Processing heterogeneous data sources for risk management by knowledge graph databases: the case of BAG-INTEL research project.

Bartolome Ortiz-Viso∗†‡, Karel Gutierrez-Batista∗†, M. Dolores Ruiz ∗† and Maria J. Martin-Bautista∗†

∗ Research Centre for Information and Communications Technologies (CITIC-UGR), University of Granada, Granada 18014, Spain

† Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18014, Spain

‡ Corresponding author bortiz@ugr.es

Abstract

Knowledge graphs stand out for their ability to encode and represent knowledge in various fields. With the latest advances in information processing and machine learning, this data structure can increasingly be incorporated more efficiently into reasoning and knowledge acquisition processes. This has made them a valuable and versatile tool capable of answering questions in many domains. In this article, we present these structures, their concepts, and applications and offer an initial approach to their use in airport security within the framework of the BAGINTEL project.

1 Introduction

The BAG-INTEL project is a 3-year Horizon Europe Research and Innovation Action powered by a multidisciplinary consortium of 24 partners from 8 European countries, including industrial players, consultancy and advisory firms, universities and research organizations, ministries, customs and tax authorities, and civil authorities. The project aims to develop an AI-based solution to support customs control staff in identifying and locating suspicious luggage for further manual inspection. Risk indicators will be produced and merged once the baggage is scanned by X-ray CT scanners and additional knowledge is extracted from the database. If the customs control staff find these markers concerning, an AI system will use high-resolution cameras to track the suspicious luggage throughout the airport, allowing customs staff to locate it and then proceed with a manual inspection. BAG-INTEL will include a feedback loop where additional and real-time information will be processed to upgrade the risk markers process continuously.

The knowledge from these tasks will be stored in a graph-based database [9], where we will collect data from the various project tasks, trial results, and risk marker information. This type of database has gained popularity since its introduction by Google as a way to manage multiple sources of knowledge with multiple applications [1] (and specifically in the security field [6]). Knowledge graph databases offer numerous advantages in multidisciplinary projects, as they can represent complex relationships, utilize structured and unstructured data sources, absorb information from ontologies (for more realistic knowledge search and representation), and allow a constant evolution in their schema as a fixed table structure does not restrict them. At the same time, incorporating all this information presents a scientific and technical challenge that requires specific design and representation to make the database as functional and practical as possible, particularly in airport security.

This work presents the concept of knowledge graphs in section 1, highlighting why they are an essential tool and their advantages and drawbacks. In section 2, we will introduce the concepts and applications of knowledge graphs. Section 3 outlines our initial proposal for creating a representative knowledge base from which valuable insights can be derived for the project. Finally, we will discuss future directions for this and similar approaches in airport security.

2 Knowledge Graph Databases

In this section, we will present the most important concepts to understand the rest of the article and the relationships between them.

2.1 Graphs, Knowledge Graphs, and Knowledge Databases

Initially, before discussing graph structures or databases, we must introduce the concept of a graph, which is fundamental in mathematics and computer science. A graph is a structure consisting of two main elements: nodes (or vertices) and edges (or links). The nodes represent entities or objects, while the edges represent their relationships or connections. Those edges can also possess a set of different properties, but in the mathematical approach to them, they have only an assigned weight.

The rest of the concepts in this work inherit this basic definition but explore its potential when we establish a bilateral function between these structures and specific concepts. Starting with a knowledge graph, this structure organizes and represents information semantically and richly. The nodes correspond to entities or concepts, and the edges correspond to the relationships between these entities, where each relationship can have a label that describes its nature (for example, "is a friend of," "works at," etc.). This allows the information to be much more flexible and enriched, as the relationships between the concepts can be complex and multifaceted.

Although there is still no broad, extended definition of knowledge graph and knowledge graph database, we could establish some specific differences between them. A knowledge graph can be viewed as a graph when considering its structure [1]. Formal semantics can be viewed as a knowledge base when interpreting and making inferences about facts.

