INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Knowledge-based machine learning approach to indirect prospecting methodologies for monument identification

Ariele Câmara

PhD in Information Science and Technology

Supervisors:
Professor Doctor Ana de Almeida
Associate Professor with Habilitation
Instituto Universitário de Lisboa (ISCTE)

Co-orientador(a):
Professor Doctor João Pedro Oliveira
Associate Professor
Instituto Universitário de Lisboa (ISCTE)

December, 2024

iscte
TECNOLOGIAS
E ARQUITETURA

Knowledge-based machine learning approach to indirect prospecting methodologies for monument identification

Ariele Câmara

PhD in Information Science and Technology

Supervisors:
Professor Doctor Ana de Almeida
Associate Professor with Habilitation
Instituto Universitário de Lisboa (ISCTE)

Co-orientador(a):
Professor Doctor João Pedro Oliveira
Associate Professor
Instituto Universitário de Lisboa (ISCTE)

December, 2024

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Knowledge-based machine learning approach to indirect prospecting methodologies for monument identification

Ariele Câmara

PhD in Information Science and Technology

Jury:
Prof. Enrique Cuenca, Associate Professor, Complutense Universidad of Madrid
Prof. José Paulo de Almeida, Assistant Professor, University of Coimbra
Prof. Maria Teresa Batista, Auxiliary Researcher, University of Évora
Prof. Luís Ducla Soares, Associate Professor, Instituto Universitário de Lisboa (ISCTE)
Prof. Ana de Almeida, Associate Professor with Habilitation, Instituto Universitário de Lisboa (ISCTE)

President:
Doctor Fernando Batista, Associate Professor, Instituto Universitário de Lisboa (ISCTE)

December, 2024

Life's defining moments

Acknowledgment

This work was partially supported by FCT – Fundação para a Ciência e Tecnologia, I.P., ISTAR-Iscte project UIDB/04466/2020 and UIDP/04466/2020, and through the scholarship UI/BD/151495/2021. I want to express my sincere gratitude to the Fundação para a Ciência e a Tecnologia (FCT), ISTAR centre and Instituto de Telecomunicações for their invaluable support in recent years. As well, I would like to express my deep appreciation to my advisors, Ana Maria de Almeida and João Oliveira, for all of the time, assistance, and valuable contributions they have provided me.

I am grateful to the Câmara Municipal de Mora for providing me with the books I have been using since my master's program and vector maps of the region.

I extend my heartfelt thanks to my husband, Matheus Silveira, for his daily support and unwavering encouragement, which have been a constant source of strength and inspiration throughout this journey.

Finally, I am grateful to all those who contributed, directly or indirectly, to the development of this work. Without everyone's collaboration and support, this project would not have been possible.

Resumo

Para o reconhecimento de monumentos arqueológicos em imagens de satélite, os especialistas utilizam conhecimento de domínio e a sua experiência. No entanto, métodos automatizados de detecção de objetos, em geral, dependem apenas das imagens, funcionando, no entanto, como "caixas pretas". Esta técnica também tende a retornar uma alta taxa de falsos positivos, especialmente no caso de monumentos de pequenas dimensões. Para que as máquinas possam utilizar o conhecimento de domínio, é necessário torna-lo interoperável mas, para isso, é necessário superar o desafio levantado pela dispersão e fragmentação dos dados. Este estudo aborda essa questão, ao converter conhecimento de domínio de fontes diversas e multidisciplinares em um formato legível por máquina e, assim, poder contribuir para reduzir os falsos positivos na detecção de objetos. A implementação de um *Knowledge Graph* (KG), usando como modelo estrutural o CIDOC-CRM, sua extensão CRMgeo e GeoSPARQL, permitiu a integração de informações sobre sítios arqueológicos e da geografia onde estes estao localizados. Combinando dados textuais com dados espaciais o KG fornece *insights* sobre elementos de uma cena que podem não ser diretamente visíveis nas imagens. Posteriormente, os resultados de uma abordagem de detecção de objetos usando técnicas de Visão Computacional foram integrados no KG para treinar um modelo hibrído *KG-Machine Learning* (KG-ML) para identificar áreas de interesse (AOIs) onde será mais provável encontrar dolmens. O modelo utiliza o conhecimento contextual da área de modo a excluir imagens com baixa probabilidade e melhorar a precisão da detecção.

Abstract

Experts use domain knowledge and experience to identify and analyze archaeological monuments from satellite images. However, traditional object detection methods often rely solely on image data and operate as "black boxes," which frequently results in false positives, especially when detecting small archaeological sites. For machines to effectively leverage domain knowledge, it must be organized in an interoperable format, addressing the challenge posed by scattered and fragmented data, particularly across multiple disciplines. This study tackles this issue by converting domain knowledge from diverse and multidisciplinary sources into a machine-readable format to reduce false positives in automatic object detection. The study links information about archaeological sites and their landscapes by implementing a Knowledge Graph (KG) based on CIDOC-CRM, its CRMgeo extension, and GeoSPARQL ontologies. This KG integrates textual data from semantic records with spatial data from vector topographic maps, encompassing (i) metadata definitions, (ii) general and specific concepts, and (iii) the geometry of each represented entity. This representation can provide insights into elements within a scene that may not be visible in images. Subsequently, the output from an object detection approach was integrated with the KG to train a Knowledge Graph-Machine Learning (KG-ML) model. This model identifies areas of interest (AOIs) where dolmens in Pavia, Mora (Portugal), are likely to be found, using contextual knowledge to exclude images with a low probability of accurate detections. The KG-ML approach effectively reduced false positives, providing contextual information that clarifies recognition decisions and enhancing the understanding of detected sites.

Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

AAT: Art and Architecture Thesaurus
ALS: Airborne Laser Scanning
A-Box: Assertional Box
AOI: Areas of Interest
CA: Carta Arqueológica
CAOP: Official Administrative Map of Portugal
CH: Cultural Heritage
CNN: Convolutional Neural Networks
COS: Carta de Uso e Ocupação do Solo
CRM: Conceptual Reference Model
CRS: Coordinate Reference System
DB: Databases
DBMS: Database Management Systems
DCN: Dilated Convolutional Network
DGADR: Direção-Geral de Agricultura e Desenvolvimento Rural
DGPC: Direção-Geral do Património Cultural
DGT: Direção-Geral do Território
DL: Deep Learning
ESA: European Space Agency
FN: False Negatives
FP: False Positives
GBT: Gradient Boosted Trees
GDB: Graph Databases
GeoKGs: Geospatial Knowledge Graphs
GEOBIA: Geographic Object-Based Image Analysis
GIS: Geographic Information Systems
GML: Geography Markup Language
KB: Knowledge-Based
KG: Knowledge Graph
KGE: Knowledge Graph Embeddings
KNN: K Nearest Neighbours
LARS: Least Angle Regression
LASSO: Least Absolute Shrinkage and Selection Operator
LPG: Labelled Property Graph

LIDAR: Light Detection and Ranging
MRQ: main reserach question
ML: Machine Learning
MLP: Multilayer Perceptron
NASA: National Aeronautics and Space Administration
n/a: Not Available
NGDB: Native Graph Database
OBIA: Object-Based Image Analysis
OGC: Geospatial Consortium
OGM: Object Graph Mapper
OWL: Web Ontology Language
PA: Portal do Arqueólogo
POI: Point of Interest
R-CNN: Region-Based Convolutional Neural Networks
RDF: Resource Description Framework
RF: Random Forest
RoI: Region of Interest
RPN: Region Proposal Network
RS: Remote Sensing
RSI: Remote Sensing Images
SGD: Stochastic Gradient Descent
SIC: Semantic Image Contextualization
SIPA: Sistema de Informação para o Património Arquitetónico
SLP: Single Layer Perceptron
SMOS: Sistema de Monitorização da Ocupação de Solos
SNIAmb: Sistema Nacional de Informação de Ambiente
SNIG: Sistema Nacional de Informação Geográfica
SNIRH: Sistema Nacional de Informação de Recursos Hídricos
STAR: Semantic Technologies for Archaeological Research
SVM: Support Vector Machine
T-Box: Terminological Box
TM: Template Matching
TN: True Negatives
TP: True Positives
UAVs: Unmanned Aerial Vehicles
USGS: United States Geological Survey
VQA: Visual Question Answering
VTM: Vector Topographic Maps
WKT: Well-Known Text
XAI: eXplainable Artificial Intelligence

YOLO: You Only Look Once

CHAPTER 1

Introduction

The expertise of specialists and information about the target is crucial in manual image analysis [9] — a combination here referred to as *domain knowledge*. Domain knowledge in archaeology combines visual and spatial analysis with ancient and modern landscape perceptions searching for greater meaning in findings [10]. In contrast, existing automatic data-driven approaches, such as object detection, process Remote Sensing Images (RSI) using algorithms and Machine Learning (ML) techniques [11], often exclude the knowledge critical to human experts. Hybrid approaches that combine data-driven and knowledge-driven methods [12] are currently seen as promising developments for automatic detection in Remote Sensing (RS) and in the field of archaeology [6] to contribute to explainability, ultimately leading to more robust object detection systems, particularly in complex scenarios. These approaches leverage existing information, relationships, and contextual understanding to make informed decisions, similar to human reasoning.

In this light, we aim at enabling machines, that is, automated systems, to leverage domain knowledge about the local landscape to enhance the location of archaeological remains through image classification. Towards this end, we explored knowledge-driven methods to reduce the number of False Positives (FP) (or incorrect predictionsFP [13]) hits returned by data-based image recognition methods. In archaeology, potential sites requires manual intervention thus misidentifying non-archaeological characteristics is costly and highly time-consuming. As datasets grow larger, the need for a more precise automated recognition also increases especially when analyzing vast areas, where the occurrence of numerous FPs can be troublesome. While the final interpretation will always rest with the expert, reducing the amount of data for manual analysis allows specialists to focus on other critical tasks, thereby optimizing the overall efficiency of the archaeological survey process [14, 15, 16].

In this chapter, first, we describe the motivation for using hybrid approaches to identify archaeological monuments in satellite images (Section 1.1). Next, we outline the problem of using domain knowledge in archaeology and define our case study (Section 1.2). Afterwards, we present the research questions (Section 1.3) and research methods (Section 1.4) used to guide and structure this thesis. After that, we reflect on the contribution and a list of the publications derived from this project (Section 1.5). Lastly, we present the dissertation outline (Section 1.6).

1.1. Motivation

The human mind can recognise a wide range of concepts through visual analysis, including complex architectural styles. Often in architecture, categories are ambiguous, overlap (e.g., castles and palaces), and include numerous subcategories (e.g., houses). In addition to common categories like houses, castles and palaces, specialized categories like megalithic monuments are less common and have fewer examples. Although limited examples are available, humans can recognize these categories through experience, generalization and inference [17]. For example, [18] manually analyzed satellite images to identify 187 dolmens recorded in Mora and Arraiolos (Portugal), with 60 monuments successfully identified in the images. It took several months to analyze images from just one year, 2017, to recognize these monuments. As remote sensing data increases, this type of manual analysis becomes not only time-consuming but even insufficient.

In light of the increasing volume of data available for analysis, data-driven approaches to recognize archaeological sites in RSI have become more and more prevalent. These typically use Airborne Laser Scanning (ALS)/Light Detection and Ranging (LIDAR) [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], satellite imaging (e.g., panchromatic, multispectral and hyper-spectral) [16, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46], or images from platforms such as Google Earth [47, 48, 49] and BING [50]. However, to achieve accurate recognition, automatic visual recognition systems require thousands of labelled examples per category [51, 52], requiring a large and diverse set of examples for training [53]. Additionally, since objects vary in illumination, occlusion, and perspective, these approaches have difficulty in recognizing real-world scenes [54].

Using ML techniques to analyze archaeological sites from satellite images when data is scarse and categorically complexity make accurate classification difficult. However, this challenge can be addressed by exploring the information humans use in manual analysis, which is often overlooked by automated methods since, for several reasons that will be discussed further along (like the nonexistence of a symbolic knowledge base gathering the until now dispersed information) it is not easily incorporated into the workflow. Unlike data-based approaches, humans leverage their experiences and previous knowledge to identify objects. For example, to identify immovable archaeological monuments in RSI, researchers traditionally recur to **domain knowledge** about the monuments' and to the topographical features of the surrounding environment (e.g., geology, soil type and use, hydrology, and archaeological features) [53, 55]. The realization that manual identification relies heavily on domain knowledge that automated systems may not fully consider led us to explore ways of incorporating domain knowledge into recognition systems' processes.

1.2. Problem Definition

Although most archaeological research relies on visual identification of past remains, the majority of the remains are hidden from view, whether buried underground by environmental processes, covered by modern towns, or by vegetation. The use of RSI helps to identify these hidden, or so-called "invisible," remains, at the same time that places them

in much larger contexts [10]. However, regarding small and non-easily distinguishable structure monuments in satellite images remains a challenge for identification, whether in manual or automated analyses [34].

The surrounding landscape may have played a critical role in the choice for the placement of monuments like megalithic structures [56], and although the landscape has evolved over thousands of years, analyzing its characteristics can help identify patterns to distinguish Areas of Interest (AOI) [18]. For example, key patterns for dolmens location include elevated terrain and proximity to natural rocky outcrops, water lines, and other similar monuments [56]. Based on this knowledge, we can infer that hydrography, hypsometry, soil type, and land use data can assist in identifying an AOI to locate dolmens. By analysing spatial relationships between these characteristics within a scene, we can identify patterns to determine how environmental factors influence monument presence and visibility. The main challenge, however, is acquiring data and information that can be made interoperable and machine-understandable for automated approaches to use as domain knowledge.

Obtaining domain knowledge for monument recognition involves integrating multidisciplinary information, such as landscape data from RS and Cultural Heritage (CH) data from archaeology. Most researchers focus on well-defined entities that can help identification, which makes the detection dependent on the specialist's interpretation that establishes the relation between the visual entities and the known concepts [55]. However, prior information and relevant data tend to be kept by different communities and institutions, which employ conceptualizations and formal definitions that usually do not match. Take the example of geoinformation, traditionally represented by Geographic Information Systems (GIS), while CH data tend to remain as records in museums and historical archives [57]. This diversity of data and data sources is problematic when the goal is to use this information in automated systems and to provide context to ML models.

To address the lack of interoperability, each community has developed ISO standard ontologies for information representation reflecting its particular focus. For example, Geospatial Consortium (OGC) (GeoSPARQL) [58] was developed for geoinformation recording using a typical formalism suited for semantic web technologies. In contrast, CH information tends to be represented using the CIDOC-Conceptual Reference Model (CRM) [59]. Since both ontologies provide standards for the specified areas, if applied together, they can enrich a model with precise and well-identified descriptions of site location — such as CRMgeo [57], which extends CIDOC-CRM.

Knowledge-Based (KB) techniques play a crucial role in achieving interoperability by representing domain knowledge and also open the path to deal with the explainability commonly amiss in traditional automated image recognition approaches [60]. For instance, the symbolic representation of Knowledge Graph (KG) has gained increasing attention in recent years due to its ability to integrate, organize, and allow for reasoning over vast amounts of domain-specific and inter-related information, enhancing AI systems'

performance in tasks requiring contextual understanding. Information in KGs is represented semantically in a linked way, emphasizing relationships between entities. While KGs excel in organizing and linking domain-specific knowledge, object detection focuses on analyzing images, that is, image data to identify objects without necessarily considering the object's underlying context. As it will be described in the next chapter with the review of the related literature, many studies explore object detection to recognize sites in RSI, and others explore ontologies and KGs to represent semantic information. Despite the growing interest in both methods, their integrated application in representing spatial data with monument information to assist in scene recognition is a recent trend and remains virtually unexplored within the field of archaeology.

1.3. Research Questions

In light of the problem definition outlined in Section 1.2, this research was driven by the following main reserach question (MRQ):

- MRQ: To what extent does the provision of landscape context information improve the precision of automated systems in archaeological site object detection?

As previously stated, incorporating domain knowledge in data-based approaches offers a promising avenue to provide contextual information that can enhance the performance of automated methods and their explainability. With this understanding, we can identify the potential of using domain knowledge with object detection. As a first step, we need to review the current state of this approach in the archaeological field, which leads us to the following more specific research question:

- RQ1: Are there approaches that combine data-based and knowledge-based methods to enhance the recognition of archaeological sites in satellite images?

After reviewing the state-of-the-art, we focus on two research avenues: one involves implementing a data-driven model, while the other aims to implement a semantic model to represent domain knowledge about the target object. This brings us to our second research question:

- RQ2: How can contextual information about archaeological sites and their surrounding environment be made interoperable to enable machine learning systems to identify and use domain-specific knowledge?

Building on the findings derived from the implementations of data-driven and knowledge-based methods, we also explore how to use the semantic model to improve the detection of the data-driven approaches to answer our third specific research question:

- RQ3: Is it possible to improve the detection of small or non-visible objects in data-driven approaches by incorporating semantic contextual knowledge for guidance?

This three-part research question approach has lead to a model implemented in order to enable answering the MRQ and thus achieve the main research goal: the proposal

of a hybrid KG plus ML method that uses symbolic knowledge to assist a sub-symbolic method in the detection of dolmens in satellite images.

1.4. Research Methods

In this dissertation, archaeological sites featuring dolmens located in Pavia (Mora, Portugal) is defined as the case study that serves as a basis for addressing our MRQ (the target object and area of interest are described in Section 2). This focused approach minimizes variability introduced by regional differences and ensures that the insights are deeply rooted in the selected AOI characteristics.

To address RQ1, we conducted a systematic literature review, reported in Chapter 3. Object-based and knowledge-based approaches in archaeology rarely perform together despite being widely used separately. As a result, we determined two lines of analysis to identify how these approaches are combined. The first focuses on understanding the data-based approaches, specifically object detection techniques, used in archaeology to identify monuments in satellite imagery. The second line of analysis investigates how knowledge-based approaches represent archaeological and spatial information, explicitly using ontologies and KGs. By combining insights from these two viewpoints, we aim to evaluate the potential for integrating data-based and knowledge-based approaches to improve the accuracy and reliability of monument recognition in satellite images.

For answering RQ2, we implemented a satellite image-based approach for recognizing dolmens, reported in Chapter 4. Next, we performed an implementation of a KG as a Labelled Property Graph (LPG) to link the information gathered and make it interoperable, which is reported in Chapter 5. Vector Topographic Maps (VTM)s with hydrological, soil types, soil uses, topographic relief information and textual fonts containing monument-related information were used to represent the knowledge. The schema model here proposed uses CIDOC-CRM, CRMgeo, and GeoSPARQL ontologies as a base to represent the information. The goal is to represent each entity (e.g., archaeological sites and surrounding environments) as structured instances, capturing their attributes and relationships.

