INSPECTr Project

Intelligence Network & Secure Platform for Evidence Correlation and Transfer

Quarterly Newsletter: Third Edition

Edition: September 2021

Dear Colleagues,

Welcome to the INSPECTr project newsletter, a guide to our latest work and news.

INSPECTr Principal Objectives Brief Summary

Intelligence
Network &
Secure
Platform for
Evidence
Correlation and
Transfe**r**

To develop a shared intelligence platform and a novel process for gathering, analysing, prioritising, and presenting key data to help in the prediction, detection, and management of crime in support of multiple agencies at local, national, and international level. This data will originate from the outputs of free and commercial digital forensic tools complemented by online resource gathering. The final developed platform will be freely available to all Law Enforcement Agencies (LEAs).

INSPECTr Newsletter Third Edition

In this, our third edition, we will provide updates on our last quarter activities, information about meetings and events attended, our upcoming events, recent dissemination activities and blogs on the subject of **Standardisation in CASE** provided by our Consortium Partners, CNR and VLTN.

Standardisation in CASE in the INSPECTr Project

Explanation of CASE Language and Reasons for its Adoption for the INSPECTr Platform - Blog provided by INSPECTr partner Panos Protopapas (VLTN)

With the CASE language having such a central role in the platform, handling of the CASE format becomes crucial. To accommodate the different data needs on the platform, the CASE language is used in three different forms by the platform and depending on the form stored in three different storage engines.

It exists as a binary json-ld file, that depending on the amount of investigative information it represents, could be of several hundred megabytes or possibly larger (e.g., if describing the file contents of a large disk drive). It is the output of the platform's parsers, the primary data input of the platform, utilised when for example an LEA user wishes to import in the platform evidential material stored elsewhere, parse reports produced by other tools in their organisation, or directly process evidential material via one of the platform's "wrappers" carrying out this functionality.

The binary form is essentially a collection of interlinked nodes forming a graph. For example, information about a (mobile) phone call between two individuals could loosely speaking be represented via a graph by three nodes; the phone call node holding information about the call itself (duration, application used, etc.), linked with two person nodes, each providing information about the individuals (phonenumber, name, etc.). Moreover, each person node could also be linked with other nodes, representing other calls that the individual took part in, messages they sent or received, etc. These binary files are stored in INSPECTr's HDFS storage system, along with all other binary files (photographs, videos, etc.), which are also described by CASE files (e.g., a photo node can list the EXIF data and provide the HDFS location of the actual photograph).

With the binary form being the output of all tools on the platform and hence the common denominator of all investigation-related data, it became evident that handling CASE data in a clear and efficient way was of crucial importance to the platform. To achieve this, binary files are "flattened", or in other words transformed from a large json-ld file describing all interlinked nodes (a large hash table) into a collection (an array) of much smaller hash tables, each describing a single flat node. The graph structure of the binary CASE form is retained via the inclusion of meta-information on each flat node. During this process, each flat node is individually stored in INSPECTr's Elasticsearch storage system. This enables tools to directly and efficiently retrieve the information needed, instead of first parsing a perhaps very large binary CASE file in order to access this information. For example, the natural language processing tool directly accessing the required sms messages (a few hundred kilobytes) without first parsing the binary CASE file describing the mobile phone's drive where these messages were retrieved from (hundreds of megabytes, possibly larger).

Finally, in order for the structured and interlinked representation of cyber-investigation information, as well as the evolution of an investigation to be easily studied and manipulated by LEAs, CASE data is also converted into knowledge graph form, and stored in INSPECTr's Neo4j storage service. This allows CASE data to be queried in a more "investigative" way; for example, a query to "retrieve all individuals having contacted suspect X via sms during period Y" can be created using the CASE language and run on Neo4j.