From this idea, graph knowledge databases store and manage data structured as knowledge graphs. These databases are useful when working with queries involving relationships between entities, as they allow direct and efficient access to these relationships. Unlike relational databases, where tables and foreign keys represent relationships between entities, the relationships are first-class entities in graph databases. This makes it easier to perform complex queries about connections or patterns of relationships, as it is common in social network analysis, recommendation systems, or search engines.

Since graph databases are not based on a strict relational model, they are often classified as NoSQL databases. In this context, they are especially useful for working with highly connected data and semantic queries where traditional databases' hierarchical or tabular structure is not optimal.

2.2 Applications

The use of knowledge graphs can be traced back to their relationship with ontologies in the field of computer science, where they have been used to process expert knowledge in computational processes [1]. Recently, the additional computational capabilities developed through graph-based machine learning have brought these types of data models to the forefront. On one hand, this is due to their ability to produce usable embeddings that can add the contained information to the system, and on the other hand, thanks to their ability to represent and integrate expert information and semantic structures within computational environments.

For all these reasons, such graphs have surged as an advanced technological approach in solving problems in areas such as question-answering systems and recommendation systems. At the same time, new computer approaches have been developed to improve reasoning over graphs, extract knowledge from graphs, and use network-based methods to obtain novel information from those structures.

Regarding the potential areas of application, knowledge graphs have been widely used in biomedical diagnosis [2], where ontologies have already been established as a powerful descriptive tool. Especially drug discovery and drug repurposing (discovering which illnesses can be treated with already known drugs)[3], but also in genomic, clinical, and multi-diagnosis problems, where these graphs enable higher-level of reasoning and connections. Knowledge graphs are also used in industry, primarily for knowledge fusion [4] not only in the maintenance, diagnosis and prognosis of distributed systems, but also in security areas.

Given the capabilities of reasoning and integration of heterogeneous data, knowledge graphs are used in threat modelling and enhance risk assessment [7]. Specially in cybersecurity [6] and financial [5] sectors, for threat prediction and fraud detection [8].

2.3 Computer tools

To conclude this section, it is important to note that, although no standardised query language for graph databases exists that is as widely adopted as MySQL for relational databases, i.e. Cypher (for Neo4j), SPARQL (for RDF graphs), and Gremlin. However, most can read, export, and work with the information represented as triples in the form (S, P, O). In this structure, the subject (S) is the node or entity, the predicate (P) is the relationship represented by an edge, and the object (O) is the related entity. This triple representation is commonly formatted in RDF (Resource Description Framework), a standard model for data interchange on the web.

In our case, the primary tool used for managing the knowledge graph, adding and retrieving information, and extracting knowledge from the graph was Neo4j.

3 Knowledge graphs in airport security

Airport security is one of the top priorities in ensuring the safety of passengers, staff, and facilities. Given the complexity of modern airport operations and the increasing sophistication of illegal activities such as smuggling, an advanced approach is required to manage and analyse the vast amounts of data related to these issues. To address this, we propose a knowledge base that efficiently organises and connects critical information regarding airport security and smuggling.

The structure of the knowledge base presented is built on a graph model that links key entities such as passengers, luggage, flights, inspections, and suspicious activities. Nodes represent essential concepts like individuals, locations, objects, or incidents, while the edges capture the relationships between them, such as movements, interactions, or known associations.

This model enables the detection of risk patterns and potential threats in real-time and facilitates the historical analysis of smuggling cases, helping to predict high-risk routes or behaviours. Additionally, it integrates information from various sources, such as luggage scanner cameras, anonymised passenger data, and customs databases (anonymised), providing a comprehensive view of security operations.

3.1 Bag Intel Proposal

In this section, we will present the proposed knowledge base schema, breaking it down into its key components. Each subsection will explain the role of different entities and relationships within the graph model, detailing how they enhance airport security and prevent smuggling activities.