To answer RQ3, the LPG was combined with the object detection outputs (bounding box with scores and coordinates of analysed images considered as Point of Interest (POI)s) to train an ML model, resulting in a KG-ML approach for identifying AOIs to recognize dolmens, reported in Chapter 6.

As a result, we present a method that integrates domain knowledge in an interoperable format, making it suitable for automatic archaeological site detection, whose applicability was demonstrated through testing, demonstrating its ability to reduce false positives.

1.5. Contributions

This project contributes to archaeology as well as information science. As a result of linking semantic and spatial information from various VTMs and semantic sources about the dolmens, its landscapes and their locations into an LPG, we created a model that

is human-machine-readable, interoperable, and capable of organizing data from various sources and formats. This model allows users to trace the origin of each piece of information, making the data reusable and, by centralizing everything in one place, easily accessible and findable. As the LPG acts as a reviewer, it operates independently of the object detection method, being agnostic and capable of utilizing outputs from any other approach, validating and refining the results from data-based recognition approaches.

To our knowledge, this is the first time an LPG has been used to review automated archaeological site recognition to provide context to identify AOIs and minimize FPs. Through the usage of well-known and accepted ontologies to implement an LPG, this research advances the integration of spatio-temporal and semantic data from multiple sources. As a result, the outputs of data-based approaches are enhanced, and the volume of data to be analysed manually is decreased. While advancing theoretical insights into knowledge representation and automated data analysis, the project's contributions offer practical benefits for archaeological research and heritage preservation.

This thesis builds on several research outcomes published in peer-reviewed venues. Each work contributes to understanding automatic approaches for recognizing archaeological monuments, knowledge representation in cultural heritage and remote sensing, and the integrated use of both methods.

- Camara, A., de Almeida, A., Oliveira, J. P., & Silveira, M. (2020, January). Photointerpretation as a Tool to Support the Creation of an Ontology for Dolmens. *In Program and Book of Abstracts XXVII Meeting of the Portuguese Association for Classification and Data Analysis (CLAD)* (p. 101).
- Câmara, A., de Almeida, A., Caçador, D., & Oliveira, J. (2023). Automated methods for image detection of cultural heritage: Overviews and perspectives. *Archaeological Prospection*, 30(2), 153-169. DOI: 10.1002/arp.1883
- Câmara, A., de Almeida, A., & Oliveira, J. (2023, May). Versioning: Representing Cultural Heritage Evidences on CIDOC-CRM via a Case Study. *In Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (pp. 363-371).* Singapore: Springer Nature Singapore.
- Câmara, A., Almeida, A. D., & Oliveira, J. (2024). Transforming the CIDOC-CRM model into a megalithic monument property graph. *Journal of Computer Applications in Archaeology.* DOI: 10.5334/jcaa.151
- Marçal, D., Câmara, A., Oliveira, J., & de Almeida, A. (2024, June). Evaluating R-CNN and YOLO V8 for Megalithic Monument Detection in Satellite Images. *In International Conference on Computational Science* (pp. 162-170). Cham: Springer Nature Switzerland.
- Câmara, A., Almeida, A. D., & Oliveira, J. (2025). A Knowledge-Graph for Portuguese Megalithic Monument-Landscape Relationships Representation and Analysis. Ed: Stefania Stellacci, Serdar Aydin. The paper has been accepted

and is forthcoming in *Endangered Heritage Sites: From Eidotypes to Enriched Representations and Design Solutions* in February 2025.

Other Publications include non-peer-reviewed works presented as posters, which are listed below.

- Câmara, A., Almeida, A. D., & Oliveira, J. (2022). KG-ML Approach Image Recognition for Cultural Heritage. In Ciencia 2022 - Science and Technology in Portugal Summit.
- Câmara, A., Almeida, A. D., Oliveira, J., & Marçal, D. (2023). Arqueologia e Comunicação na era da Big Data: do sítio arqueológico ao registo de monumentos e paisagens. Será este um dia FAIR? *In IV Congresso da Associação dos Arqueólogos Portugueses*.

1.6. Document organization

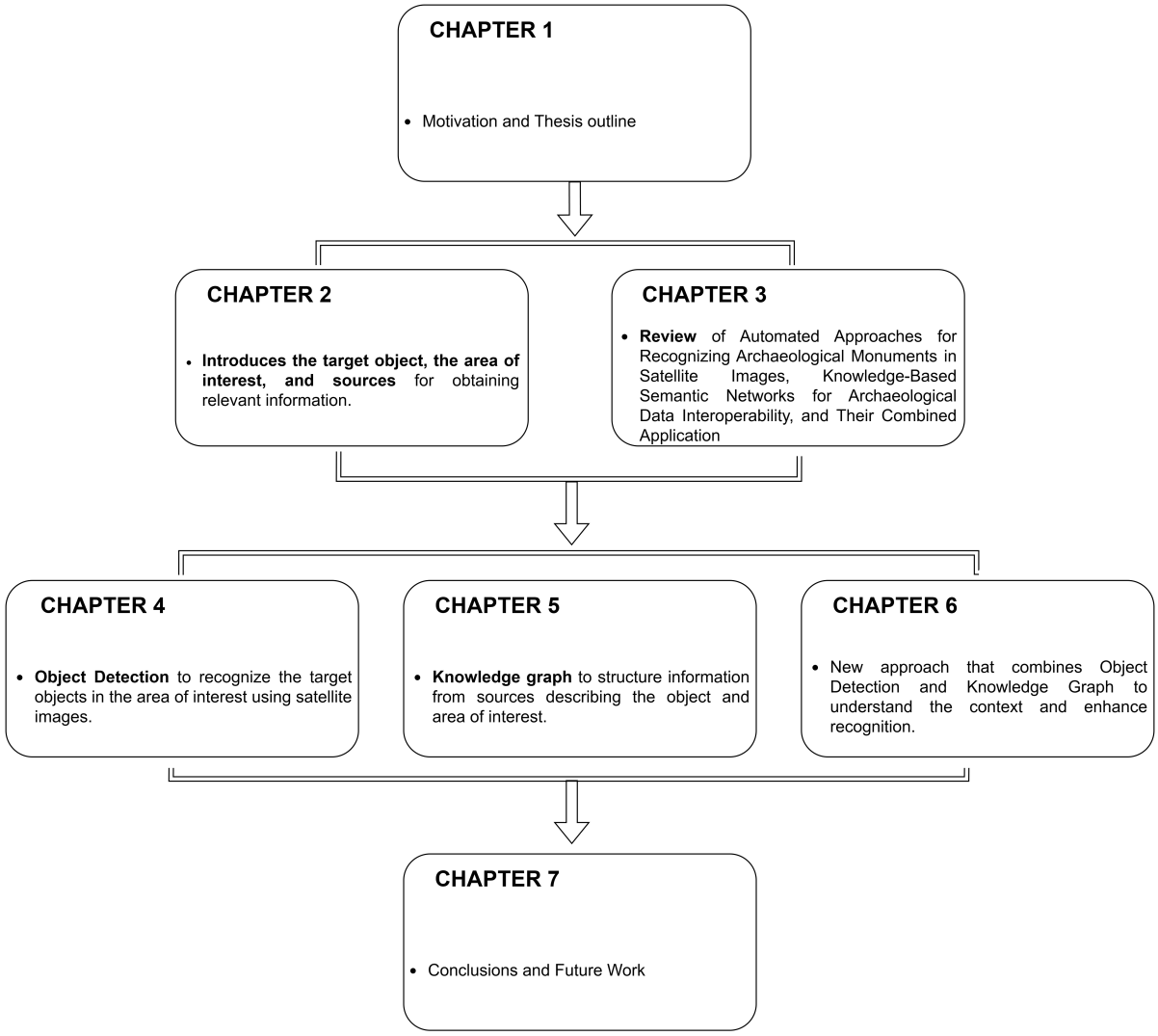


FIGURE 1.1. Dissertation outline

This dissertation is divided into seven chapters as depicted in Figure 1.1. The first chapter (this chapter) introduces the problem addressed, methods and contributions. In the

following, Chapter2 describes the AOI, including the monument and landscape features. Chapter 3 presents a review of the state of the art, focusing on object detection approaches to recognize archaeological monuments and knowledge graphs to contextualize archaeological information. Next, the model implementation is subdivided into three major chapters: implementing an object detection approach based on ML (Chapter 4), detailing the technical approach and processes involved, implementing the KG (Chapter 5), explaining its design and integration, and implementing the KG-ML approach (Chapter 6), in which results from object detection are combined with the KG. Finally, a conclusion is shown in Chapter 7, which summarizes the research findings and suggests research avenues for the future.

CHAPTER 2

Area of Interest: Definition and Data Sources

This case study uses information on dolmens, specifically those located in Pavia in Mora in the Alentejo region of Portugal — our Areas of Interest (AOI). This city is part of the Mora municipality, including Mora, Brotas, and Cabeção. Pavia is situated at the northern edge of the Évora district, near the borders of Santarém and Portalegre [56]. It consists of an area of 185 km2 [2] and has been occupied since prehistoric times, as evidenced by its numerous megalithic monuments such as dolmens.

In this chapter, we present in detail our target object (Section 2.1) and the landscape description of the AOI where they are situated (Section 2.2) to understand the context of what we are looking for and the relevant information. We also surveyed the geospatial and semantic data — sources used to gather information about the target object and the surrounding landscape (Section 2.3). Data and information from these sources can help derive domain knowledge about the target object and the AOI.

2.1. Dolmens

Dolmens are megalithic monuments primarily recognized for their funerary functions, typically serving communal purposes [61, 62]. In Portugal, these structures, built during the Neolithic and Chalcolithic periods (4000-5000 BC), were used to facilitate burial practices and serve as collective memorials [61, 63, 64, 65, 66]. Portugal's Alentejo region has one of the highest concentrations of megalithic sites in Europe [67, 21]. In the Alentejo region, these structures typically consist of a chamber formed by three or more uprights (orthostats) supporting a single cover-stone (capstone) [68].

These vertical stones bear the weight of the overarching capstone and demarcate the chamber's confines. It also may have a corridor as an entrance composed of orthostats. These structures may have been covered with earth and stone (burial mound or tumuli) [69, 63]. This human-made mound, raised over the dolmen, possibly played protective and symbolic roles. In Figure 2.1, it is presented a well-preserved example of a dolmen captured in a 3D model [1]. The image reveals its large chamber with eight orthostats (1), a capstone split in half that originally measured about 3.85 meters in length (2), a corridor (3), and a well-preserved tumulus (4).

It has been observed that dolmens had their opening facing the rising sun and generally diverged from East to South by 10°-20°. The differences in orientation, evidently based

[1]The image can be accessed at: https://sketchfab.com/3d-models/anta-grande-da-comenda-da-igreja-5bd4c1bddaf64c38937f6c47a71a79e6 [Last accessed in 09/12/2024]

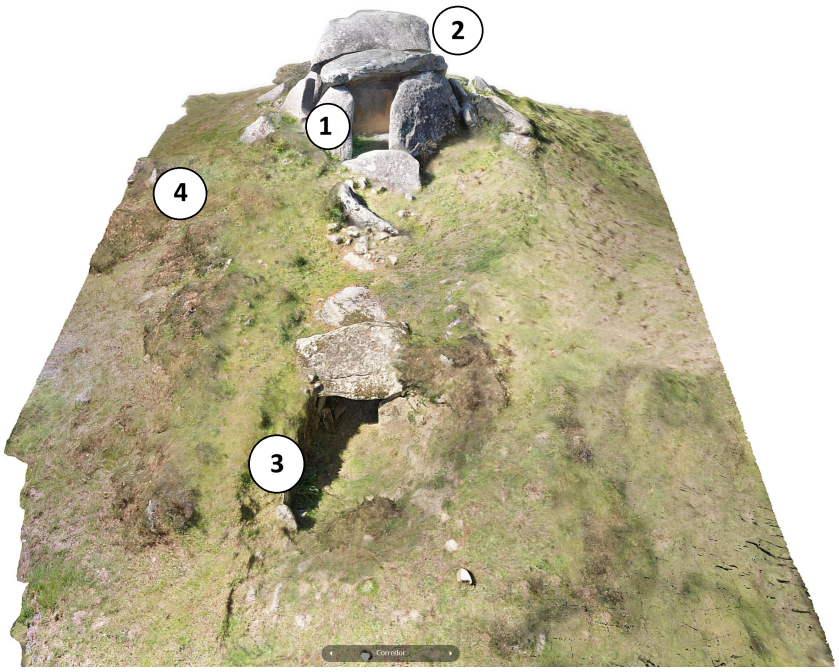


FIGURE 2.1. Anta Grande da Comenda da Igreja. A three-dimensional reconstruction. Source: Sketchfab. License: CC0 Public Domain.

on astronomical observations, could be attributed to the time or the season of the monument's construction [61, 70]. However, the theory remains debated and is not universally accepted, with a larger-scale analysis needed to definitively attribute this characteristic to these monuments [69].

Studies on megalithic monuments have existed in the Pavia area since earlier, with notable works dating back to the early 20th century [70, 71]. Recently, efforts have been made to map these burial types using LIDAR data in the Alentejo [21] as well as a plan to classify and protect them (Official Gazette No. 39/2022, Series 2 of 25/02/2022) [72]. Currently, 94 monuments are recorded in the area of Pavia [73, 8]. The map depicted in Figure 2.2 showcases Portugal with a detailed view of Pavia, signalling the locations of the dolmens that have been analysed.

A typical example of this monument type in the region features a chamber diameter of two to five meters and a variety of shapes (polygonal, circular, semicircular, or quadrangular). It is usually built from granite or schist [74, 75]. The chosen construction materials, mainly granite and schist, imparted a distinct aesthetic and fortified them against the relentless march of time. Schist or granite soils are acidic, leading to the complete decomposition of organic matter over time. As a result, no organic remains survive, leaving only the architectural traces of these structures to endure through the ages [76].

2.2. Landscape

Portuguese continental geomorphology is characterized by three major morphostructural units: the Ancient Massif, the western and southern Meso-Cenozoic fringes, and the Cenozoic Tejo-Sado basin [77]. Alentejo falls primarily within the Ancient Massif unit,

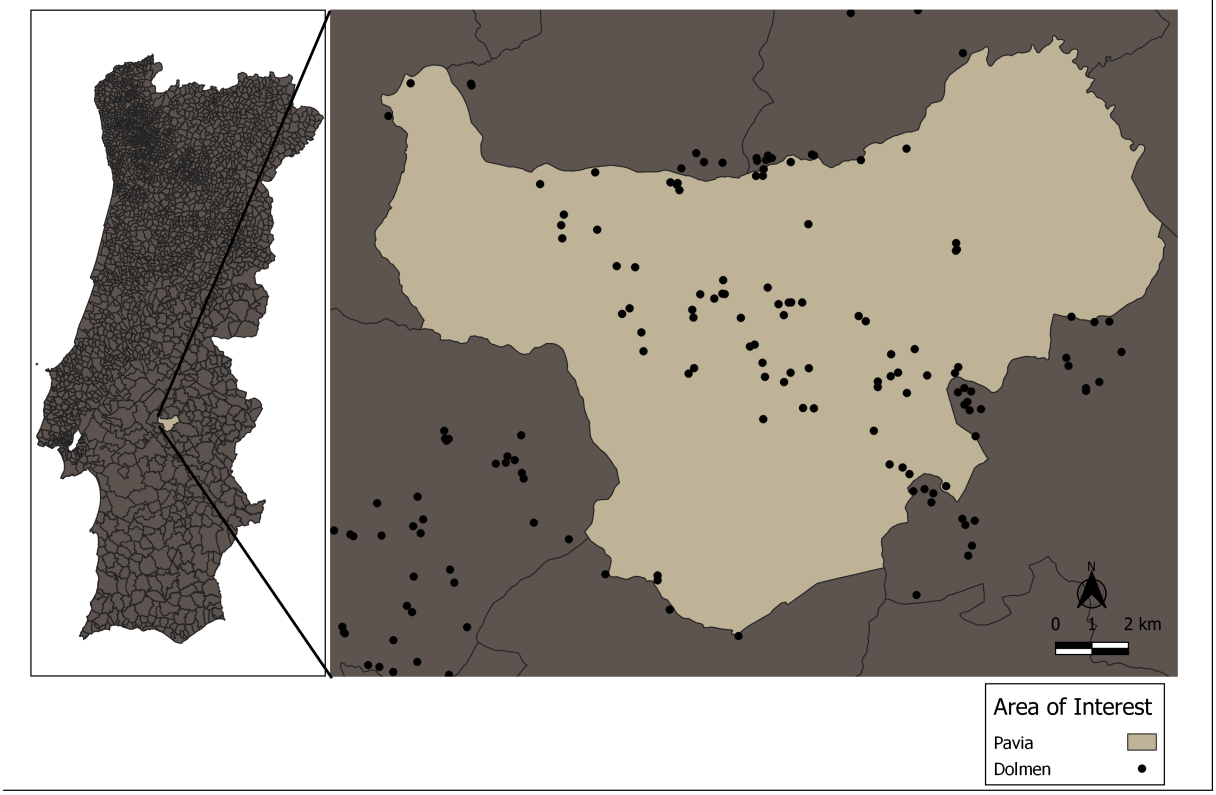


FIGURE 2.2. Map highlighting Portugal with a detailed inset of the Pavia region, situated within Mora in the Alentejo area.

which is characterized by granite and schist, as well as some quartzites and metamorphic rocks [77]. From a geological perspective, the Pavia area encompasses two formations: the edges of the Tertiary Tejo basin and the ancient substrate, which is mostly made of granite. Both formations are intersected by a dense watercourse network [77, 56]. The Tertiary cover includes flatter areas, sometimes forming residual reliefs (*W-S*), while the ancient substrate presents a more irregular topography. However, it is predominantly flat with extensive granite outcrops [56]. There are generally gneisses, mica-schists, metamorphic schists, and granites in this region (*Ossa Morena*), as well as lesser amounts of marble, quartz, quartzite, clay, sandstones, and conglomerates [56, 78, 79].

Despite the relatively flat topography, numerous watercourses cut through the area [77]. Central Alentejo has three major river basins: Tejo, Guadiana, and Sado. The Pavia area is part of the Sorraia river basin, a subsidiary of the Tejo left bank [56]. The river exhibits high irregularities, resulting in more severe droughts and more concentrated runoff [80, 56]. According to the VTM to water lines from *Agência Portuguesa do Ambiente (APA)* at scale 1:25,000, this river is considered to be artificial, and its left bank tributaries are the Raia, Mora, Matalote and Tera rivers. In addition, the Tera River, a tributary of the Sorraia River, traverses the entire central region of Pavia in a north-south direction [4]. According to Ramos (1994), the basin extends in an ESE-WNW direction, following the flow direction of the primary collector, which begins on the Estremoz Plateau and flows

to the Raia River [80]. Among the tributaries of the Tera are the Cré River on the left bank and the Freixo River on the right bank. Other natural streams in the region include Azenha and Divor rivers [4].