Handling of Standardised Evidence (CASE) by the Platform - Blog provided by INSPECTr partner Fabrizio Turchi (CNR)

One of the call's requirements was to adopt a common format for data homogenisation, data discovery (linked cases) and data exchange. The INSPECTr project opted for the open-source Cyber-investigation Analysis Standard Expression (CASE, https://caseontology.org) language, a community-developed ontology designed to serve as a standard for interchange, interoperability, and analysis of investigative information in a broad range of cyber-investigation domains, including digital forensic science, incident response, counter-terrorism, criminal justice, forensic intelligence, and situational awareness. The CASE Community is a consortium of for-profit, academic, government and law enforcement, and non-profit organisations that have created a new specification for the exchange of cyber investigation data between tools.

CASE provides a structured specification for representing information that are analysed and exchanged during investigations involving digital evidence. To perform digital investigations effectively, there is a pressing need to harmonise how information relevant to cyber-investigations is represented and exchanged. CASE enables the merge of information from different data sources and forensic tool outputs to allow more comprehensive and cohesive analysis. The main benefits of using CASE are:

- fostering interoperability: to enable the exchange of cyber-investigation information between tools, organisations, and countries. For example, standardising how cyber-information is represented addresses the current problem of investigators receiving the same kind of information from different sources in a variety of formats;
- establishing authenticity and trustworthiness: based on the clear representation of the Chain of Evidence (provenance) and the Chain of Custody. A fundamental requirement in digital forensics is to maintain information about evidence provenance while it is exchanged and processed;
- enabling more advanced and comprehensive correlation and analysis. In addition to searching for specific keywords or characteristics within a single case or across multiple cases, having a structured representation of cyber-investigation information allows more sophisticated processing such as data mining, or NLP techniques. This can help, for instance, to overcome linkage blindness that is the failure to recognise a pattern that links one crime to another, such as crimes committed by the same offender in different jurisdictions;
- helping in dual/multiple tools validation or results in order to evaluate their completeness and correctness/accuracy;
- automating normalisation and combination of differing data sources to facilitate analysis and exploration of investigative questions (who, when, how long, where).

An investigation generally involves many different tools and data sources, creating separate storerooms of information. Manually pulling together information from these various data sources and tools is time consuming, and error prone.

Tools that support CASE can extract and ingest data, along with their context, in a standard format that can be automatically combined into a unified collection to strengthen correlation and analysis. This offers new opportunities for searching, contextual analysis, pattern recognition, machine learning, and visualisation.

Project Activities and Events between July 2021 – September 2021

- INSPECTr Monthly Project Meetings
- INSPECTr Weekly Technical Meetings
- INSPECTr LSG Monthly Meetings
- Ethics Work Package Monthly Meetings

INSPECTr Monthly Project Meetings

These continue to be held monthly where an overview of the activities undertaken in each work package are reported on by work package leaders to the Consortium.

INSPECTr Weekly Technical Meetings

Weekly technical meetings are held in order to support the finer detail of the project development, give close attention to particular issues, and bring about resolution so that the project continues to develop on track between the monthly meetings.

INSPECTr Law Enforcement Steering Group (LSG) Monthly Meetings

These meetings provide a forum for collaboration between our law enforcement, technical, and ethics partners. It is of primary importance that that technical features under development in the INSPECTr platform are developed in a way that is relevant to and mirror the LEA investigative workflow. An ethical oversight at these meetings ensures that privacy and data protection requirements from applicable legal frameworks are observed, developed, and embedded.

INSPECTr Network of Living LEA Living Labs

During the last quarter there has been a great deal of increased engagement between the INSPECTr technical and Law Enforcement Steering Group partners in order to finalise Use Case mocked but realistic evidence, consolidate that evidence and prepare it for experimentation and testing of the INSPECTr platform's functional and non-functional characteristics. More information on developments in the INSPECTr Living Labs will follow in the first quarter of 2022.

Ethics Work Package Monthly Meetings

These meetings continue and are held to reinforce the project's Ethics-by-Design approach and allow time for deeper consideration and exploration of ethical issues that arise.