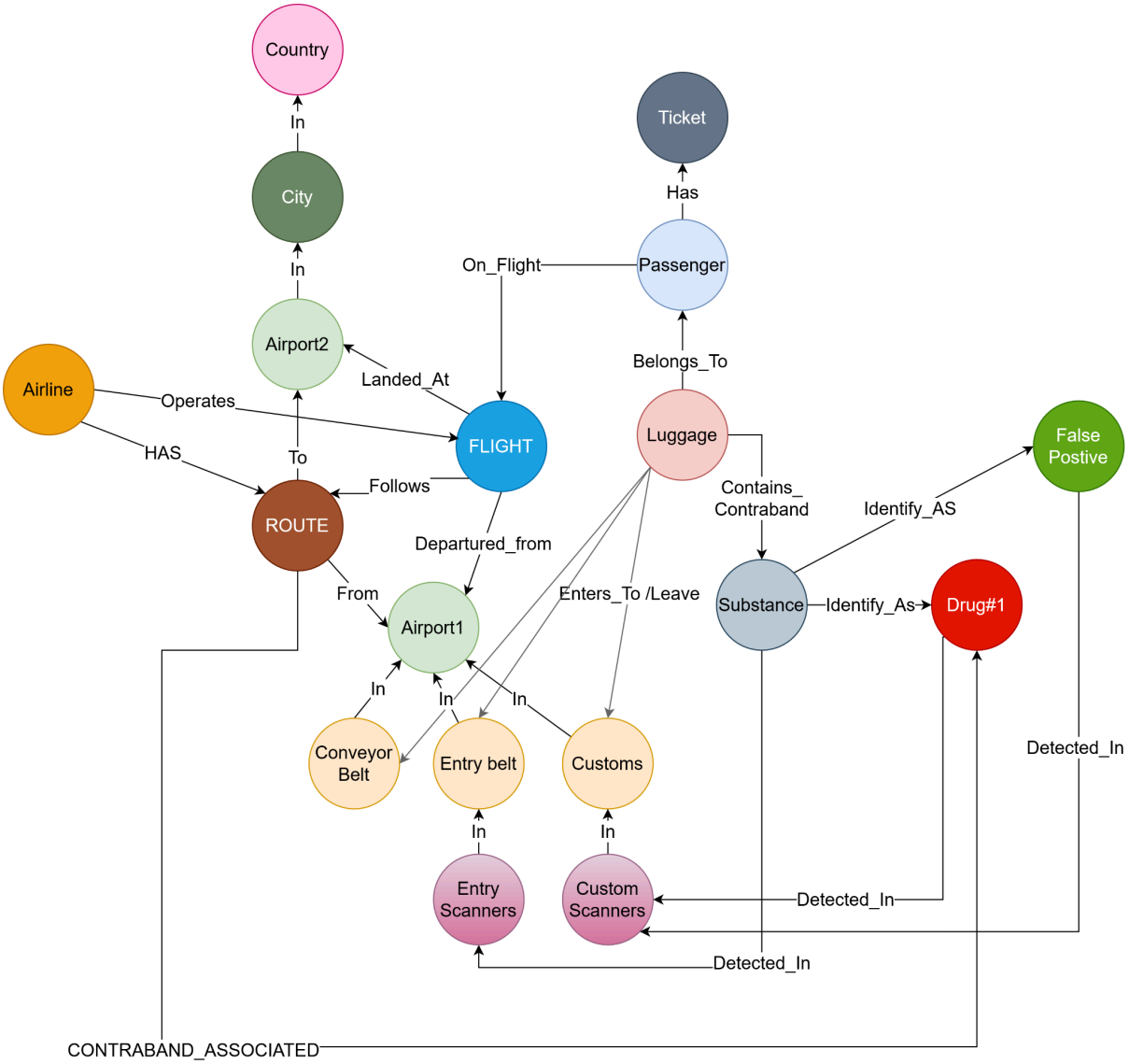

Figure 1: Full Knowledge-base schema. Nodes presented are from the Arrival airport perspective, thus each airport will have this or similar representation, but no departure airports (out of the scope).

3.1.1 Airport distribution and lines

Our main interest in developing a knowledge base from this project is analysing smuggling information and the flights where it is detected. Establishing a node for each aeroplane and passenger is logical, allowing us to account for the type of traffic found in each terminal. However, thanks to the potential of graph databases to incorporate heterogeneous data sources, we can enrich this information in various ways. Our proposal involves describing a geographical sector where we have geographic information about airports. Each airport will be linked to different airlines operating multiple routes. These routes will constitute specific route nodes, which we can use to assess their risk levels based on historical data.