The interaction between the geological characteristics of a watershed and its drainage network profoundly influences soil types and water movement across the landscape [81]. Over time, various factors can significantly alter watercourses, including deforestation, dam construction, and natural hydrological variations [80]. Moreover, precipitation patterns also affect changes in the landscape. In this region, the average annual precipitation ranges between 650 and 700 mm, slightly increasing to 700-800 mm in the northeastern part. Rainfall is evenly distributed throughout the year, primarily between October and May [77].

Solos Litólicos refer to soils generally associated with rocky outcrops, unsuitable for any agricultural use except with manual labour [82]. Much of the soil in the central region and surrounding the Tera, Cré, Matalote, and Divor rivers consists of *Solos Litólicos* and *Solos Argiluviados Pouco Insaturados* [83], as shown in Figure 2.3. These soil types have low natural fertility due to their parent material, mainly granite, its derivatives, and schists. The soils are characterized by low cation exchange capacity and high acidity, further limiting their fertility [84]. Other soils present in the region, include *Solos Incipientes*, *Barros*, *Solos Hidromórficos*,and *Solos Calcários* [3]. *Solos Incipientes* and *Solos Calcários* are weakly developed soils, the former being minimally evolved with shallow organic layers, while the latter forms from limestone with varying carbonates. *Barros*, in contrast, are well-developed, clay-rich soils with high plasticity and firmness. *Solos Hidromórficos* face challenges due to temporary or permanent water saturation, leading to reduction phenomena in the soil profile [85, 82].

Regarding soil quality, granitic formations generally result in low agricultural potential, typically classified as Class D and Class E soils. In contrast, tertiary formations exhibit two distinct scenarios: the Oligocene clayey-limestone terrains support good areas of agricultural soils, classified as Class B and Class C, while the Miocene-Pliocene complex of sandstone-clay conglomerates in the plateaus predominantly consists of soils with very low agricultural suitability, classified as Class E [56]. Soil classification encompasses various categories based on their usability for different purposes. Class A soils have very high usability with minimal limitations and erosion risks, suitable for intensive agricultural use. Class B soils have high usability but moderate limitations and erosion risks, making them ideal for moderately intensive agriculture and other uses. Class C soils offer moderate usability, significant limitations, and high erosion risks, and they are suitable for light agricultural use. Class D soils have low usability with severe limitations and high erosion risks. They are often unsuitable for agriculture except in exceptional cases but may be used for grazing, woodland, or forestry. Class E soils have very low usability, severe limitations and high erosion risks, rendering them unsuitable for most uses, often designated for natural vegetation or protection forests [82].

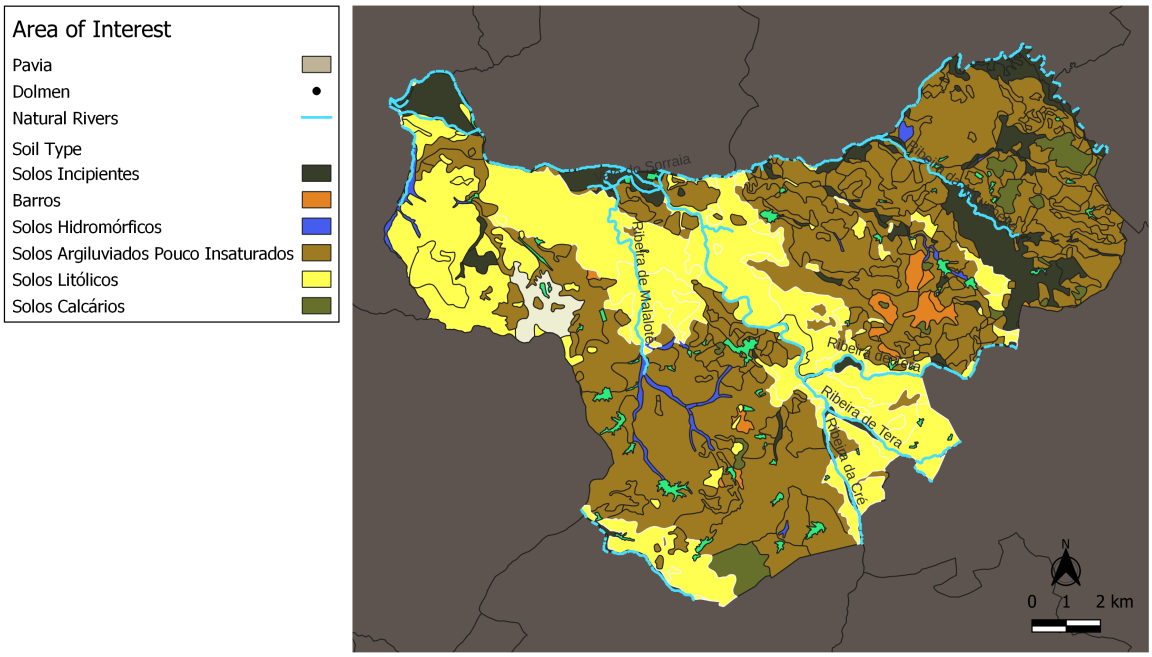


FIGURE 2.3. Map highlighting the soil types in Pavia. The map was created using VTMs [2, 3, 4]

In general, this region is relatively flat with elongated relief and gentle slopes near watercourses, along with small hills and spurs [86], with altitudes ranging from 50 to 205 meters as shown in Figure 2.4. Among the areas in the region, Pavia stands out as having the highest proportion of elevated areas above 145 meters. The northern region, including the eastern and western parts, features the lowest altitudes. Conversely, the central to southern regions exhibit the highest elevations. Rivers like the Tera are situated in low-lying areas but are often surrounded by higher elevations, creating a varied landscape. The region's general flattening results from typical water erosion, which has gradually smoothed and levelled the landscape into a peneplain [5].

Regarding current land use, the region is predominantly characterized by areas devoted to cereal crops, olive groves, and vineyards [86]. These agricultural practices dominate the landscape, contributing to the region's economic activity. In addition to these cultivated areas, there are significant forested regions, including *montado* systems, which consist of oak and cork oak forests [87]. It is important to note that land use practices can influence soil visibility. For instance, the extent of vegetation cover and agricultural activity can obscure or alter the appearance of soil surfaces.

This landscape has profound utilization practices that can influence soil visibility and megalithic monuments, which are strategically positioned near watercourses, rock outcrops where they could access raw materials for constructions and elevated points to

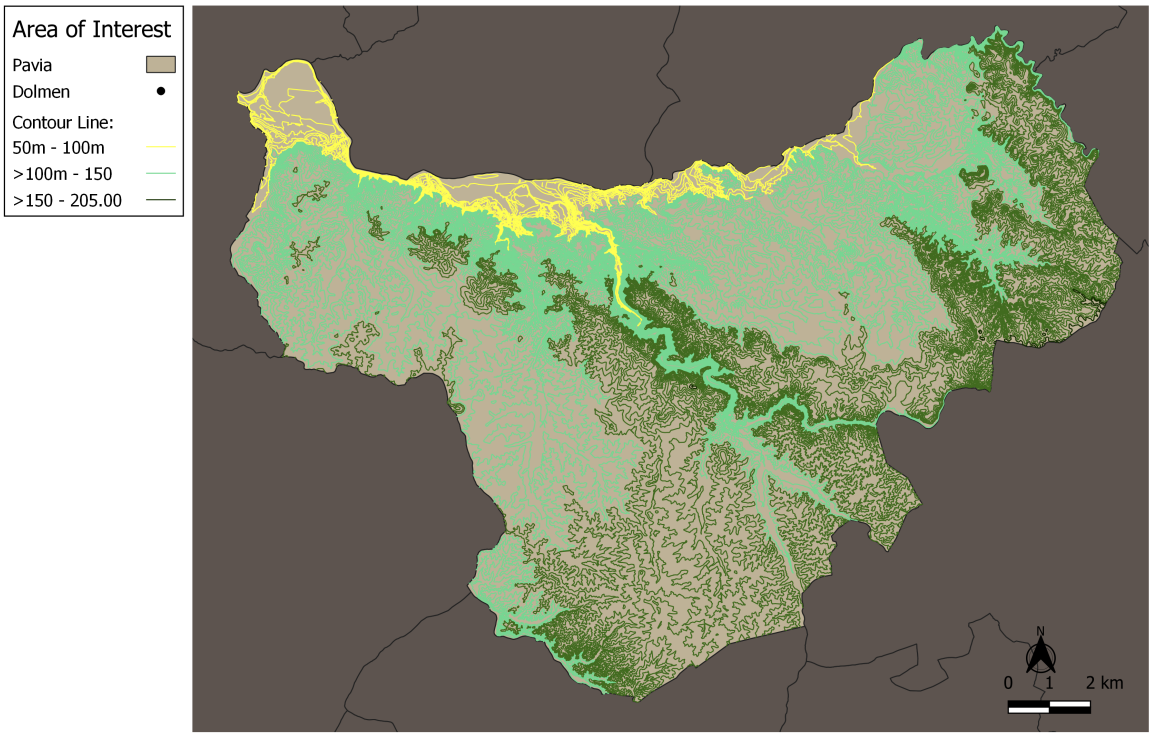


FIGURE 2.4. Map highlighting the topographic relief in Pavia. The map was created using VTMs [2, 5]

maximize visibility [56, 86, 88]. At a macro-geographic scale, the placement of tombs is generally considered deliberate, with site selection potentially based on practical reasons or cultural factors. These results rely on most studies on astronomical or landscape observations [76]. While current land use does not directly influence the original selection of these sites' placement, it could impact their preservation. The analysis of previous data shows that many monuments were destroyed in Alentejo during the 20th century, mainly between the 1970s and 1990s. These destructions were primarily driven by intensive agriculture and reforestation efforts, which have significantly altered the landscape and compromised the integrity of these sites [67].

Neolithic communities were among the first to develop and rely on agriculture, albeit on a small scale—more akin to horticulture as we understand it today. They prioritized settling on light, well-drained soils near permanent watercourses where they could cultivate crops. Heavy, clayey soils were generally avoided, even if close to water, as these soils were challenging to work with using the technology available at the time. Therefore, Neolithic communities did not favour soils classified as A and B [76, 86]. In areas with Class B and C soils, burial sites tend to be limited [76]. Conversely, soils with poorer agricultural suitability, such as Classes D and E, which are often schistose, tend to host more burial sites [56, 76].

The archaeological landscape is, in essence, a product of long-term interactions between humans and the natural environment [6]. Over time, the remains of past civilizations have become buried or obscured by natural processes and later human activities, often leaving monuments and structures concealed beneath layers of earth. This process of layering, or stratification, is fundamental to archaeological excavation. In this ongoing process, the uppermost layer represents the most recent interaction between humans and the environment [89].

2.3. Data Sources

The information contained in cultural heritage sources (e.g. archives, museums, records and databases) and from public entities responsible for the management and coordination of policies related to the territory, such as those responsible for creating Vector Topographic Maps (VTM)s, is crucial as a source of knowledge about the territory. We can interpret, understand, and extract accurate information from reliable and relevant data sources from which domain knowledge is consistently derived.

Detecting archaeological sites in an image requires domain knowledge that involves a multidisciplinary effort to understand the monuments and their surrounding landscapes — involving geospatial and semantic data and relating archaeology, geography and Remote Sensing (RS). The landscape is a dynamic entity that evolves and houses various geographical objects and features. Archaeological sites can be viewed as a subset of geographical objects, enclosed by significant areas, representing remnants of the original archaeological record at a specific morphogenetic stage [6]. We infer that geographic objects and features encompass all physical elements above the Earth's surface, including natural landscape features, built structures, and archaeological monuments.

Data sources with information on geographic objects and features from different scopes are found in disparate sources. In the sections below, we describe the data types and the sources to obtain information about our target object and the AOI described in the previous sections.

2.3.1. Archaeological Data

Archaeological analysis employs a range of data types, including domain literature, RS data, images, and field observations. The analysis of these data types is influenced by various interrelated factors, such as methodological, observational, contextual, and semantic biases, which can impact the outcome [90]. So, *"The archaeological record is constructed by archaeologists, and patterns within that record will inevitably reflect the activities and interests of archaeologists as much as any reality in the past. This is particularly apparent in the ongoing process of the interpretation and re-interpretation ..."* [91, p.15]. This perspective highlights how knowledge is inherently subjective to continuous revision and argumentation [92]. Researchers unfamiliar with the described taxon and related literature may find it challenging to interpret due to semantic ambiguities in terminology, which can vary by taxonomy, author, and time frame [93].

As evolving interpretations shape records, Cultural Heritage (CH) geographic objects such as sites change over time. Dynamic concepts, such as time-sensitive changes, are integral to these processes [94, 95]. As new findings emerge or the condition of cultural assets shifts, descriptions are updated to create new data versions that maintain the continuity of the object's history while incorporating these evolving interpretations [96]. Alternatively, new records may be produced separately, resulting in different documents about the same cultural asset. The process generates extensive data, often organized to meet specific research needs and stored in institutional repositories or the cloud. Additionally, records are frequently presented as unstructured text across various sources, impeding their findability and accessibility [97, 93]. The lack of standardization in records renders the data non-interoperable, and when combined with proprietary systems, which often restricts access to the data, it further complicates data reuse. [98, 99, 92, 100].

The lack of standardization in records results from the fact that, for a long time, knowledge has been created or maintained by different individuals and institutions with different objectives within a framework of varied knowledge and understanding [101, 102]. Some institutions have dedicated efforts to structuring data. For example, the Portuguese database Endovélico —- Archaeological Information and Management System — managed by the Direção-Geral do Património Cultural (DGPC), started as a manual inventory that began to be digitized at the end of the twentieth century and now has more than 35,000 archaeological records registered by different experts over the past 40 years [103]. Since 2013, the digitalization of technical and scientific archaeological reports has been underway at DGPC to increase transparency and equality of access [103, 67]. Data from these collections is available through *Portal do Arqueólogo (PA)* [8] — a digital platform intended for professionals and researchers in archaeology. Few European countries have repositories with the necessary expertise and mechanisms to ensure archaeological data is freely and openly available for future research [98]. In Portugal, the PA serves as a valuable source of digital information, alongside other useful resources, such as books published by local municipal councils, like those from Mora — Carta Arqueológica (CA) [73].

However, the information is semi-structured and not interoperable, making it difficult to integrate and analyze across different systems; for example, to use this information to understand the structure of a monument, it would be necessary to make the data granular and extract details from the text, such as size, shape, dimensions, and other relevant attributes.

Different researchers have been highlighting the need to make data findable, accessible, interoperable, and reusable to enhance its usability and integration across various platforms and systems [104, 92, 102].

2.3.2. Geospatial Data

Geospatial data is obtained through RS, where information about the Earth's surface is acquired using sensors without direct contact with the objects being studied [9, 105, 106].

This process typically involves satellites, manned aircraft, and Unmanned Aerial Vehicles (UAVs) that measure the electromagnetic radiation emitted or reflected by ground objects. The data collected varies according to these objects' physical and chemical characteristics [106, 107].

There are numerous benefits to using RS data in archaeology: as a non-invasive technique, it preserves archaeological heritage without affecting sensitive objects directly, while it provides a bird's-eye view of archaeological sites within their broader landscape context, revealing the intricate interaction between natural and human-made elements [108]. Archaeological remote sensing relies mainly on the active-sounding technique known as Airborne Laser Scanning (ALS) and passive air and spaceborne imaging in the optical spectrum. While airborne imaging remains the preferred approach for detailed analysis, satellite reconnaissance is well suited for mapping extensive landscape features and detecting monuments in challenging environments [109]. The ability of satellite imagery to cover large areas and provide broad contextual information highlights its significant advantage in archaeological research.

Satellite Images

Satellites offer a unique category of RS platforms, distinct from aerial vehicles such as human-piloted aircraft and UAVs. These artificial satellites collect and transmit electromagnetic data by orbiting the Earth or other celestial bodies, generating images of large areas accessible at various resolutions [106, 17]. The four main resolution characteristics in sensors are spatial, temporal, spectral, and radiometric [109, 106, 9].

Spatial resolution indicates an instrument's ability to distinguish between neighbouring objects, with higher resolution allowing for more detailed images. Radiometric resolution refers to the sensor's ability to detect differences in intensity related to the bit depth. Temporal resolution indicates how frequently an imaging system revisits the same area, which is critical for tracking changes over time. Spectral resolution measures a sensor's ability to distinguish wavelengths of light [109]. Images can be panchromatic, capturing data from a single spectral band at high spatial resolution, multispectral, capturing data from several discrete bands at moderate spatial resolution, or hyperspectral, which captures data from many narrow, continuous bands with high spatial resolution [109, 110, 106, 111]. Further, images can be fused, for example, high-resolution optical panchromatic images can be combined with low-resolution multispectral images to add color and spectral information, a procedure known as pansharpening [15, 39].

Since the launch of Sputnik 1 in 1957 and the capture of the first image by Explorer 6 in 1959, satellite technology has profoundly enhanced Earth observation capabilities [112]. Today, various satellites serve various purposes, including scientific research, meteorology, and imaging [110, 112]. As a result, there is an ever-increasing amount of data in the environmental sciences and cartography, which may contain archaeological information [108, 111].

Raster and Vector

Satellite imagery can be found in different catalogues maintained by major agencies and regional Earth observation initiatives like the European Space Agency (ESA) [113], the National Aeronautics and Space Administration (NASA) [114], and the United States Geological Survey (USGS) [115]. Additionally, platforms such as Google Earth offer interactive satellite imagery through an intuitive interface, enhancing exploration and visualization [7]. This platform is widely used for applications such as archaeological site detection, landscape analysis, and monitoring changes [48, 32, 47, 49]. While its accessibility and ease of use make it popular among both casual users and professionals, Google Earth has limitations for large-scale quantitative research due to inconsistent image quality, lack of metadata, inadequate measurement capabilities (e.g., the spectral analysis), and limited analytical tools [48].

The satellite imagery is typically stored as raster data and can be integrated with vector data [30]. Raster and vector datasets are two geospatial data types [116, 111]. Raster data consists of grid cells (or pixels), such as satellite images [106]. In contrast, vector data comprises distinct geometric entities representing discrete objects from the real world with high spatial determination, such as rivers, elevations, soil types, and archaeological monuments. This representation provides a compact representation of real-world features within Geographic Information Systems (GIS) [117]. In this, geographic features are illustrated using one of three types of geometries: (i) points, which depict individual geographic locations; (ii) lines, used for linear features such as rivers; and (iii) polygons, which describe enclosed areas like islands [116]. Each geometric entity can store associated attributes, providing detailed descriptive information about spatial features. This data type offers advantages such as lower data volumes, improved spatial resolution, and the preservation of topological relationships, which enhance the efficiency of operations like network analysis [111].