INSPECTr Publications

The proceedings of the 4th International Conference on Intelligent Technologies and Applications (INTAP 2021)

Title: **Iterative Learning for Semi-automatic Annotation Using User Feedback**

Authors:
Meryem Guemimi, Daniel Camara
Center for Data Science, Judiciary Pôle of the French Gendarmerie, Pontoise, France.

Ray Genoe
Centre for Cybersecurity and Cybercrime Investigation, University College Dublin, Dublin, Ireland.

Abstract:
With the advent of state-of-the-art models based on Neural Networks, the need for vast corpora of accurately labelled data has become fundamental. However, building such datasets is a very resource-consuming task that additionally requires domain expertise.

The present work seeks to alleviate this limitation by proposing an interactive semi-automatic annotation tool using an incremental learning approach to reduce human effort. The automatic models used to assist the annotation are incrementally improved based on user corrections to better annotate the next data.

To demonstrate the effectiveness of the proposed method, we build a dataset with named entities and relations between them related to the crime field with the help of the tool. Analysis results show that annotation effort is considerably reduced while still maintaining the annotation quality compared to fully manual labelling.

Further Opportunities for INSPECTr Dissemination and Cross-Project Learning and Collaboration

COPKIT FINAL EVENT - September 16th, 2021
Attended and reported on by INSPECTr Partner Fabrizio Turchi (CNR)

Our partner from INSPECTr partner CNR, Fabrizio Turchi, attended the Copkit final event and was one of the round table delegates attending the afternoon session: A Shared Journey - Stories from Fellow Travellers.

Fabrizio Turchi (CNR) made a short introduction about the INSPECTr project. One of the INSPECTr call's requirements was to adopt a common format for data homogenisation, data discovery (linked cases) and data exchange and Fabrizio explained how the INSPECTr project has opted for the open-source CASE, a community-developed ontology designed to serve as a standard for interchange, interoperability, and analysis of investigative information in a broad range of cyber-investigation domains.

For the round table discussion Fabrizio highlighted three important features to bear in mind during the development and the application of the Standard/Knowledge Representation model:
i) Sustainability: it's crucial to guarantee the maintenance and the updating of the Knowledge Representation model (standard) in the long term, to keep pace with the evolving needs of the technology.
ii) Appropriateness/Pertinence: the capacity to represent as much as possible the data/metadata of the Subjects and Objects involved in the domain of applicability.
iii) Feasibility: to implement the model/standard to a real environment. Within the INSPECTr project it has been of utter importance to take this step to address technical issues but also to improve the model (ontology), whilst in previous projects the task was limited to the development of Proof of Concept applications where there were less constraints to respect (i.e. ontology consistency).

Another interesting and relevant point raised by other panelists highlighted the importance of involving the Tech Giants (Amazon, Apple, Google, Facebook, and Microsoft) in the development of a standard as without their approval the standard is unlikely to succeed. However, hurdles would be present of imposing that standard upon them or suggesting its adoption. Also, by relying on Tech Giants, it would be difficult to develop an open source solution suitable to all involved stakeholders.

Conferences, Workshops and Future Events

INSPECTr Consortium Attendance at Conferences and Workshops ....

- INSPECTr Project Mid-Term Review held on Friday 2nd July 2021.
- Copkit Final Event held on Friday 16th September 2021.

Forthcoming Events ....

- The INSPECTr Consortium is currently preparing submissions for the Autumn/Winter events of:
- Octopus Lightning Talks 17-18 November 2021, hosted by CEPOL.

- FEF 2021 Conference (Forensic Experts Forum – Online Conference) 22-26, November 2021, hosted by Europol.

Closing

We look forward to updating you further in January 2022 with our fourth edition of the INSPECTr Newsletter. In the interim, communications from our readers are welcome and if you wish to contact us or subscribe to our Newsletter you can e-mail us directly at inspectr@ucd.ie. Further information and updates can also be found on our project website https://inspectr-project.eu/.