Therefore, the nodes representing flights will be connected to the route nodes. These nodes are designed to highlight, within the database, sets of routes commonly used for smuggling. This feature will allow us to adjust the risk levels of a given flight based on intelligence about intra-city, country, or continent routes provided by security experts.

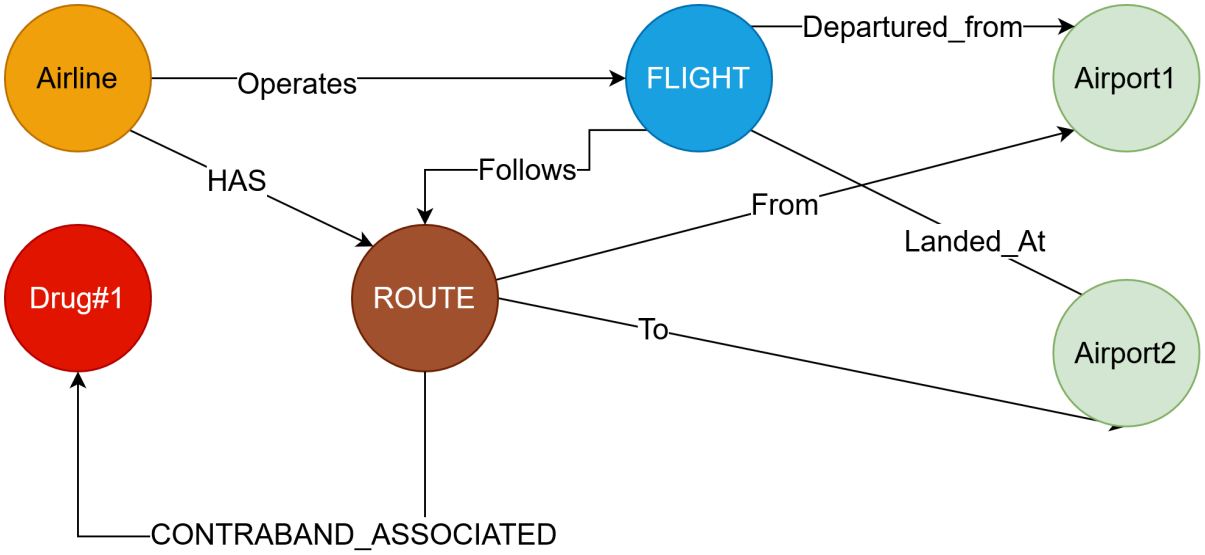

Figure 2: Airport distribution and routes schema connected to drugs classification.

3.1.2 Bag transportation across the terminal

A particular aspect of our project is the re-identification of luggage. At each phase of our project, additional information may be generated that informs us about how a suitcase has moved through the airport. Therefore, we include a node for each suitcase and establish a set of nodes related to the spaces where the luggage may be located. This set of nodes, when we focus on a specific suitcase, can help us confirm its exact location, as well as identify bottlenecks in the system through the timestamps assigned to the edges marking the entry and exit of each room