Various platforms and agencies create and provide access to digital vector data. At the national, regional, or local level, the national agencies responsible can make vector maps available. These maps generally provide more granular information about specific areas, though the level of detail may vary depending on the mapping project's region, scale, and purpose. For instance in Portugal, the Sistema Nacional de Informação de Ambiente (SNIAmb) [118] and the Sistema Nacional de Informação de Recursos Hídricos (SNIRH) [119] offer hydrographic data, while the Direção-Geral do Território (DGT) [120] and Direção-Geral de Agricultura e Desenvolvimento Rural (DGADR) [3] provide cartographic information on land use and land cover. Most of these cartographic datasets can be accessed through the online portal of the Sistema Nacional de Informação Geográfica (SNIG) [121]. It allows users to share, search, and access geographical information produced by both public and private entities in Portugal. Many of the datasets on this platform are available for public access without restrictions under the CC-BY-4.0

license. Other platforms, such as the Sistema de Monitorização da Ocupação de Solos (SMOS) [122] – e.g. viSMOS, COScid, and COSvgi portal provide a quick and easy way to view VTMs from DGT and satellite images.

The general mapping methodology consists of a systematic interpretation, fieldwork, image analysis, and map preparation. It emphasizes the integration of Remote Sensing Images (RSI) with ground observations to ensure accurate mapping [123]. This approach can generate vector maps reflecting the real-world conditions and features of the mapped area. These maps contain fine-granule details and quantitative representation of the Earth's surface and its natural and artificial features. On such a map, the features are labelled, and they integrate multiple elements (e.g., features differentiated by colour and symbols, labels for feature names, and contour lines showing the terrain changes) to provide a comprehensive view of the terrain [54].

CHAPTER 3

Literature Review

3.1. Introduction

In archaeology, the identification of monuments involves inspection, which includes intrusive (e.g., ground surveys) and non-intrusive prospection approaches (e.g., remote sensing techniques like image analysis) to find and study archaeological monuments [124]. Image interpretation is one of the methods used to extract information from Remote Sensing Images (RSI) [123].

The increasing availability of satellite imagery has created overwhelming data for manual image interpretation [125, 126]. However, modern land use practices and natural changes over time are altering landscapes [89, 50]. While many archaeological sites have been identified, others remain undiscovered and risk disappearing due to natural decay or human activity [34]. These challenges have rendered traditional manual image analysis methods insufficient. In response, automated classification systems are being implemented to accelerate data analysis and archaeological discovery, aiding in the protection of these sites [19, 40, 20, 23, 49, 31, 24, 43, 35, 33, 16, 42, 30, 34, 32, 50, 29, 36, 46].

Despite the progress made with automated approaches, there is still a wide gap between humans and machines regarding learning. Automated approaches can find patterns using training data. Nevertheless, those approaches are not taking advantage of the vast amount of existent background knowledge [127]. Since images are reflections of the world, exploiting background knowledge can be helpful and enrich image interpretation [127]. Integrating knowledge systems with data-driven approaches to add context to scenes via semantic networks has been considered a promising direction to explore [128, 6, 129].

In this chapter, we analyse two distinct lines of research: one focused on object detection methods for identifying monuments in RSI (3.2), and the other on using a Knowledge Graph (KG) to represent and contextualize the existing information on the monuments targeted for detection (3.3). There is extensive research in each of these areas, but combining KGs with object detection is, as far as we know, a more recent development and, we believe, with limited application in archaeology.

For describing the state of the art in this thematic, a literature search was conducted using Scopus and Google Scholar, with keywords such as: "knowledge graph" AND "semantic" AND "image" AND "contextualization" AND "archaeology" AND "site" along with terms related to "machine learning" OR "remote sensing" OR "scene understanding". We then broadened our search to include references cited by the authors in the identified papers, ensuring a comprehensive exploration of the relevant literature.

3.2. Archaeological Object Detection Methods

Object detection is a computer vision task that involves identifying and localizing specific objects within an image [130]. Typically, algorithms identify the objects of interest, draw bounding boxes around them, and classify them into categories. The primary goal of using automated methods to recognize archaeological sites in RSI is to reduce manual labour, standardize analysis processes through replicable workflows, and increase the likelihood of successfully detecting archaeological sites in large areas [30, 24]. The focus is on developing techniques that enable machines to analyze and interpret visual data.

Object detection approaches are fundamentally data-based and often focus solely on the visible traces of extant or subsurface structural remains [127], relying heavily on the quality and quantity of the data as input to perform accurate recognition and classification [23, 16]. The classification is based on initial knowledge, statistical information, patterns, or a combination of these [11]. The most common approaches include pixel-based [131, 38, 45, 39, 47, 23, 49, 43, 35, 33, 16, 42, 32, 132, 37] and Object-Based Image Analysis (OBIA) [22, 38, 40, 46, 36, 20, 21, 31, 24, 29, 30, 50].

Pixel-based approaches involve assigning each pixel, or group of pixels, to a specific target class based on their values [123]. These methods rely on the separability of classes and establish relationships between pixel attributes to form relevant features for classification [11]. These techniques effectively distinguish objects based on differences in reflectance between the pixels corresponding to the target object and those of the background [35, 45]. However, pixel-based analysis often performs poorly in heterogeneous environments with mixed vegetation and soil contrasts [133, 38]. Low contrast between archaeological features and the background may be responsible for this, as well as image noise [134, 47, 43].

In contrast, OBIA starts with image segmentation, which groups pixels into meaningful objects rather than classifying individual pixels [6, 33]. OBIA incorporates additional components such as nearest neighbour classifiers, expert knowledge, and feature space optimization [135, 136]. This method considers shape, texture, and morphology, bridging the pixel world with the vector world [125, 33]. Both pixel-based and OBIA methods have proven successful in archaeological applications, each offering unique advantages depending on the specific requirements of the analysis.

The rapid advancements in Remote Sensing (RS) technologies and computer vision have significantly enhanced the potential for automated detection and classification of archaeological sites above or below ground level [137, 38]. According to Cheng and Han (2016), data-based approaches for object detection in remote sensing images can be classified into five main categories: Template Matching (TM), OBIA, Machine Learning (ML), Deep Learning (DL) and Knowledge-Based (KB) [138]. Table 3.1 presents the methodologies used for detecting archaeological monuments, as well as the data types and target objects identified in each case. It's important to note that these methods are not mutually exclusive. In fact, combining different techniques within the same project can enhance

outcomes and help determine the most effective approach for each specific case [40, 31]. For example, OBIA has been used together with ML [40] or TM methods [31], with promising results. These five main categories are detailed in the following subsections.

Ref	Methodology	Data	Goal
[131]	MV + ML/RF (PB)	STRM/SI	Identify Mounds (tells)
[22]	TM (OB)	LIDAR	Identify Burial Mounds
[38]	ED + KNN (PB/OB)	SI	Identify archaeological features
[45]	TM (PB)	SI	Identify Burial Mounds
[39]	ML/k means (PB)	GE	Identify looted areas
[47]	TM/CHT (PB)	GE	Identify tops of qanat shafts
[40]	GEOBIA + ML (OB)	SI	Identify buried remains
[46]	DL/CNN (OB)	SI/Aerial	Identify ruins of enclosures
[36]	Sup/Uns ML (OB) - HCAL	SI	Identify damage in sites
[20]	GEOBIA (OB)	LIDAR	Identify Barrows
[23]	ML/RF (PB)	LIDAR	Identify burial mounds
[49]	DL/CNN (PB)	GE	Identify Barrows
[21]	GEOBIA (OB)	LIDAR	Identify Barrows
[31]	GEOBIA + TM (OB)	LIDAR	Identify Mounds and shell rings
[24]	DL/R-CNN + CS (OB)	LIDAR	Identify hitherto unknown sites
[43]	ML/SVM:RF (PB)	SI	Identify farm communities sites
[35]	TCT + PCA Matching (PB)	SI	Identify Buried remains
[33]	HBE+ML/SVM+OBIA (PB)	SI	Predict cultural deposit location
[16]	DL/CNN (PB)	SI	Identify qanats
[42]	ML/RF (PB)	SI	Identify Mounds
[29]	DL/R-CNN (OB)	LIDAR	Identify various sites
[30]	DL/R-CNN (OB)	LIDAR	Identify hollow roads
[32]	PCA/LISA (PB)	SI	Identify ancient roads
[50]	HDBSCAN + PCA	Bing	Identify funerary monuments

TABLE 3.1. Methodology, data, and goals for RSI-based identification of archaeological monuments [a]

[a] Abbreviations: **i) Methodology**: CHT (Circular Hough Transform), CS (Citizen Science); ED (Edge Detection); HCAL (Hierarchical Categorization And Localization); HBE (Theoretical Model From Human Behavioral Ecology); HDBSCAN (Hierarchical Density-Based Spatial Clustering Of Applications With Noise); KNN (K-Nearest Neighbours); LDA (Linear Discriminant Analysis); ML (Machine Learning); MV (Morphometrical Variables); OB (Object-Based); PB (Pixel-Based); PCA (Principal Components Analysis); RF (Random Forest); Sup (Supervised); SVM (Support Vector Machine); TCT (Tasselled Cap Transformation); TM (Template Matching); Uns (Unsupervised). **ii) Data**: LIDAR (Light Detection And Ranging); SI (Satellite Imagery), GE (Google Earth).

3.2.1. Template Matching

In object detection, TM, is a straightforward and widely used technique in computer vision, particularly effective in archaeology, when monuments have distinct geometric

shapes like circles or rectangles, which are rare in natural landscapes [138, 24]. This method involves creating a template of the target object, manually or from existing data, and searching for it across an image by adjusting the template's orientation and position. Its simplicity and effectiveness in detecting consistently shaped objects make it a valuable tool for automated archaeological site detection [22, 45, 25, 26, 31].

Despite its advantages, template matching can struggle with variability and complexity in real-world environments. Variations in shape, size, orientation, and factors like noise and overlapping features can complicate accurate detection [138, 24].

3.2.2. Object-Based Image Analysis

Unlike TM, which focuses on detecting predefined shapes within images, OBIA involves segmenting images into meaningful objects or regions, allowing for the analysis of complex and variable object features. When specifically applied to geographic data, this technique is referred to as Geographic Object-Based Image Analysis (GEOBIA) [139]. Since the early 21st century, GEOBIA has gained popularity, leading to the development of numerous applications and methods [6]. This approach leverages object-based methods in various subfields, such as feature extraction, often in combination with statistical algorithms, to achieve good detection results [20, 40, 31].

3.2.3. Machine Learning

Object detection approaches using ML are becoming increasingly popular, treating it as a classification problem to improve analysis and data management. These methods can be divided into supervised and unsupervised techniques [140, 141].

In supervised approaches, features are selected from labelled data to train a model using domain expertise to fine-tune a learning algorithm. This allows for precise recognition of objects within a defined feature space [142]. In contrast, unsupervised ML explores unlabelled data to identify patterns and group similar objects, usually employing clustering methods. Unlike supervised methods, unsupervised approaches do not require predefined classes, making them helpful in discovering unknown relationships and needing less prior data [11, 40, 36]. Supervised learning is typically preferred for satellite data analysis, where the goal is to detect specific objects, among many other features.

3.2.4. Deep Learning

Recent developments in object recognition include Deep Learning methods, a subset of ML, which has advanced the field by providing more precise results [143]. Deep learning uses artificial neural networks to perform complex computations on large datasets, enabling machines to learn patterns and features from examples. These networks are composed of artificial 'neurons' organized into layers: input, hidden, and output [144]. Popular DL algorithms includes Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN)s [143]. The latest applications of DL for detecting archaeological sites in RSI often involve CNNs combined with other methods [49, 24, 27, 46, 140, 16, 29, 30].

CNNs are particularly popular in image classification. This task becomes increasingly complex when an image contains numerous objects, such as in satellite images. This leads to challenges such as occlusion, where objects obscure one another, target blur, which involves a loss of clarity in objects, and issues related to rotation, scale, and complex backgrounds. Additionally, there can be occlusions between objects and their backgrounds [54].

A CNN encapsulates two main stages: feature extraction and classification. In feature extraction, convolutional layers apply filters (kernels) to the input image to detect patterns such as edges and textures. Pooling layers follow, reducing the size of the extract feature maps while retaining the most relevant information, typically using max pooling. In classification, fully connected layers learn to map the extracted features to class labels, often using a softmax function, allowing CNNs to generalize from labelled data [143, 144]. Learning occurs by adjusting connection weights employing techniques like backpropagation scores based on the error between predicted and actual outputs [143].

The CNN architectures can be classified into single-stage and two-stage detectors [145]. Single-stage detectors, such as You Only Look Once (YOLO), streamline the detection process by predicting class labels and bounding boxes in one step, resulting in faster performance but often with reduced precision [145, 146]. On the other hand, two-stage detectors, including Faster Region-Based Convolutional Neural Networks (R-CNN), employ a more complex approach by first proposing regions of interest through a Region Proposal Network (RPN) and then classifying these regions and refining the bounding boxes in a second stage to detect objects [147, 130, 146]. This method achieves higher performance and less False Positives (FP)s, making it particularly effective in detecting archaeological features in RSI [19, 24, 30, 29].

Various studies have demonstrated the effectiveness of CNNs in archaeological detection, showcasing their strengths and limitations [24, 46, 27, 45, 28, 19]. However, a key critique of these models is their "*black-box*" nature, where outputs are difficult to interpret, and they often require large amounts of labelled data to perform well [129]. This is particularly challenging in cultural heritage, where training samples are frequently limited [148, 30, 149]. Strategies like transfer learning [143], and data augmentation are commonly employed to reduce the need for extensive labelled data and extend the use of DL to fields previously constrained by smaller datasets [108, 16].

However, these techniques do not address the challenge of explainability in model outputs. Semantic technologies offer a promising solution for enhancing the understandability and interpretability of results, providing human-understandable insights into how specific outcomes are achieved [60].

3.2.5. Knowledge-Based systems

"A holy grail of computer vision is the complete understanding of visual scenes: a model that is able to name and detect objects, describe their attributes, and recognize their relationships" [150, p. 34]. Consequently, different approaches focus on adding semantic

information to data-based object recognition systems in RSI [150, 151, 152], namely for object-detection-based approaches. Incorporating contextual and semantic knowledge into these processes can improve object detection by considering complex relationships between observed properties [127, 153].

Classification/detection using knowledge concerns on how information is represented and formalized to be integrated with data-driven approaches [128], thus combining the strengths of knowledge-based and data-based methods [154]. It also works as a bridge where knowledge offers explainability in these models [60]. This integration aims to improve the interpretation of results and performance in handling complex tasks [60, 155]. The goal is to use context to enhance the interpretation of the detection of objects by traditional object detection methods, that typically return a bounding box that indicates candidate objects within an image. Knowledge-based models leverage this output by mapping detected objects to corresponding nodes in a scene graph — a structured representation that captures the semantic and spatial relationships between objects and their attributes within the scene — integrating object detection with relevant contextual data [156, 157], thus integrating high-level semantic information with low-level numerical data from images [33, 139].

In knowledge-based object detection methods for RSI, Cheng & Han (2016) highlight the use of geometric and contextual information, widely used as prior knowledge. These approaches generally translate object detection into a hypotheses-testing problem by establishing various knowledge sets and rules [138]. Rule-based knowledge representation is a method that articulates knowledge and features through structured "if-then" statements. They rely on predefined patterns or fixed knowledge that require specialist interpretation, who then convert them into rules or guidelines for analysis. Experts frequently encode their visual perceptions into symbolic classification rules. However, this knowledge remains mainly implicit as it is often applied directly based on trial and error or domain expertise, that is, without formalization [128, 55].

A variety of approaches have been developed to add semantic information to data-based object recognition systems [150, 151, 152], with these hybrid approaches presenting promising developments for remote sensing and archaeology [128, 6, 129]. This is suitable for expressing relational knowledge, associating concepts with entities and facilitating information sharing through standardized vocabulary and semantics [158, 129]. Given that knowledge represented through ontology's and KGs is a recent trend in semantic networks [129], their application to contextualize scenes in remote sensing is also an emerging approach. Different hybrid approaches combine KG-ML techniques to improve the recognition of objects/Areas of Interest (AOI) in satellite images [159, 160, 161, 152, 162]. Other authors are using hybrid approaches for Visual Question Answering (VQA), eXplainable Artificial Intelligence (XAI) [60, 17], information retrieval [159], Semantic Image Contextualization (SIC) (semantic Referee [160]), semantic image classification [163, 162], semantic segmentation [152], and land use/cover [164].

Building on these approaches, using semantic knowledge with object-based image analysis can enhance decision-making by integrating relevant background information and situational factors, refining object detection processes. This type of model provides information, although abstract, that can help contextualize scenes [127]. However, to the best of our knowledge, KG-ML approaches have not yet been used to provide context to RSI for archaeological object detection.

3.2.6. Considerations

Data-based approaches — TM, OBIA, ML, and DL — share processes, including feature extraction, pattern recognition, and classification, and transform higher-dimensional vector spaces into lower-dimensional vector spaces [165]. Feature extraction involves identifying and isolating relevant information from images [11]. Pattern recognition is a set of techniques that makes use of *'feature extraction, discriminant analysis, principal component analysis, cluster analysis, neural networks and image processing to search for data with a set of predefined characteristics'* [22, p.245] [144]. Classification then assigns objects to specific categories based on their extracted features and recognized patterns [143].

Automatic approaches for detecting archaeological monuments from RSI began over 30 years ago. However, they saw a decline in research interest until recent advancements in computational power and improvements in aerial and satellite imagery quality revitalized the field [166]. Since the early 21st century, automation in archaeological detection has achieved significant success, with new methods, particularly those based on ML and DL, greatly enhancing the precision and efficiency of identifying archaeological features [39, 40, 36, 23, 43, 42, 46, 49, 24, 16, 29]. It is fair to say that computer vision systems are being driven by enhanced computer processing power. Together with increasing image resolution levels (either spatial, spectral, radiometric, or temporal) and faster availability of data, technological advances are ensuring greater use and acceptance of automation in image analysis [167, 14].

Despite successful applications, most of the object-based archaeological site detection research has focused on relatively simple and uniform structures, as highlighted in [166, 6]. Researchers have focused on using these methods for recognition based on geometric features such as circular shapes [131, 22, 45, 47, 39, 39, 20, 23, 49, 31, 16, 29, 50] to identify structures like mounds [131, 22, 45, 47, 20, 23, 49, 31, 24, 42, 42, 29], linear shapes [38, 39, 40, 46, 30] or rectangular [40, 46] to detect features such as roads [30, 32] and walls [38, 39, 40, 46], and other patterns to recognize landmarks and significant terrain features [43, 35, 33, 42]. Sites such as mounds are generally easier to detect because they exhibit consistent features and can be defined with a limited set of descriptors and parameters (e.g., those presenting a circular, standardized pattern with dimensions $\geq$5m diameter) [6]. However, automated approaches usually return a high number of FPs [30, 19, 16, 97]. Barrows, in particular, are detected with a degree of uncertainty, and their identification is only considered accurate after reviewing common FP associated with manual and automated detection methods [21, 20]. The manual evaluation of data

returned by these approaches is highly time-consuming, limiting the amount of test data that can be thoroughly analysed [168].

