topic. The main actions in the action plan are tackling that area.

The answers provided by LEA respondents provided on the side, the insight that the setup of advanced training for senior analysts or trainings on specialists or emerging topics is challenging. This information will be exploited by the COPKIT project as part of its exploitation effort, possibly in the context of defining the mission of the COPKIT Live Lab (COPLAB)

Overall, while the advantages of using standards and certification are clear, a good balance between flexibility and standardisation is critical for the type of tools developed in the COPKIT project. In addition, careful assessment of the maturity of the various areas and in particular the risk of changes rendering specific implementation obsolete should be performed before engaging in implementation.

Annex I: Questionnaire sent to LEA to collect information on experienced challenges related to Standardisation and certification

This section presents the questionnaire that was distributed to COPKIT LEA partners to obtain insights on the situation in their organisation with respect to standardisation (in the areas relevant for the COPKIT project) and on the challenge that they encountered with respect to existing standard or best practices or the lack thereof. In some case, the questionnaire was followed with by interview with the contact point for clarifications.

Introduction (Optional)

Matter	Answers
LEA organization (COPKIT Acronym)	
Contact point: (optional)	
Do you allow the COPKIT technical team to contact you in case of follow-up questions?	

Standards and practices for (intelligence) analysis process

Appropriate processes for intelligence analysis have been proposed in the past and are applied in most LEA organizations. A typical example would formalize the steps of: collection plan definition, collection, collation, evaluation and dissemination (see for example: https://www.app.college.police.uk/app-content/intelligence-management/analysis/getting-started/).

(Intelligence) analysis process	
Does your organisation use a standard to define the process of intelligence and analysis work? (Yes/No/Information not available)	
If yes:	
If yes, which one?	
If yes, do you find it satisfactory (grade from 1 not at all to 5 very satisfactory)	
If yes, indicate why you are satisfied or not with it	
If no:	
If not, how much does it affect (negatively) your ability to carry out analysis work? (grade from 1 not at all to 5 very problematic + explanation)	
Additional details (free)	

Standards and practices for (intelligence) analysis techniques, methodologies and tools

A number of guidance and best practices regarding analysis techniques (for instance, hypothesis generation, network analysis), methodologies (for instance SWOT analysis, Force field analysis) and tools (see “Analyst toolbox”, https://it.ojp.gov/documents/analyst_toolbox.pdf) have been proposed in the past and are applied in most LEA organization.

Analysis Techniques and tools	
Does your organisation use a standard to define the analysis techniques and tools? (Yes/No/ Information not available)	
If yes:	
If yes, which one?	

If yes, do you find it satisfactory (grade from 1 not at all to 5 very satisfactory)	
If yes, indicate why you are satisfied or not with it	
If no:	
If not, how much does it affect (negatively) your ability to carry out analysis work? (grade from 1 not at all to 5 very problematic + explanation)	
Additional details	

Standards and practices for training of analysts

A number of guidance curricula for analysts have been proposed in the past and are applied in most LEA organization. For instance, in the UK: see here https://www.college.police.uk/What-we-do/Learning/Curriculum/Intelligence/Foundation_analysis/Pages/Analyst-Foundation-Module.aspx or, with a detailed syllabus for the US here: https://it.ojp.gov/documents/d/minimum%20criminal%20intelligence%20training%20standards.pdf)

Training curricula For analysts	
Does your organisation use standard training curricula for analysts? (Yes/No/ Information not available)	
If yes:	
If yes, which one?	
If yes, do you find it satisfactory (grade from 1 not at all to 5 very satisfactory)	
If yes, indicate why you are satisfied or not with it	
If no:	
If not, how much does it affect (negatively) your ability to carry out analysis work? (grade from 1 not at all to 5 very problematic + explanation)	
Additional details	

Standards and practices for data format, data exchange and input / output of analysis tools

LEA analysts use a variety of data types (ranging from raw unstructured data to numerical structured data) and tools to process them. The format in which these data are organised is an important condition of the capacity to use these data and exchange them. While standardised format for some type of data exists, there does not seem to be unified, generally agreed upon way to represent any data. Therefore in the question below, data are divided in categories corresponding to a “nature” of the information. For each category, please indicate, if your organisation implements a standardised approach or uses existing standard and which one, if so, how satisfying it is from your point of view. If your organisation does not use such standardisation, please indicate how much this absence negatively affects your analytical work and why. To minimise the effort in answering, categories presented are limited to the ones relevant for the current COPKIT tools.

Notes:

- Indicate NA for Not applicable or INA if information is not available.
- If some categories seem to be missing use the lines at the bottom to add.

Category of data	**Standard use if any**	**Satisfaction Grade (1-Very high, 5 not at all)**	**Negative impacts Grade (1-Very high, 5 not at all)**	**Rational and comments**
html /website data				
Annotation for text				
Metadata for text				

Formalised criminal knowledge representation	
Knowledge base models	
Spatial temporal data	
Geographic information	
“graph” data (relation networks)	
AI models representation (inc. but not limited to NLP)	
Others:	

Challenges encountered related to standardisation

This section contains open questions so that you can express your comments and your priorities on matters that have not been addressed so far. Think of gaps, priorities in terms of standardisation, or other issues such as adoption by industry, inadequacy etc…

Challenges, priority and comments related to standardisation
What other challenges related to standardisation and certification do you experience in fields relevant for the COPKIT project? Indicate how high their impact is (grade from 1 not at all to 5 very problematic).
Do you have suggestions for standardisation in COPKIT? Priority? Actions?
Are there any challenges related to standardisation and certification in fields outside of the scope COPKIT project that you would like to mention?

Optional: Standards and practices for IT system and digital tools

Note: the question is about the IT environment in your organisation. It may be difficult to answer for an analyst. Still this information is useful for the COPKIT project (for the tool deployment strategy among other). We would appreciate if you could address this question to the relevant colleague (probably IT department)

LEA organisations use numerous IT systems and digital tools. In many cases, they use off-the-shelf products, frequently already implementing state of the art technical standards. In this section, the different categories of components of IT systems relevant for the work of analysts are mentioned.

For each category, please indicate:

- if your organisation implements a standardised approach or use existing standard and which one and, if so, how satisfying it is from your point of view. This question may seem ambiguous: if your organisation uses several well defined (commercial) implementation, please indicate yes and cite a few. For instance, if your organisation uses database made by Oracle and SAP, answer yes and name both. Indicate NA for Not applicable or INA if information is not available.
- If your organisation does not use such standardisation, please indicate how much this absence negatively affects your analytical work and why.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved

Notes:

- Indicate NA for Not applicable or INA if information is not available.
- If some categories seem to be missing use the lines at the bottom to add.
- Standards for data exchange format are treated in the previous section

Category of system	**Standard use if any**	**Satisfaction Grade (1-Very high, 5 not at all)**	**Negative impacts Grade (1-Very high, 5 not at all)**	**Rational and comments**
Communication intra agency				
Communication inter agency				
Hardware				
Operating system				
System architecture and communication middleware (between tools)				
Programming language (inc. web-development and framework)				
Database software				
Database language and client software				
Communication security				
Others:				

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the COPKIT project partners. © 2018 – All rights reserved