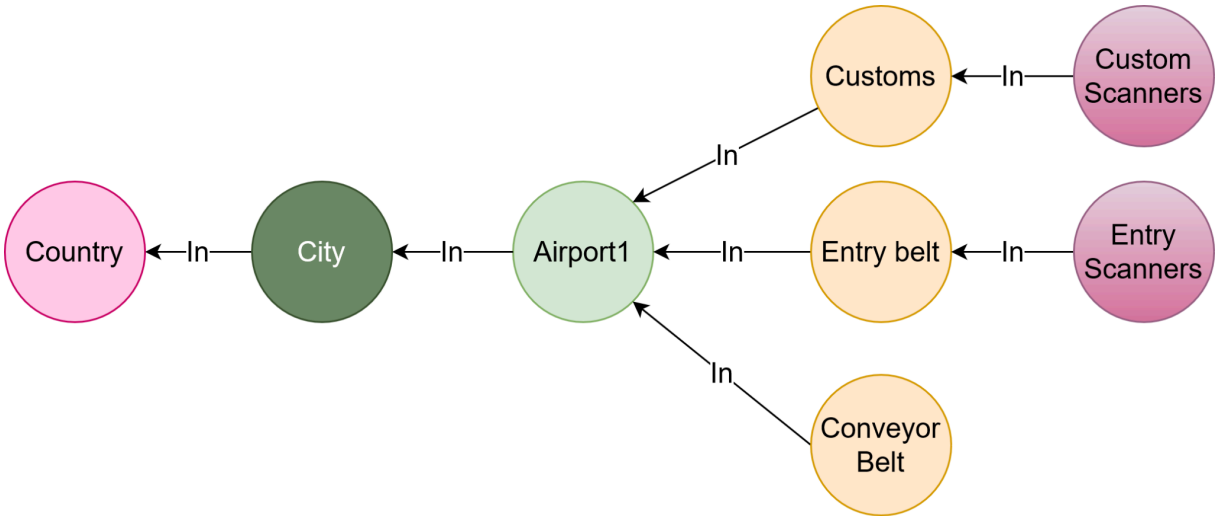

Figure 3: Spaces schema for tracking and storing luggage transportation. These nodes are created and analysed on the Arrivalal airport perspective during the arrival phase of the flight.

3.1.3 Contraband information

The project primarily focuses on gathering information about smuggling routes. This generates a set of additional information that must be incorporated into the knowledge base. To achieve this, our schema includes the relevant ontologies for describing the type of smuggling we are dealing with. Often, a. Thusled technical or chemical description is not feasible during the detection process, which is why our knowledge base separates possible substances by the type that the scanner can detect. Once the content is confirmed, it is linked to the detected substance based on the ontology proposed by the International Narcotics Control Board for Drugs and Psychotropics.

This approach allows us to connect different substances with specific flight routes while providing specific insights into the methods used to smuggle the substances. Furthermore, these substances are identified and linked to the detect method that led to its discovery, therefore it could help us address weaknesses in the contraband seizure.

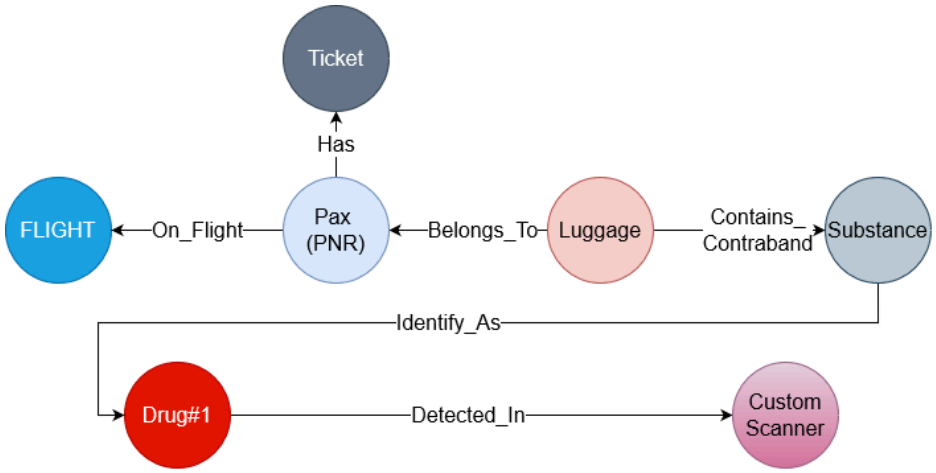

Figure 4: Contraband and Passenger information schema.

4 Ethical and legal aspects

The process presented here is a way to process and store information that can later be used to provide artificial intelligence systems with a baseline dataset that contains semantic and ontological relationships. This adds context and additional information to the processing of that data. It is worth noting that most of the datasources are connected with security factors, for this reason, ethical and legal considerations must be taken into account during our research. Four main aspects are presented here: data privacy, bias and discrimination, transparency, and dual-use risks.