In satellite imagery, detecting small objects is particularly challenging due to low pixel resolution, which can lead to a loss of critical details that make it difficult to distinguish these objects from the background [146]. Furthermore, the large scale of these images, often consisting of hundreds of millions of pixels, complicates the task of separating the target from the background in complex terrains [46, 108, 146]. Other factors can hinder recognition, notably the characteristics of the site itself, such as the monument's size, its state of preservation [21, 23] and resolution/scale/time of the dataset used [23, 40]. These factors complicate feature extraction and parameter selection, making it challenging for automated systems to recognize ancient sites.

Detecting small-scale cultural heritage sites, such as tombs in RSI, is even more difficult, specially when the construction material of the structure matches the surrounding landscape [20, 6]. Their spectral similarity to surrounding imagery, combined with erosion and collapse over time, indicates that new methods are needed to effectively detect these objects that blend into their environment.

3.3. Knowledge Bases in Archaeology

Data-based approaches focus on recognizing visual features without considering whether the found solution is semantically correct or leveraging the semantic advantages associated with geo-data [129]. For example, neural-based algorithms are trained by adjusting model parameters to minimize a cost function over the data [143]. These models excel in perceptual tasks, such as image classification, but often struggle with more complex cognitive tasks, like understanding and interpreting the deeper meaning or relationships within an image [24]. These cognitive tasks involve reasoning, not just recognition. To be successful at cognitive tasks, models need to understand how objects interact and relate to one another [128, 150].

In contrast, KB approaches address this limitation by using or integrating feature information, namely details on the characteristics of the objects to be found or on the site where they are located. In literature, structured rule-based interactions are often used. These approaches leverage information such as spectral features (e.g., spectral indices) and environmental knowledge (e.g., precipitation, temperature, topographical features, phenological stages) to inform model decisions [46, 33]. Rule-based approaches are one of the simplest forms of machine-understandable expressions [129], by which domain-specific knowledge is represented in the form of rules.

Conversely, knowledge-driven approaches based on semantic networks to interpret RSI are considered one of the most promising directions [128, 6, 129]. A well-structured semantic database is essential for hybrid approaches to provide context to images. As

indicated by [129], compared with rule-based systems, semantic networks excel in constructing organic relationships between complex entities and characterizing spatial distribution and relationships. Ontologies and KGs are modern implementations of semantic networks [129].

As noted by Magnini et al. (2019) [6], there is a growing urgency in establishing standardized methods for publishing semantic networks that can be semi-automatically or automatically applied in archaeological investigations. They proposed a theoretical ontology-based framework to formalize expert archaeological knowledge. Their approach introduced the Diachronic Semantic Model (DhSM), designed to explain long-term landscape evolution and applied it to a data-driven approach using OBIA for archaeological predictive modelling, incorporating DEM-based techniques to identify areas for human occupation and territorial control based on info such as slope, local dominance, and solar radiation [6]. This example highlights how landscape information can be valuable. However, the use of semantic approaches to support object detection in satellite data, particularly in archaeology, is still in its infancy, emphasizing the need for well-structured semantic models.

A domain knowledge component for the interpretation of an image can include (i) the real-world instances, consisting of tangible entities (e.g., barrow X, castle Y, etc.); (ii) the conceptual domain, which represents these entities based on expert knowledge; (iii) the digital domain, associated with the virtual representation of the instances; and (iv) the spatiotemporal domain, indicating the chronological depth inextricably linked to archaeological research [6] (see Figure 3.1). This involves knowing the structural details of the archaeological sites as well as their relationship with the environment.

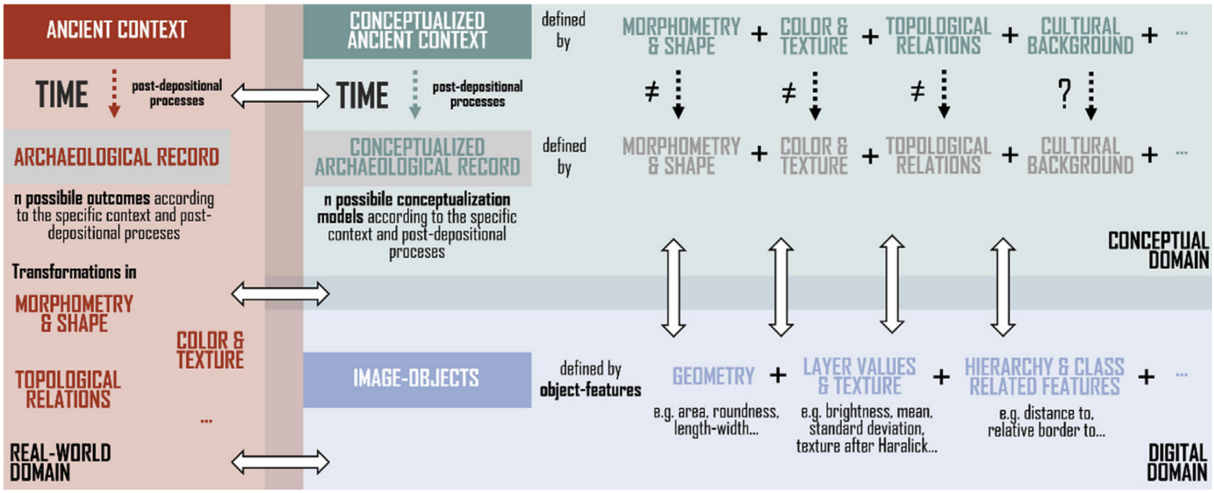


FIGURE 3.1. Graphic presentation of the features to be represented. (Source: [6, p.14])

However, data with archaeological and environmental information to be used as domain knowledge are kept by different communities and institutions and these entities employ conceptualizations and formal definitions that usually do not match [57, 169]. There

has been a traditional division in the management of data about archaeological monuments. Archaeological data, typically overseen by cultural heritage institutions, often emphasizes descriptive and interpretative information. On the other hand, geoinformation has been primarily handled through Geographic Information Systems (GIS), focusing on visual and spatial representation rather than detailed semantic content [57]. This separation is mirrored in the various standards for information representation: Geospatial Consortium (OGC)'s GeoSPARQL [58] — a well-established ontology for representing geoinformation — while CIDOC-CRM [59] (ISO 2117:2023) is a well-established ontology recognized as an official ISO standard for representing cultural heritage data. When these standards are applied together, they have the potential to significantly enrich Cultural Heritage (CH) data by providing precise descriptions of both the site locations and the geometries [57].

Semantic networks are a widely used formalism for the representation of knowledge in archaeology and RS [129, 128]. In the geospatial domains, various Geospatial Knowledge Graphs (GeoKGs) [170] were implemented, such as GeoLink [171], Geonames [172], GeoGraphVis [173], KnowWhereGraph [174], DBpedia [175], Yago2geo [176], Wikidata [177], Event KG [178] and GeoKG [179]. In general, the KGs with geographic information such as DBpedia, YAGO, and Wikidata focus mainly on entity-centric information and may not offer the same level of granularity as national/local Vector Topographic Maps (VTM)s, particularly regarding detailed landscape features and local context. Others focus on solving specific problems, for instance, constructing models and applying them to disaster preparedness and resilience [173, 174].

One notable application using KGs in RSI is narrative cartography, which integrates event-based spatial information to provide semantic context for scenes in a given temporal space [180, 116]. Mai et al. (2022) highlight several advantages of using knowledge graphs for narrative cartography, including improved data acquisition, integration, and semantic content management. Despite these benefits, the use of semantic web technologies for geo-visualization remains underexplored [180]. In some studies, we find KGs, such as LinkedGeoData [181], WorldKG [154] and GeoKG [154], converting OpenStreetMap (OSM) data into an Resource Description Framework (RDF) KG. Others are extracting information from VTMs as historical maps and representing them as spatial-temporal KGs [116]. Semantic representations of vector maps as a KG present inherent challenges due to the large, heterogeneous, and ambiguous schema, and several recent projects are addressing this issue [154, 180, 116, 182, 183].

Rather than focusing on 'Relationships', a cartographic map focuses primarily on 'Entities', where each entity is explicitly defined by its location, geometry, and semantics descriptions [182]. Identifying and partitioning vector data into interconnected geospatial entities — referred to as "building block" geometries — is fundamental for capturing geographic features from vector maps and representing them in a liked way. This approach enables a detailed and efficient representation of features such as railway networks or

wetlands [116]. The process involves tasks like partitioning and entity matching and can be enhanced through schema matching, ontology alignment, and data exchange. Schema matching involves aligning constituent elements of different relational schemas, such as attributes and relation names. Ontology matching focuses on aligning classes and properties, while data exchange requires complex mapping specifications to transfer data between source and target schemas. In this context, a direct concept attribute maps a database column to a concept attribute, also known as a data property or schema entity [184].

Knowledge-based methods for RS data still need standards for storing information to deal with the data flow that is constantly generated [128]. The lack of standardization in terms of collection, storage and interpretation, and the data dispersion, inconsistency and inaccuracy are reasons cited as restricting the use of automated approaches in this field [185, 42]. In archaeology, different organizations have been concerned about data storage methods and management, forming partnerships at both national and local levels to create a comprehensive database for data collection and storage [186]. For example, various national archives (e.g., the Direção-Geral do Património Cultural (DGPC), Sistema de Informação para o Património Arquitetónico (SIPA)) and international archives (such as UNESCO and ICOMOS) present information on archaeological monuments. However, there is no consistency standard for how data is organized and modelled, nor is there a single, centralized database that contains all the data in one place.

To address these challenges, several projects are underway to enhance interoperability among diverse databases [187, 188]. Initiatives such as ARCHES, an open-source information system for heritage inventory and management [189], ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe), which integrates diverse archaeological datasets to enhance research through unified access and advanced technologies [190] that was expanded as ARIADNE Plus to include other public organizations which includes DGPC in Portugal, STAR (Semantic Technologies for Archaeological Research), which achieved semantic interoperability among archaeological datasets [191], OpenArchaeo which is a tool designed to query archaeological datasets in the Linked Open Data cloud [192], EPISA [187], a project that migrates data from the National Archive of Torre do Tombo to a relational database model [188], ArCo KG of Italian CH [96] and CARARE [193] connecting archaeology and architecture. These projects provide digital infrastructure for research data by creating interoperable datasets and representing the datasets in a linked way, providing information on both a semantic level (e.g., descriptions and interpretations) and a graphical level (e.g., photos and drawings).

Many of the semantic network representations for archaeology are compatible with CIDOC-Conceptual Reference Model (CRM) [189, 190, 191, 192, 187, 194, 188, 96, 155, 195, 196, 197, 198, 199, 200, 193, 201]. There is an extensive body of literature demonstrating the benefits of the CIDOC-CRM in representing building and architectural heritage in archaeology [189, 202, 200, 193, 201, 169]. For instance, Hansen & Fernie (2010) describe the CARARE metadata schema, which focuses on the record of a detailed description

of heritage, events, and digital resources [193]. Carlisle et al. (2013) highlight the benefits of incorporating CIDOC-CRM standards into the design of Arches [189]. Ronzino et al. (2016) present CRMba, an extension of CIDOC-CRM for encoding metadata about archaeological building documentation [203]. Gergatsoulis et al. (2022) utilize CRM and CRMba to represent archaeological buildings based on fieldwork data, including records, provenance, and images [200]. Santos et al. (2022) focus on representing megalithic monuments at a granular structural level using CRM [201]. Ranjgar et al. (2022) develop a Point of Interest (POI)s-based data model for Iranian heritage sites by integrating CIDOC-CRM with GeoSPARQL to merge spatial semantics with heritage information, enabling users to explore and utilize location-based services and applications [169].

The CIDOC-CRM and its extensions have been successfully applied to represent archaeological sites instances [96, 169] in a granular way [201, 200, 189, 204]. Some approaches focus on retrieving information [169, 204]. Others focus on represent spatial data [169] focusing on topological spatial relations [204, 197, 155, 205]. In this case, "internal spatial relations" are explored, which refer to interactions among the structure (e.g., walls with cave roofs) [204] or between the finds in the structure [197]. Most research utilized SQL [189] or NoSQL models, often employing the RDF and Web Ontology Language (OWL) [155, 201, 169, 198], with few of these studies using Native Graph Database (NGDB) to create KGs based on CIDOC-CRMs [197, 204].

3.3.1. Considerations

To the best of our knowledge, there are only a few approaches that integrate CIDOC-CRM with ML [206, 168, 205, 199]. Even fewer apply semantic networks to contextualize images [205] and improve data-based object detection [148, 168]. Ontologies/KGs with data-based approaches to provide explainability to imagery are being applied to retrieve image information [168, 199], improve data-based object detection [148, 168], generate training data [207] and VQA [208]. The usage of KG-ML models to deal with RSI has been introduced recently and is seen as a good method for improving automatic detection in archaeology [6].

Recently several ontologies have been developed. However, spatial and contextual information seem to be discussed separately, as do archaeological and landscape data. A scene involves a complex network of cultural and historical contexts that must be integrated and interconnected to fully capture its significance and provide comprehensive information [169]. This perspective aligns with Tobler's first law of geography, which states, *"everything is related to everything else, but near things are more related than distant things."* [209].

A remote sensing scene consists of features such as land covers (physical states) and land uses (functional objects), which are closely linked [17, 152]. Objects in the topographic space include both landscape and archaeological features. The geographical features considered in RS information extraction are classified into spatial, physical, and regional categories. Spatial knowledge encompasses spatial vision features (e.g. texture),

spatial geometry features (e.g. shape), spatial distribution (e.g. position) and spatial relationship. Physical knowledge contains sensor information (e.g. imaging mode, sensor performance), model (e.g. physical model), and spectral features (e.g. spectral index). Regional knowledge includes temporal knowledge (e.g. land/use change) and environmental knowledge/features (e.g. topographical features) [129]. Although data and knowledge about these elements exist, they are not always organized effectively for integration and utilization.

The domain knowledge of geographical features can be obtained from a variety of sources (As presented in section 2.3). The composition of a scene can be derived from semantic, raster and vector data. Pre-made official regional VTMs containing landscape data are a valuable resource because they already contain derived and processed information, which saves time and effort in extracting and interpreting raw data. These maps provide immediate insights into features like land cover, vegetation, or topography and, contain info on structures (entities) and descriptions (characterization of identified entities) [81]. The info is provided in a structured format which allows for easy information extraction. On the other side, data from archaeological sites is often found in textual or structured repositories. Although this contains valuable information, it can also be incomplete and disjointed due to data incompatibilities and the subjectivity of scientific observations [97]. This information is often fragmented, with separate knowledge representation models dedicated either to landscape information or to archaeological heritage. Furthermore, specific data about monuments and landscapes within a given region are typically stored in isolated sources and are not always consistent.

Although there is data and information about a place, what is missing is the structured integration of this information in a way that can be effectively utilized to leverage KG-ML models. Considering this, we wonder if combining these data into a semantic model could make it interoperable for contextualizing scenes using machine learning methods.

3.4. Conclusion

Upon reviewing studies on automated approaches for recognizing archaeological sites in RSI (Section 3.2), it became clear that most automatic methods being used are data-driven. Additionally, they tend to return FP [21, 131, 45, 25, 27], particularly for complex or small sites [21]. Knowledge-driven approaches that add domain knowledge to provide context to RSI scenes via semantic networks have been considered an important direction to explore together with data-driven approaches [128, 6, 129]. However, these approaches remain largely unexplored in archaeology. Nevertheless, semantic networks are gaining attention for their role in representing and standardizing information otherway dispersed across various sources [210]. There are many ontologies available, as presented in Section 3.3, but to fully leverage them for contextualizing scenes it is crucial to gather the relevant data and information to populate the ontology and implement a model to deal with this information at an instance level. However, data is often scattered, diverse, and sometimes inaccessible, making it challenging to effectively integrate and use.

The difficulty in acquiring diverse datasets poses challenges in the analysis of archaeological sites and their surrounding environments. Data collected from various sources often lack consistency, resulting in different interpretations of the information. Semantic and geospatial data are essential for understanding archaeological and topological features and their relationships. Together, these types of information can enhance domain knowledge about the elements present in a location and help contextualize a scene. We identified two issues: one regarding the different sources of information related to archaeological and topological features, and the other concerning both semantic and geospatial data formats.

The representation of spatial and contextual knowledge as linked data is seen as a promising approach. It has been applied to standardize, improve accessibility and interoperability of data, and enhance the contextualization and interpretation of images [128, 182, 17]. Currently, different approaches are combining KG and ML techniques to improve the recognition of objects/AOI in satellite images [159, 160, 161, 152, 162]. In these works, KGs are applied either at the beginning, to train the model, or at the end to review and refine the results. Thus, KGs act as (1) a Reviewer - validating and refining visual model outputs [160], (2) a Trainee - used to create semantic embeddings to align them closely with visual embeddings [148], (3) Peer, which combines semantic and visual data into a hybrid space for a comprehensive representation [34, 211, 161, 168, 212], or (4) guide the visual recognition where the KG acts as a Trainer [133]. These diverse roles are illustrated in Figure 3.2 and align with the categories defined by Monka et al. (2022) [165]. Most of them, however, are used at the beginning of the process to aid in training and model development, with no attention paid to using domain knowledge in an automated way to validate outputs.

Although hybrid KG+ML approaches are being used [159, 160, 161, 152, 162], they have not yet been used to provide context to RSI for archaeological object detection. Knowing this, a promising direction would be to implement a KG that centralizes available information about the monuments and their surrounding landscape, encapsulating domain knowledge in an interoperable format. This would allow machines to effectively access and utilize information that would otherwise remain scattered and disconnected. To address the gap in archaeology, where semantic databases have not yet been utilized to provide context for RSI scenes and improve object detection outputs, we aim to create an interoperable semantic database. This database will consolidate dispersed information about both the target objects and the AOI. The purpose is to implement aKG to provide contextual knowledge that can be used to refine and improve the object recognition results. By incorporating this contextual information, we aim to train a model capable of enhancing detection and evaluating whether the additional context improves overall performance.

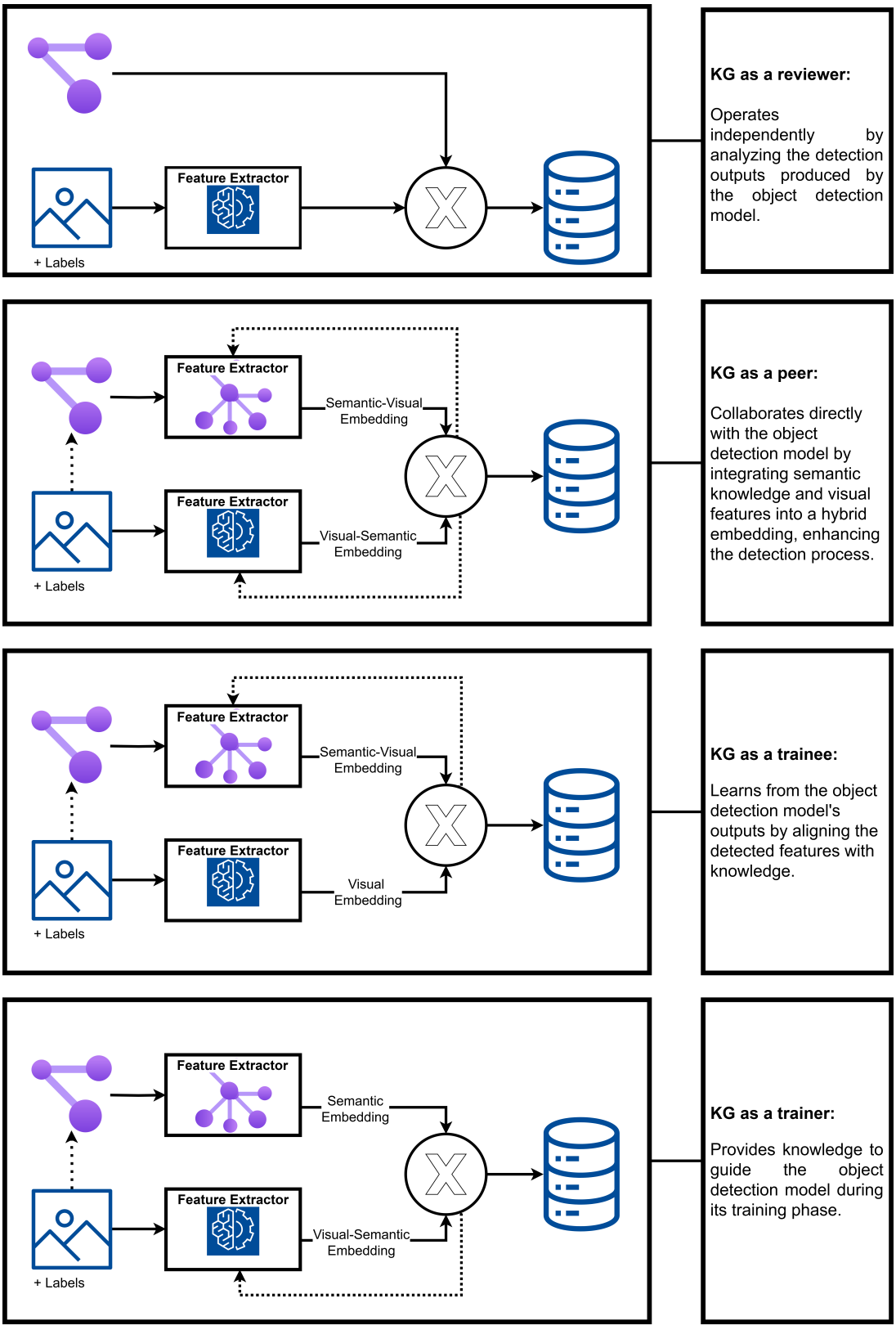


FIGURE 3.2. Diagram showing the roles of knowledge graphs in object detection: reviewer, peer, trainer, and trainee.

CHAPTER 4

Dolmen Detection

4.1. Introduction

Dolmens are one of Portugal's most representative prehistoric cultural features. Object detection for recognizing these monuments in Remote Sensing Images (RSI) in the Iberian Peninsula has already been performed. Specifically, in our study area (Section 2), we have contributed with a sequence of three significant works related to image analysis for identifying dolmens. The first involved photo-interpretation, resulting in the recognition of 60 monuments within the regions of Mora and Arraiolos (Portugal) [18]. This was followed by two distinct automated approaches, both using satellite images over the 60 recognized monuments identified in the previous work as a dataset to train models for automatic monument detection. One was a pixel-based approach that used hyperspectral, multispectral, and panchromatic images from ESA's PROBA-1 and WorldView-2 (WV-2) satellites [37] and the other an object-based approach employing 8K resolution Google Earth images [213]. In the first case, there were many False Positives (FP)s, while the object detection-based model showed better performance, achieving an F1 score of 0.78 and a precision of 0.93 using Convolutional Neural Networks (CNN)s. The test set consisted of 3 monuments. However, it's important to note that as more images of the terrain are analyzed, the number of false positives may increase, especially considering that the surrounding landscape predominantly comprises the same material as our target object.

The problem of false positives remains significant. Other research, such as the one in [20], which used LIDAR data to detect megalithic monuments on the Iberian Peninsula, also resulted in several FPs. Even using different types of data, identifying small-scale prehistoric structures appears to be challenging and often results in false positives.

This chapter details our object detection approach for dolmens in satellite images for this thesis. Given that the object-based approach in [213] showed good precision, despite the small test set, we adopted it as our model and extended the tests. The focus is on our study area (Pavia), where we will use the 16 identified dolmens within this area for testing and the remaining 44, located in the surrounding areas, for training. Additionally, we extended the analysis to use images with no monuments in or testing AOI to assess the model's performance and evaluate if the number of false detections is maintained.

The remainder of this chapter outlines the process for data gathering (Section 4.2), dataset preparation (Section 4.3), and the proposed algorithm (Section 4.4). Subsequently, the implementation section describes how the algorithm was executed (Section 4.5), and the results section presents the outcomes of the experiment (Section 4.6).

4.2. Data gathering

Data for this study comes from previous photointerpretations that identified 60 dolmens from satellite images within the regions of Mora and Arraiolos (Portugal) [18]. These dolmens are shown in Figure 4.1. Knowing the location of this targets objects, the images were sourced from Google Earth at an 8K resolution, captured at a scale of 1:400 meters. For each known archaeological monument, five images were obtained, each showcasing the monument from different positions, resulting in a total of 300 images. An additional 70 images with non-archaeological monuments were collected.

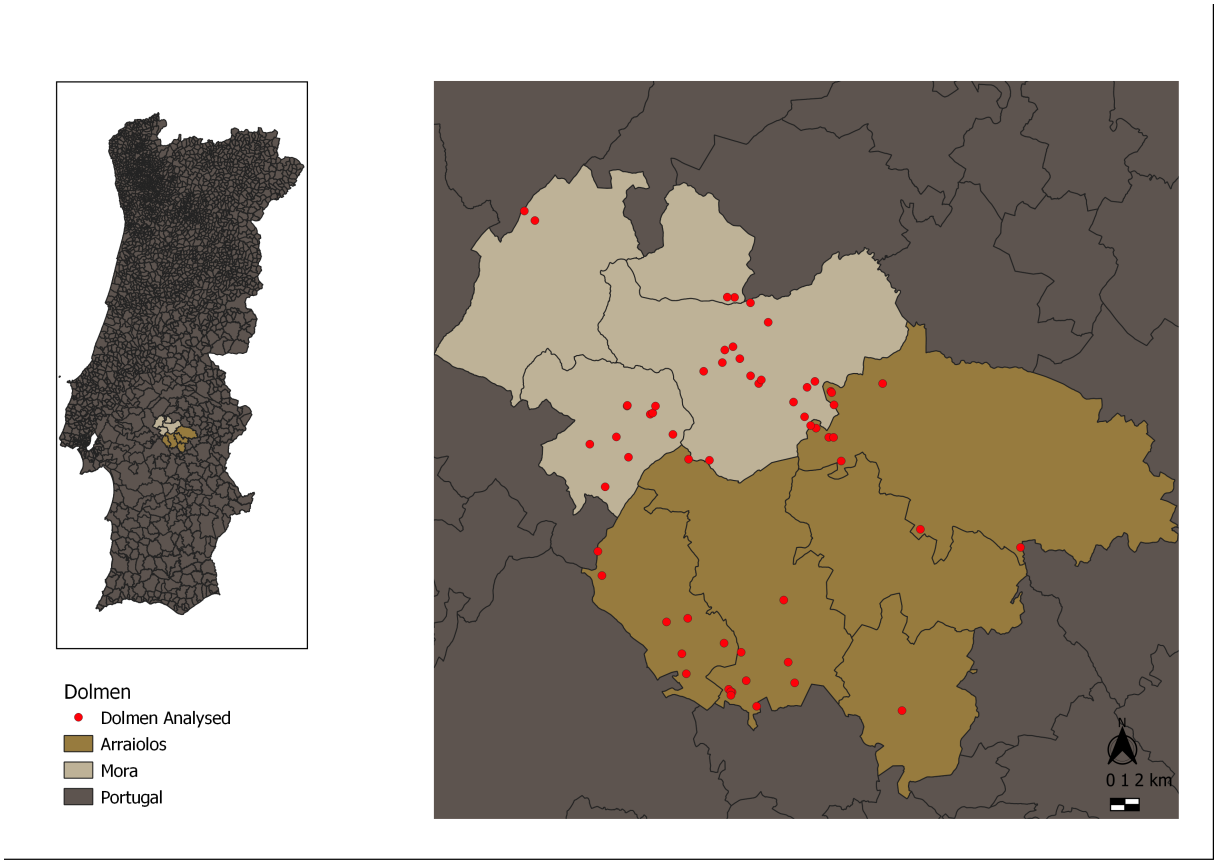


FIGURE 4.1. Map of Portugal showing detailed regions of Mora and Arraiolos on the right, with analyzed dolmens marked in red. The map was created using VTMs [2, 5]

Since its launch in 2005, Google Earth has leveraged geospatial technology to deliver precise location data based on latitude and longitude, allowing users to pinpoint specific locations with ease. As a free and user-friendly platform, it supports the storage and sharing of location data through KMZ files, which package multiple files into a single, compressed format for easier distribution and quicker downloads [214]. Google Earth Pro offers high-resolution images in pan-sharpened format [213], which can be accessed at various resolutions and from different time periods, making it a valuable tool for analysing and visualizing geographic data.

Google Earth provides historical images of the region, spanning from 1995 to the present. Since 2015, each year presents multiple timeframes within the year, varying in

quality throughout the timeline. The quality of the images may not necessarily improve over time but rather depends on which satellite took the photographs. For example, images between 2013 and 2017, especially during August and September, show better resolution four our Areas of Interest (AOI) (see Figure 4.2d). Studies have shown that the optimal time of year for object detection in images is typically during the summer or fall [35, 215]. We chose images from August 2017 because they provided the best resolution and clarity among the available historical images.

(A) August 2011

(B) August 2017

(C) August 2018

(D) September 2023

FIGURE 4.2. All images from Google Earth are in 8k resolution and at a scale of 1:400m. Credits: [7]

4.3. Dataset Preparation

After gathering the data, the process involved preparing the data for training, including enhancing data quality, preprocessing, and model training. This was done using Roboflow, which facilitated data preparation, enhancement, and preprocessing for model training. Roboflow is a platform that simplifies the management of computer vision datasets. The Workspace Image Search API offers a robust and flexible mechanism for searching and labelling images, and an extensive range of augmentation and preprocessing options [216].

Labeling involves marking objects in images. Bounding boxes and polygons can be used to annotate images [216]. A total of 300 images were organized and annotated with binary labels to create the dataset for the project *"Dolmen"* and *"non-dolmen"*. dolmens are a positive class, while images without dolmens are marked as *null* — to indicate the absence of the object of interest.

Once the images were annotated, we enhanced and augmented the dataset. Preprocessing steps involve preparing the data for model training by normalizing images,

resizing them, and ensuring consistency and quality. Image augmentation is a computer vision technique that artificially increases the size and diversity of training datasets by applying various transformations to the original images. These transformations introduce variations that enhance the model's ability to generalize and improve its robustness to new data. Augmentation techniques, such as rotations, flips, and colour adjustments, are essential for enhancing dataset diversity. For instance, cropping involves selecting a portion of the image, with adjustable minimum and maximum ranges. Rotation allows for rotating the entire image by a specified angle in degrees. Hue adjustment modifies the colour hue by specifying degrees, while saturation adjustment changes the image's saturation level in percentage terms [216].

In the preprocessing stage, we applied the following Auto-Adjust Contrast to enhance image quality through adaptive equalization. The auto adjust contrast feature enhances low-contrast images by adjusting contrast based on Contrast Stretching [1], Histogram Equalization [2] and Adaptive Equalization [3] — being this last one used in our dataset. Augmenting training datasets artificially with a variety of transformations is a technique in computer vision that increases training dataset size and diversity. Here, we experimented with different parameters, including replicating the approach indicated in [213], to determine the most effective preprocessing and augmentation strategies.

To augment the dataset and improve the robustness of the model, we applied the following configurations, which yielded the best results, producing three outputs per training example: Crop with 0% minimum zoom and 30% maximum zoom, Rotation between -15° and +15°, Hue adjustment between -25° and +25°, Saturation adjustment between -34% and +34%, and Brightness adjustment between -24% and +24%. In comparison, the replicated approach used the following augmentations: Crop with 0% minimum zoom and 62% maximum zoom, Rotation between -20° and +20°, Hue adjustment between -35° and +35°, Saturation adjustment between -99% and +99%, Brightness adjustment between -55% and +55%, and Bounding Box Exposure adjustment between -35% and +35%. These resulted in a dataset with 943 images.

To properly evaluate the model's performance, the dataset was divided into training, validation, and test sets based on unique objects. The training set includes one annotation file and 867 images (comprising 47 dolmens, or 73%, and 47 null images). The validation set contains 44 images (with 4 dolmens, or 3%, and 14 null images) and one annotation file. The test set consists of 32 images (16 dolmens, or 23%, and 8 null images). For the test set, special attention was given to the 16 dolmens located in the AOI, allowing for a targeted evaluation of the model's performance in identifying these specific objects. This

[1]Contrast Stretching: Rescales the image to cover the full range of intensities between the 2nd and 98th percentiles [216].

[2]Histogram Equalization: Distributes intensity values more evenly across the image, achieving a roughly uniform pixel colour distribution [216].

[3]Adaptive Equalization: Contrast Limited Adaptive Histogram Equalization (CLAHE) enhances local contrast by applying histograms to different regions of the image, improving details in both dark and light [216]

dataset was downloaded in the COCO format [216]. This set is referred to as Dataset 2, and in addition to it, we also utilized the dataset from [213] — referred to as Dataset 1 — which consists of three test images.

4.4. Proposed Algorithm

Fast Region-Based Convolutional Neural Networks (R-CNN) enhances the two-stage detection process by introducing a Region of Interest (RoI) pooling layer, which standardizes the feature maps generated by the backbone network. The process starts with the Region Proposal Network (RPN) identifying potential object regions. Fast R-CNN then classifies these regions and refines their bounding boxes [130]. The entire image is initially processed through convolutional layers to produce a feature map. The RoI pooling layer extracts fixed-length feature vectors from the proposed regions, which are then fed into two fully connected layers: one for box regression (reg) to adjust bounding boxes and one for box classification (cls) to determine object categories [147]. The network outputs classification probabilities and bounding box adjustments, all optimized through an end-to-end multi-task loss function that simultaneously addresses both classification and regression tasks [130].

In training RPNs, the objective is to minimize various loss functions to improve model performance [147]. Several metrics are used to evaluate this process. The *Classification Loss for RPN* measures the log loss over two classes (object vs. not object), assessing the precision of the RPN in classifying anchors as either objects or backgrounds. The *Regression Loss for RPN*, evaluates the error in predicting bounding box coordinates using a robust loss function to refine the predicted coordinates. The *Box Regression Loss* measures the error in the final bounding box regression performed by the detection network [147]. The *Classification Loss for Detection Network* evaluates the precision of object classification within the proposed regions, computed as a log loss over class predictions for detected objects. These metrics are combined into a multi-task loss function that optimizes both classification and regression objectives, ensuring effective training of both the RPN and the detection network [213].

4.5. Implementation

To implement the Faster R-CNN framework, we employed Detectron2 with two backbone networks, ResNet-101 and ResNet-50, and evaluated Dilated Convolutional Network (DCN), configured using the COCO detection pre-trained weights. Additionally, two training schedules, 1x and 3x, were used to explore their impact on detection performance

The model was trained with a maximum of 5000 iterations, a base learning rate of 0.001, and an evaluation period of every 200 iterations. The dataset annotations were visualized to ensure quality, and the output directory was organized dynamically based on the training setup and timestamp. Batch sizes and the number of workers were adjusted

to optimize for the available computational resources. Additionally, the mask format was set to bitmask for compatibility with the dataset annotations

4.5.1. Training Performance

Various training sessions were conducted with different dataset configurations, involving changes to both the training and test sets. This approach focuses on the model's ability to correctly identify dolmens in the area of interest while also accounting for the impact of dataset variations. To better understand the model's performance, the training metrics will be examined, as illustrated in Figures 4.3 – 4.4.

Figure 4.3 presents three graphs depicting the model's performance over time: Classification Accuracy, False Negatives (FN), and Foreground Classification Accuracy. Classification accuracy shows that after approximately 3,000 steps, it stabilizes near 1, indicating that the model correctly classifies nearly all training samples, with minor fluctuations but an overall high accuracy of around 0.9922 at the final measurement. The false negative rate similarly approaches zero by about 2,000 steps and remains low throughout the training. Lastly, foreground classification accuracy stabilizes after approximately 2,000 steps, reflecting the model's effective precision in detecting objects of interest.