Regarding privacy, the system proposed here complies with the legal requirements of the countries that use it, in this case, the GDPR. All the information currently in the system is synthetic, and the necessary anonymisation methods have been analysed to ensure that no PNR data which is not required for risk assessment is used, so there is no possibility of collecting parameters that uniquely identify a passenger (names, addresses, etc., are not used). Customs officers and data engineers will be the only individuals accessing this data, with the latter always seeing only anonymised information.

As for the risks involved, all the data in the database currently comes from official sources. Storing information in knowledge graphs allows for a better understanding of the decision-making system that uses the collected data. Still, the biases in the system can be analysed using the reinforcement tools proposed in the project. This falls outside the scope of this article, where we focus on how to represent and store the information.

The system also introduces the possibility of storing the final decision made by customs personnel. Therefore, besides including only anonymised information, the graph allows for storing information when future risk assessment mechanisms fail (again, here we address how to represent the information, not how to evaluate or use it for predictions). However, this option ensures that future AI-based systems remain accountable, and we can modify or remove any information in the knowledge base.

Finally, it is noteworthy that there are certain dual-use risks. Although the potential operation of this technology with millions of users is not foreseen, biased risk analysis or the inclusion of biased risk parameters could lead to unintended uses and biases. Our framework allows for reviewing and updating these risks in terms of their representation, providing a more useful interpretation of those included in the database thanks to the additional information. This can help make the algorithms using these data more explainable. However, the origin of these risks lies in the original information sources, whether they are reports or expert opinions.

5 Conclusions

This work presents knowledge bases, their relationship with modern data exploitation techniques, and their typical applications across various fields. Among these, fraud detection and cybersecurity threat identification have been prominent, but to our knowledge, they have not yet been extensively applied to airport security. Within the European project Bag Intel framework, our proposal represents a first step that meets the project's needs and provides a framework and data model that can be expanded to other airports. This model can represent the complexity of airport smuggling situations and enable reasoning techniques through complex semantic relationships to efficiently utilise the data.

Our proposal offers a functional model with the ability to address many uncertainties. In future work, using historical data, we will explore potential applications and the reasoning capabilities of the model to generate insights and enhance the fight against smuggling in commercial flights.

Acknowledgments

Funded by the European Union (Project 101121309—BAG-INTEL). Views and opinions expressed are, however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

1. Ji, S., Pan, S., Cambria, E., Marttinen, P. & Yu, P. S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems* **33**, 494–514 (2022).

2. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* **18**, 1414–1428 (2020).

3. James, T. & Hennig, H. Knowledge Graphs and Their Applications in Drug Discovery. *Methods Mol Biol* **2716**, 203–221 (2024).

4. Buchgeher, G., Gabauer, D., Martinez-Gil, J. & Ehrlinger, L. Knowledge Graphs in Manufacturing and Production: A Systematic Literature Review. *IEEE Access* **9**, 55537–55554 (2021).

5. Ye, X. *et al.* Application of Knowledge Graph in Financial Information Security Strategy. *Proceedings of the 8th International Conference on Cyber Security and Information Engineering*. 192 (2023). doi:10.1145/3617184.3630130.

6. Liu, K. *et al.* A review of knowledge graph application scenarios in cyber security. (2022) doi:10.48550/ARXIV.2204.04769.

7. Zhang, K. & Liu, J. Review on the Application of Knowledge Graph in Cyber Security Assessment. *IOP Conf. Ser.: Mater. Sci. Eng.* **768**, 052103 (2020).

8. Zhang, L. *et al.* Auto Insurance Knowledge Graph Construction and Its Application to Fraud Detection. in *Proceedings of the 10th International Joint Conference on Knowledge Graphs* 64–70 (ACM, Virtual Event Thailand, 2021). doi:10.1145/3502223.3502231.

9. Besta, Maciej, et al. "Demystifying graph databases: Analysis and taxonomy of data organization, system designs, and graph queries." *ACM Computing Surveys* 56.2 (2023): 1-40.