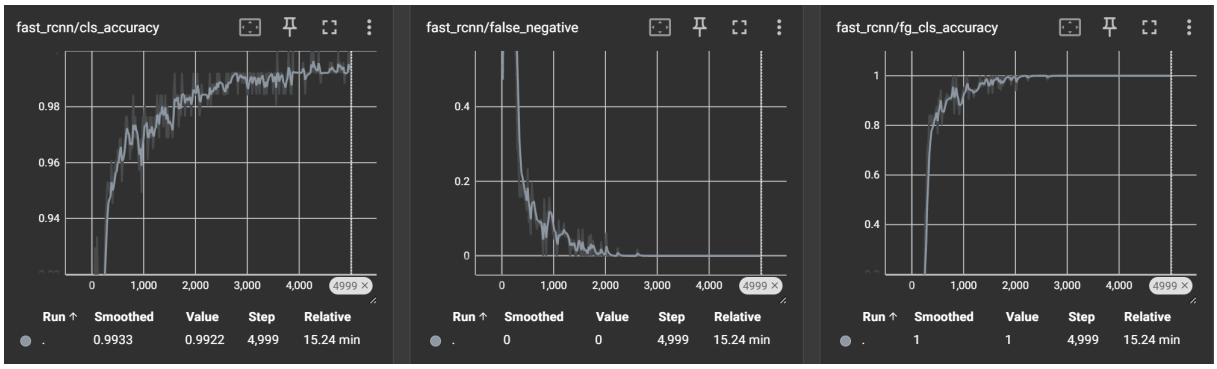
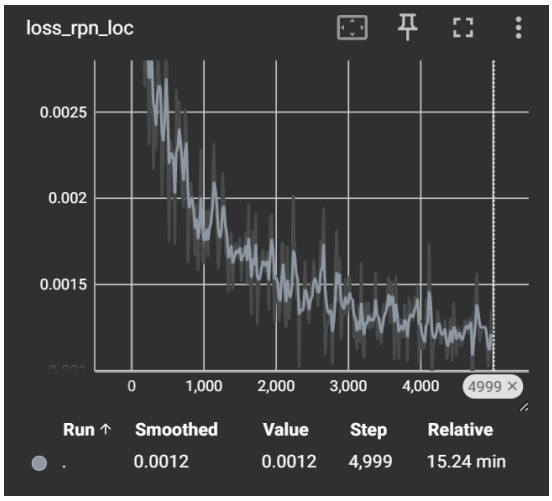


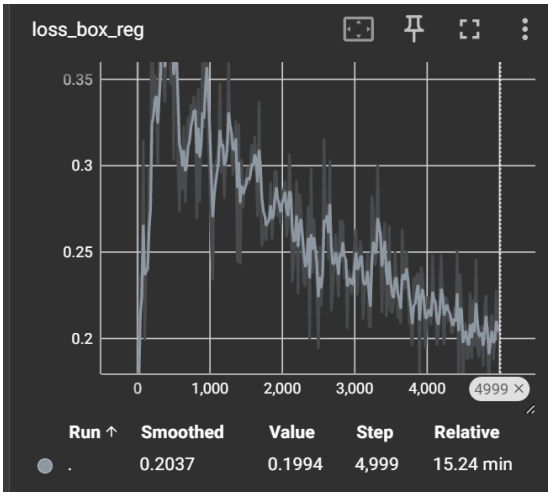
FIGURE 4.3. Training metrics showing classification accuracy, false negatives, and foreground classification accuracy

The evaluation of the model's performance is reflected in several loss metrics. Figure 4.4a shows a low loss_rpn_loc value of 0.0009, indicating excellent performance in predicting bounding box coordinates. Figure 4.4b presents the loss_box_reg value of 0.1632, which, while higher, still suggests reasonable accuracy in final bounding box regression. Figure 4.4c highlights a very low loss_rpn_cls value of 0.0001, demonstrating high precision in classifying anchors as objects or backgrounds. Finally, Figure 4.4d shows a loss_cls value of 0.0091, indicating strong classification performance in detecting objects within proposed regions.

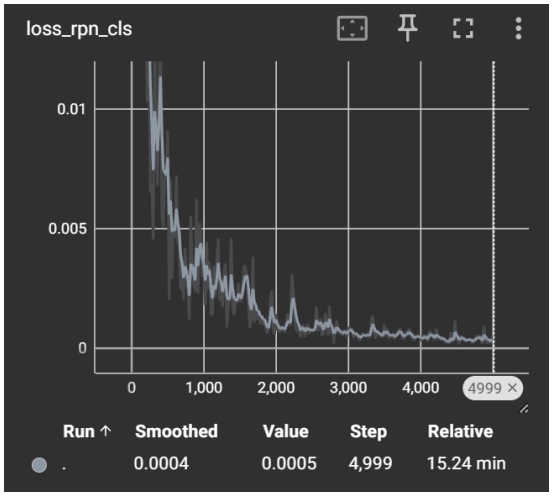
Using Google Collab, each algorithm required approximately 2 hours to train for 5000 iterations or 40 min for 2000 interactions, aiming to minimize total loss and achieve the optimal learning rate for each algorithm. The default GPU for Collab is an NVIDIA Tesla K80 with 12GB of VRAM (Video Random-Access Memory).



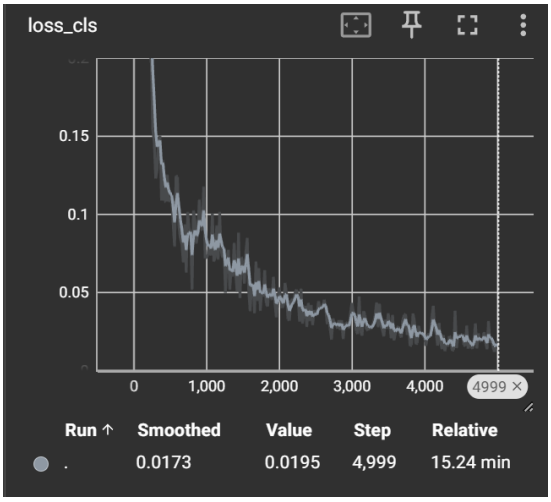
(A) Location Loss in RPN.



(B) Box Regression Loss.



(C) Classification loss in the RPN.



(D) Classification Loss for detection.

FIGURE 4.4. Training metrics are shown: Location Loss for RPN (A), Box Regression Loss (B), Classification Loss in the RPN (C), and Classification Loss for the Detection Network (D).

4.5.2. Evaluation Metrics

When evaluating an object detection model, several key metrics are used to assess its performance, depending on the application's goals. The model's performance is evaluated by counting True Positives (TP)s, where the model correctly identifies and locates objects; FPs, where the model incorrectly detects objects that aren't present; FN, where the model fails to detect objects that are present; and True Negatives (TN)s, where the model correctly identifies the absence of objects. A confusion matrix provides a detailed summary of these metrics. It presents TP, FP, FN, and TN in a tabular format, allowing for a comprehensive view of the model's performance across all categories [141].

These metrics are then used to calculate precision, recall, and F1 scores. Precision, calculated as Precision = *TP / (TP + FP)*, measures the proportion of true positive predictions among all positive predictions made by the model. This metric reflects the robustness of the model's positive detections, accounting for both correctly identified

objects (true positives) and incorrectly detected objects (false positives). Recall, given by Recall = *TP / (TP + FN)*, measures the proportion of true positives among all actual positive instances, including both correctly identified objects (true positives) and those that were missed (false negatives). This indicates how well the model detects all relevant objects. The F1 Score, calculated as F1 Score = *2 * (Precision * Recall) / (Precision + Recall)*, combines precision and recall into a single metric, balancing the trade-off between accuracy and completeness. It provides a comprehensive measure of performance, which is especially useful for evaluating models on imbalanced datasets.

4.6. Results and Discussion

Various training and testing scenarios were conducted using the two Faster R-CNN models, R_50_DC_1x and R_50_DC_3x, to assess their performance. For the approach using Dataset 1, the methodology closely followed [213], employing the same dataset and hyperparameters. In contrast, Dataset 2 expanded the test set from 3 to 16 dolmens and increased the non-dolmen images from 2 to 8 in the test scenario while also adding non-dolmen images to the training and validation steps. This expansion was designed to test the model's ability to recognize all dolmens within the Pavia area. Various tests were conducted for each model and dataset combination. Across all training sessions, the R_50_DC_1x architecture consistently outperformed its counterpart. Notably, the replicated approach in Dataset 2 achieved better results, with an overall precision of 55.6%. Based on these outcomes, the R_50_DC_1x model was selected as the preferred pipeline for further analyses and prediction of new datasets. The results are presented in Table 4.2.

Algorithm	**Dataset**	**Overall Precision**
R_50_DC_1x	Dataset 1	44.6%
R_50_DC_1x	Dataset 2	55.6%
R_50_DC_3x	Dataset 1	43.7%
R_50_DC_3x	Dataset 2	51.5%

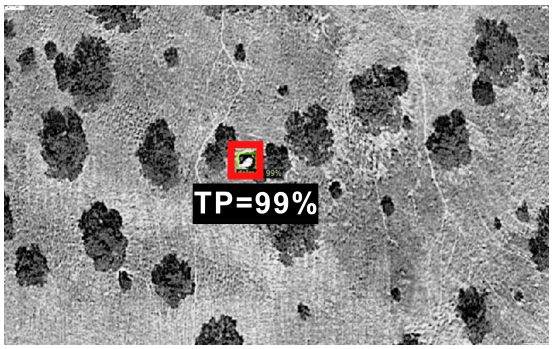
TABLE 4.2. Performance metrics (Overall Precision) for R_50_DC_1x and R_50_DC_3x models across different datasets

The aggregated results are presented in the confusion matrix in Table 4.4. This matrix reveals a high number of false positives (77) compared to true positives (46), with nearly five false positives for each correct detection. Switching the dataset to include only large and easily detectable monuments improves performance; however, the focus was on detecting the dolmens of Pavia. While the model performs exceptionally well during training, its performance on test data is less favourable. This discrepancy highlights the challenges posed by the small size of the dataset. A limited dataset may not provide sufficient information for effective model training, leading to suboptimal performance and difficulty reducing errors such as false positives.

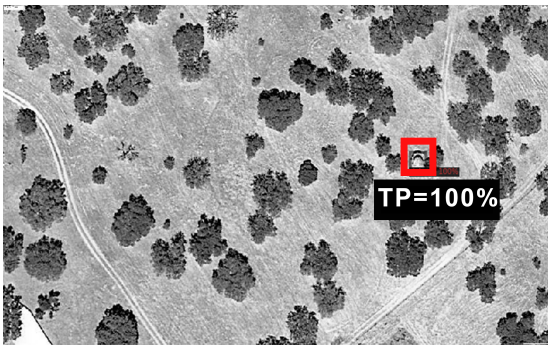
Classification	True	False
Positive	57%	43%
Negative	82%	17%

TABLE 4.4. Aggregated confusion matrix – dataset 2

The model demonstrated high performance in detecting monuments that are well-preserved and located in areas with minimal rocky outcrops. For instance, Figures 4.5a (Dolmen Adua 1) and 4.5b (Dolmen Antoes 3) illustrate successful detections of such monuments, achieving precision between 95% and 100% — this is similar for other monuments such as — Remendo 2, Goncala 2, Sao Miguel 4 and Anta Capela de S. Diniz.



(A) Dolmen Adua 1 – was recognized in all tests with precision superior to 95%.



(B) Dolmen Antoes 3– was recognized in all tests with precision superior to 95%.

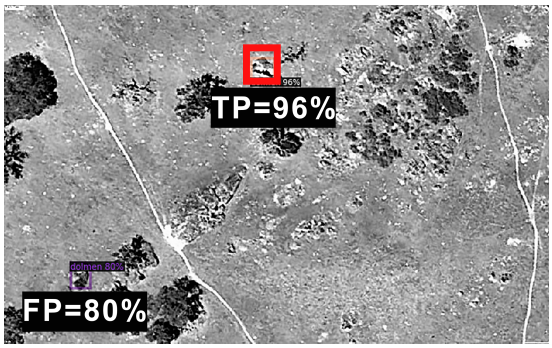
FIGURE 4.5. The figures A and B show examples of monuments that are consistently recognized.

Challenges arise when the environment includes rocky outcrops, as illustrated by Figures 4.6a (Dolmen Matalote 1) and 4.6b (Dolmen Casa Branca). Other examples occurred with Cre 2. In such cases, the algorithm tends to generate a high number of false positives. This issue is particularly pronounced in areas with substantial granite outcrops, which are characteristic of the Alentejo region.

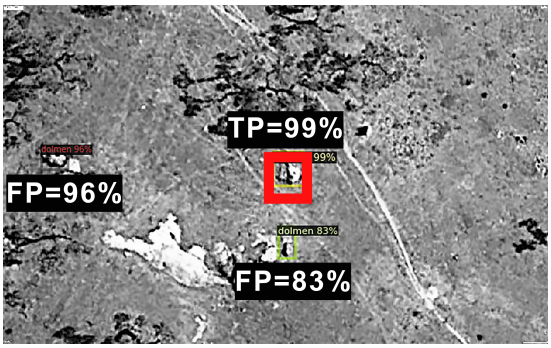
When images with no dolmens were analyzed, especially in these rocky areas, the model frequently returned false positives - exceeding 50% (Figure 4.7).

Conversely, the model has difficulty detecting monuments that are either covered by vegetation, such as dolmen Lapeira 1 (4.8b), Adua 4 or in very poor condition, like dolmen Tera 12 (4.8a), Adua 6, Alcarou de Baixo and Lapeira 2. In these situations, the algorithm's performance significantly declines, making efficient detection challenging. The presence of dense vegetation or advanced deterioration often results in missed detections or a high rate of false negatives.

When applying the model to predict new dolmens in a new data set (prediction set) with no dolmens, the number of false recognitions tends to increase, particularly in areas with significant granite presence. This additional test included 100 images, of which 64 images returned FPs. This resulted in an increase of false detections, reflecting the

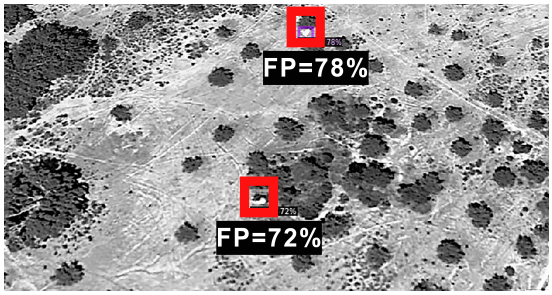


(A) Dolmen Matalote 1 recognized in 60% of the tests with precision above 85%.



(B) Dolmen Casa Branca 1 recognized in most tests with high precision.

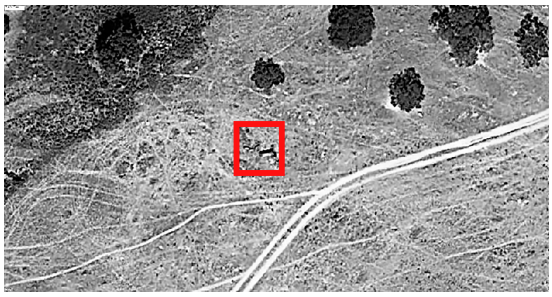
FIGURE 4.6. The figures show examples where monuments are present and identified by the model but also include false positives.



(A) Image with no target object showing false detections.

(B) Image with no target object showing false detections.

FIGURE 4.7. The figures show examples where no monuments were present but were incorrectly classified as such.



(A) Dolmen Tera 12.

(B) Dolmen Lapeira 1.

FIGURE 4.8. The figures show examples where monuments are present but were not identified by the model.

challenges posed by the rocky environments and the model's tendency to misidentify granite natural formations as dolmens. In the figure 4.9, a map is presented with the data used in the area of interest; the green points indicate the 'dolmens' in the test set, the red points represent the 'non-dolmens' set, which was divided into training and testing, and the black points correspond to the new dataset that we used after training the model.

The nature of the landscape plays a crucial role in the visibility and analysis of archaeological features. Flat terrains, with their smooth and unbroken surfaces, provide

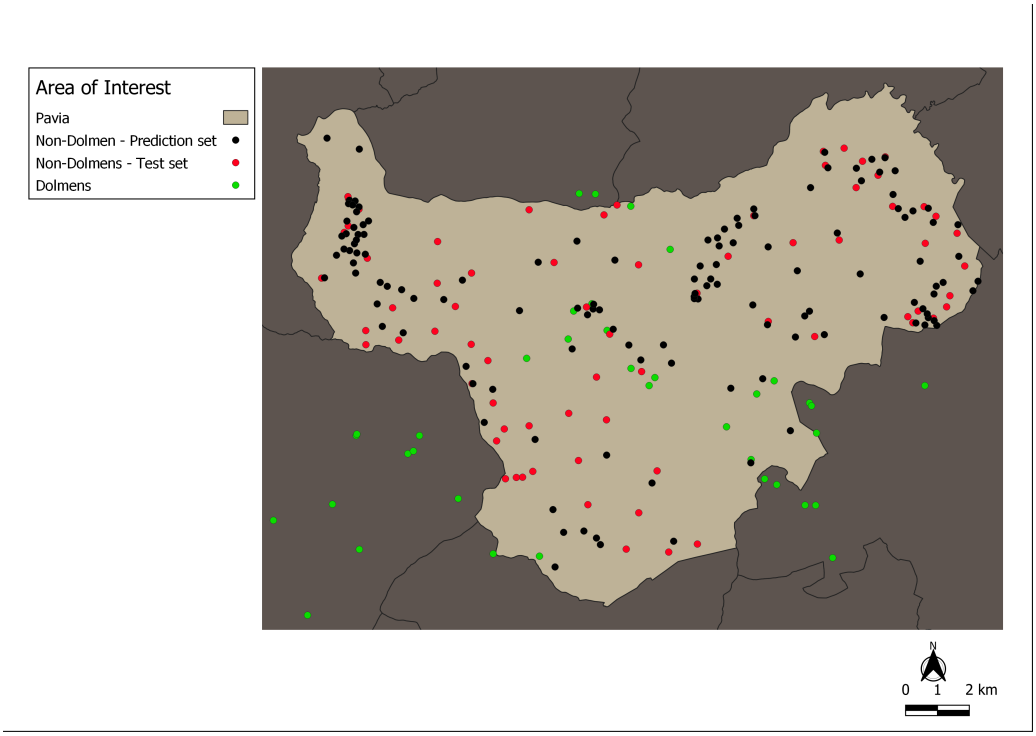


FIGURE 4.9. Map displaying 'dolmens' (green points), 'non-dolmens' (red points), and the new dataset (black points) used after model training.

an ideal backdrop for detecting these monuments. In these environments, the distinct outlines and features of the monuments stand out clearly against the background, making them more noticeable. In contrast, rugged and undulating terrains pose significant challenges due to variations in light, shadows, and potential camouflage effects. These factors can obscure or distort the visualization, making detection more difficult. Furthermore, an unobstructed view is essential for clear visualization. Elements such as dense vegetation, debris, or other environmental obstructions can greatly reduce the recognition. Thus, areas with a clear line of sight to the dolmen's structure are more favourable for successful detection.

Archaeological features are generally more visible on exposed ground than in cultivated fields or areas with irregular vegetation growth, which weather conditions can influence. Notably, some traces become visible due to anomalies in vegetation, such as changes caused by varying health states of plants. This phenomenon is especially evident during winter under dry and rainless conditions [35].

CHAPTER 5

Knowledge Graph: Monuments and Landscapes

5.1. Introduction

There are many platforms and formats for presenting information and data about cultural heritage and landscape. An image analyst can consult a variety of layers of Vector Topographic Maps (VTM)s to understand better the landscape and spatial relations between the geographic objects and features on the surface. Additionally, contextual information about the target object is used to derive knowledge. This process is further enriched by the specialist's own experience and expertise.

According to previous studies, dolmens tend to be found in elevated areas, near watercourses and rocky outcroppings and in regions with poor agricultural capacity [56, 69, 18]. In addition, these monuments tend to be spatially near similar monuments, and they are more visible from the ground in rural areas with a flat surface and little overlap. While domain knowledge of cultural heritage information can be derived from textual records, these landscape features can be derived from hydrography, hypsometry, soil type, and land use modern VTMs of the Areas of Interest (AOI).

Traditionally, these data are not interoperable because they consist of different types of information and formats. Ontologies and Knowledge Graph (KG)s for representing archaeological culture heritage and landscape features have been receiving more attention to deal with the lack of interoperability of data. There are ontologies and KGs with information about the landscape and cultural heritage, but they are rarely addressed together, and models are sometimes incompatible or non-open. Additionally, when it comes to dolmens located in our AOI, we do not find open graphs containing such data.

Due to the 1) lack of a model linking VTMs features and contextual information on the modern landscape with archaeological monuments, including spatial and contextual data, and 2) the absence of a semantic model populated with detailed, granular information on our target object and AOI (at least not openly available) — we chose to implement our own KG as an Labelled Property Graph (LPG). This was implemented to efficiently and coherently display extensive archaeological data related to monuments and landscapes, systematizing and integrating them. The goal is to represent knowledge granularly by capturing specific instances and their relationships. This approach enables detailed queries on particular aspects and information retrieval focused on instances. Using this knowledge structure, we aim to organize information interoperably, providing a solid foundation for future analysis and pattern recognition.

This chapter examines the implementation of the LPG, designed to represent information about the scene by linking instances of entities present in the AOI where dolmens

is located. Section 5.2 introduces the fundamental concepts of knowledge representation, providing the theoretical foundation for the chapter. The remainder of this chapter outlines the process for data gathering (Section 5.1), dataset preparation (Section 5.3), and the proposed schema model (Section 5.4). Subsequently, the implementation section describes the model implementation (Section 5.5), and the information retrieval section presents the analysis performed and conclusions (Section 5.6).

5.2. Knowledge-Based Semantic Networks

A knowledge base is a repository of information [217]. Systems based on Knowledge-Based (KB)s utilize domain-specific knowledge and rules to guide decision-making. A system is considered knowledge-based if it relies on a collection of symbolic structures that represent the information used for reasoning and operation [217, 218].

Semantic networks are a form of knowledge representation that uses graph structures to model relationships between concepts. These networks enable an intuitive and structured method for capturing and visualizing knowledge, with **vertexes** representing entities and **edges** defining relationships between them.

5.2.1. Knowledge Representation

Knowledge representation in the context of semantic networks refers to the process of encoding information about the world into a structured format that computational systems can understand and manipulate, enabling them to model, interpret, and reason about complex domains. On the other hand, knowledge reasoning involves using these representations to solve problems and derive new insights, mimicking human cognition's logical and analytical capabilities. Knowledge representation and reasoning refer to how knowledge is symbolically represented and automatically manipulated by reasoning programs [217].

Various standards have been proposed to communicate knowledge between knowledge-based systems and represent knowledge in a standard format. Ontologies and KGs are both methods for knowledge representation, with ontologies defining domain concepts and relationships in a formal structure, while knowledge graphs capture entities and their interrelationships in a graph-based format [217, 219]. Figure 5.1 illustrates how knowledge-base, knowledge-representation, ontologies and knowledge-graph concepts are related. The following subsections provide a detailed description of these knowledge representation models.

Ontologies

Ontologies (a.k.a. for knowledge-bases) are *'a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, which are useful to our purposes'*[220, p.2], defining *'explicit specifications of conceptualizations'.* [220, p.8] in a common used and powerful way to represent domain knowledge [221]. For Bianchi et al. (2020), it is "*a formal specification of the meaning of types and relationships expressed as a set of logical constraints and rules, which support*

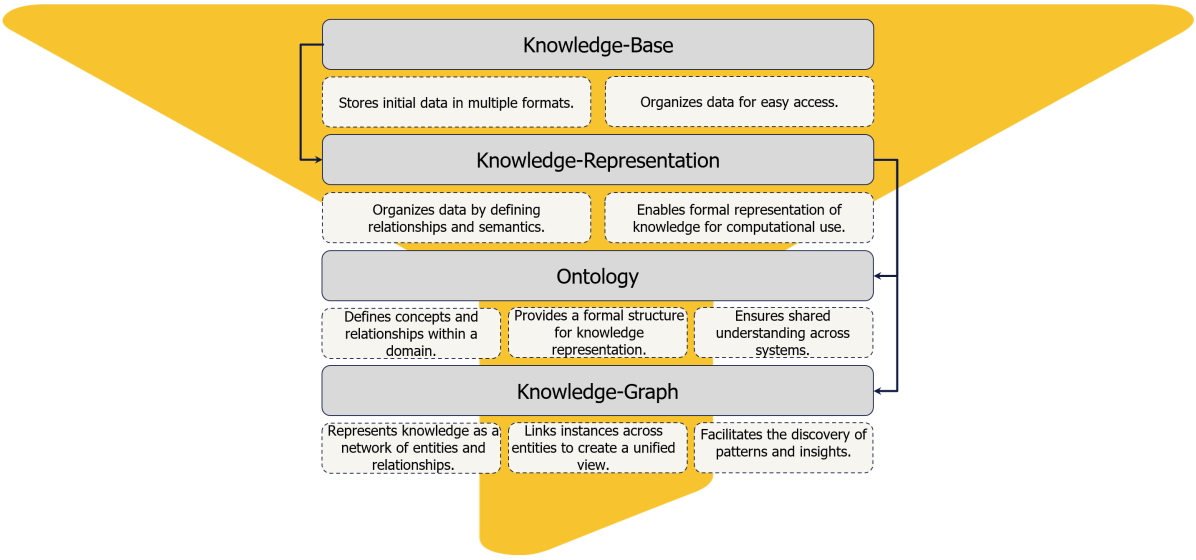


FIGURE 5.1. The diagram shows how knowledge bases, knowledge representations, ontologies, and knowledge graphs are interconnected.

automated reasoning."[222, p.2]. While there is a consensus on the importance of ontologies, opinions vary regarding their precise definitions and scope. In essence, ontologies are structured to define how to represent entities and their relationships. This structure semantically represents fundamental concepts and their relationships across various levels of abstraction, increasing the sharing and reuse of information while increasing the shared understanding of knowledge of a domain in an interoperable way [223].

An ontology is expressed in logic through logical axioms, which define the constraints of semantic modelling [217]. It provides a formalism designed to represent structured knowledge about a domain. The focus is on language and semantics to structure information in a way that ensures interoperability [217, 59]. Description Logics play a crucial role in the logical formalization of knowledge. It constitutes a family of logic rather than a single specific logic. They are based on three fundamental elements: individuals (e.g., Lisbon), classes (e.g., City), and properties (e.g., flight). Description Logics enable the formulation of assertions, known as axioms, about these elements. Assertive axioms can represent unary class relations in individuals [224]. Ontology's structure traditionally comprises two parts: the terminological part Terminological Box (T-Box), which includes definitions and axioms, and the assertive part Assertional Box (A-Box), which specifies individuals or instances [225, 184].

Different languages have been implemented to express these ontologies. The World Wide Web Consortium (W3C) has adopted several languages to represent ontologies, such as the Resource Description Framework (RDF) and Web Ontology Language (OWL). RDF provides a basic structure for describing data on the web composed of three elements representing two vertices connected by an edge: subject-predicate-object [219]. The subject represents a node or resource, the object a node or literal value, while the

predicate represents an edge. In this model, nodes or edges are identified by a Uniform Resource Identifier (URI), which is a unique identifier where the subject and predicate are URIs, and the Object is a URI or literal. So nodes and edges are purely unique labels, with no internal structure [226]. The OWL, defines how to model the RDF knowledge semantically [227, 228, 229]. These international languages represent the most widely used ontologies in information systems [226].

Some ontologies do not supply a vocabulary of concepts, so thesauri and glossaries provide controlled terminology required for a semantic link between different terms for the same concept [210]. A thesaurus is a controlled vocabulary with a structured semantic network of distinct concepts that improves information retrieval through categorized queries. It standardizes and harmonizes terminology across various sources, ensuring consistency and clarity in concept representation [230]. Different thesauri are available for defining specific domain features such as archaeological terms, such as ROSSIO [231], an open-access, free platform to aggregate, organize, and connect digital resources in the fields of Social Sciences, Arts, and Humanities, designed by Portuguese higher education and cultural institutions — utilizing structured vocabularies and widely recognized ontologies such as the Getty's Art and Architecture Thesaurus (AAT); GETTY AAT provides a comprehensive vocabulary for art and architecture, facilitating the standardization of terminology in cultural heritage documentation; and the Forum on Information Standards in Heritage (FISH) thesaurus has as main focus has been on developing content and data standards for use in the heritage sector [232].

Knowledge-Graphs

Information from heterogeneous sources can be represented, retrieved, and integrated using KGs. Even though the term *"knowledge graph"* has been used at least since 1972, its modern incarnation was introduced by Google in 2012 [233], followed by announcements of knowledge graphs from Airbnb, Amazon, Facebook, LinkedIn, Microsoft, and more [224, 234]. In recent years, more scientific literature has been published on knowledge graphs due to the growing industrial uptake of the concept. This model is designed to accumulate and convey real-world knowledge, focusing on representing relationships between entities [165, 224, 235].

A KG is structured around two primary components: **nodes** (or vertices/entities in ontological terms) and the **relationships** that connect them (known as edges or properties in ontologies). These nodes and relationships are not just abstract concepts; they can have specific instances (known as individuals in ontologies) with attributes or labels that provide additional context or details. Knowledge within a KG is typically expressed through factual triples, creating a web of interconnected entities and relationships that constitute the graph's structure [165].

The structure of a KG is often a source of confusion, particularly regarding its relationship with ontologies that comprise the T-Box and A-Box. However, the literature frequently mixes the distinction. Nys et al. (2018) suggest that, in the realm of computer

science and KGs, an ontology formally describes the types, properties, and relationships between real-world entities [155]. A similar idea is presented by Ferilli et al. (2021), who states that when the definitions/concepts/axioms T-Box are considered in conjunction with the instances/individuals A-Box, the result is a KG [235]. This interpretation underscores ontologies as the foundational framework, establishing the rules and structure that can be used to build KGs, while KGs focus on visualizing these relationships, emphasizing the representation of instances. From our perspective, ontologies emphasize logical structures, concept definitions, and semantic relationships. In contrast, KGs are primarily concerned with representing and linking real-world data—instances.

Additionally, KGs can integrate information into an ontology and use reasoning mechanisms to generate new insights [218, 159]. Ontologies, semantics, and reasoning are critical for extracting new information [236, 224]. Ontologies provide a formal representation of knowledge domains, while semantic technologies interpret the meaning and context of data, enabling advanced reasoning. Reasoning involves applying logical rules to existing knowledge, allowing for the derivation of new conclusions, inferences, and insights from both explicit and implicit data relationships. There are various KGs models for representation. For example, LPGs is a model where nodes and relationships have a uniquely identifiable ID and a set of key-value pairs, or properties, that characterize them, so in this model, nodes and relationships have internal structures [237] — with the 'key' being the property name and the 'value' containing the corresponding data (instance). Unlike the RDF structure where the nodes are atomic, LPG carries information allowing a compact structure, has unique identifiers for relationship instances, allowing different instances of the same relationship to be distinguished between the same pair of entities, and instances of relationships can have properties.

Graph Databases

In semantic networks, knowledge representation based on ontologies and knowledge graphs provides the foundation for effectively structuring and organizing information. Databases, then, offer a powerful tool for storing, managing, and querying interconnected information. SQL and NoSQL are two distinct types of database management systems. SQL is a relational database where data is stored in a highly structured format within tables, which consist of rows and columns with predefined data types. These databases require a strict schema design where tables are interconnected through referential integrity, typically using primary and foreign keys. When retrieving data from multiple linked tables, a JOIN operation combines rows based on matching keys [238]. Conversely, NoSQL databases offer more flexible solutions for handling unstructured or highly scalable data. Traditionally, NoSQL models are associated with Basically Available, Soft State, and Eventual Consistency (BASE) properties, whereas relational databases are known for their Atomicity, Consistency, Isolation, and Durability (ACID) principles [188]. NoSQL databases can be divided into five categories: key-value (e.g. Redis, Azure Table Storage, DynamoDB), column-based (e.g. Cassandra, HBase), document-based (e.g. MongoDB,

Couchbase), graph-based (e.g. ArangoDB, GraphDB, Neo4j) and multi-model (e.g. OrientDB). Choosing the proper database depends on the final objective and the type of data [239, 238, 240, 241, 242].

Graph databases are an excellent option for handling data in graph form. They are specifically designed to analyze relationships among data points more efficiently, which makes them ideal for knowledge graph applications. A Graph Databases (GDB) is a database that uses a graph structure, not data schemas or rows and columns, to represent the information. This model is specifically designed to manage large volumes of graph data and supports running efficient queries involving multiple levels of relationships between instances [240, 238]. Employing graph-based knowledge abstraction offers several advantages over relational models or NoSQL alternatives. Graphs provide a concise and intuitive abstraction for various domains, where edges and paths capture different and potentially complex relationships between entities within a domain [224, 238].

Data connections can be explored and graphically represented using different languages, such as RDF and OWL, which are ontology languages that use SPARQL for querying. Both languages represent data as a graph and are focused on exchanging data [219, 226] being ideal for incorporating disparate datasets and creating data ontologies. In most cases, these triple stores are indicated to be used with slow-changing, if not immutable, additive datasets. Scalability, storage optimization, efficient handling, mining, and browsing data are some of the advantages of this Database Management Systems (DBMS)s [235]. However, operational and transactional use cases were not intended for them. Alternatively, graph query languages (GQL), such as Cypher, are designed to query and represent data instances in graph databases [237].

Native Graph Databases

The terms native and non-native databases can be used to describe GDBs. Non-native graph databases use graphs as a bolt-on afterthought technology. Instead of being engineered explicitly for graph data, non-native graph storage uses relational, columnar, or other general-purpose databases. Performance and scalability are affected by graph data stored in non-graph storage. As in relational databases, relationships between rows are not physically stored, and foreign keys can be used to refer to a row from another row instead. That is a foreign key acts as a pointer. Using JOIN-like operators, relationships between rows can be calculated at query time. The cost of this type of operation increases exponentially with the size of the table, and the fact that the data is stored in a different format starts to build a gap between the conceptual model and the model that is stored and queried [237, 226].

Regarding of native graph storage (index-free adjacency), its purpose-built stack is managed for performance and scalability [226]. Using a Native Graph Database (NGDB), the focus is on efficient storage, querying and fast traversals across the connected data — since it is designed to maximize the speed of traversals during arbitrary graph search algorithms [237, 188]. The graph itself provides a natural adjacency index technique, so

NGDB are not dependent on indexes for searching linked data. Graphs are traversed by "walking" along it. Linked relationships point to a node at both the start and the end [226]. Each vertex or edge in the graph stores its own "mini-index" instead of a global adjacency index, allowing vertices to be found quickly. Without indices, determining if an element has a particular property would require a linear scan of all elements. Due to this model, the size of the graph has little or no effect on performance, and we can walk over the graph following these relationships without requiring JOINs [243], traversing a very large number of nodes per second [244]. Figure 5.2 shows the difference between the graph and relational models for the order management dataset.

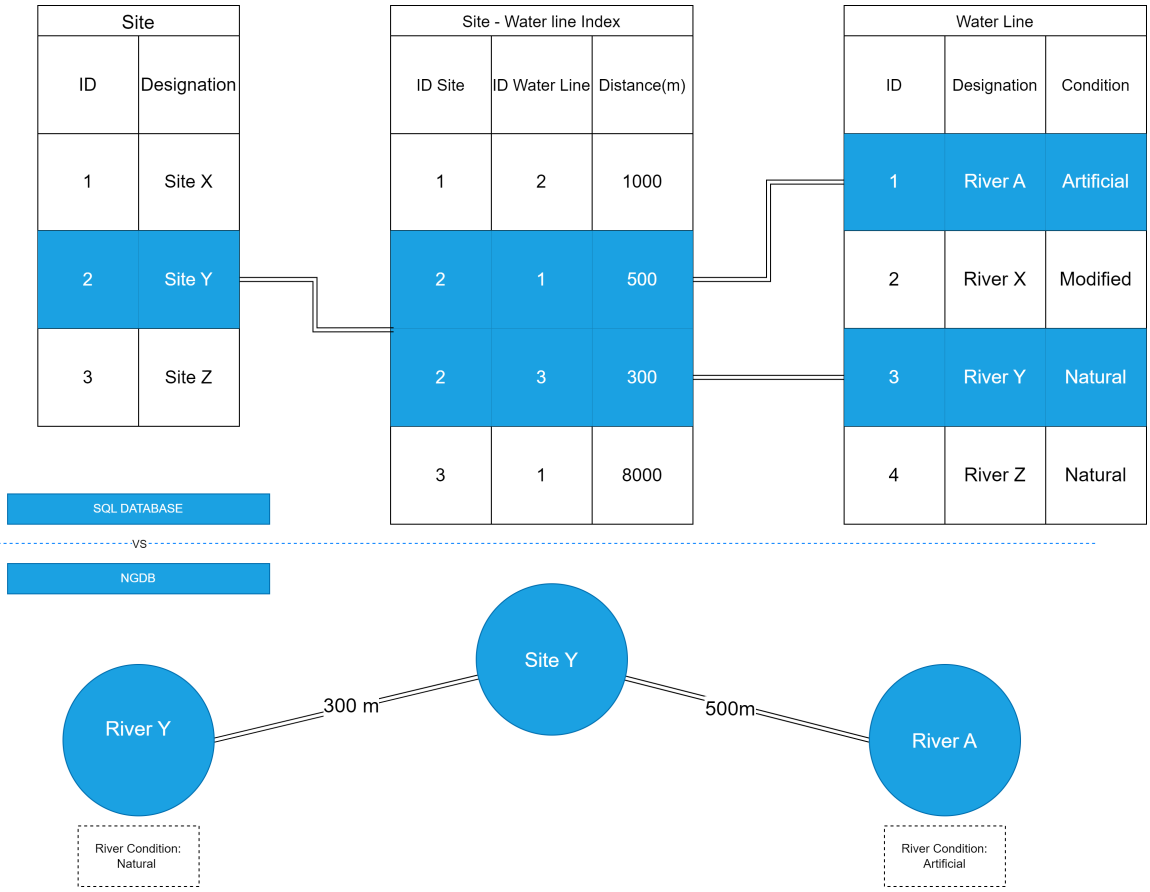


FIGURE 5.2. Comparison of search methods in SQL *vs.* NGDB.

The structure of a NGDB is explicitly built for storing graph-like data, ensuring that data is stored efficiently by writing nodes and relationships. At the same time, its processing is performed using index-free adjacency [237, 226]. It is the fastest way computers have to look at relationships since graphs have direct physical RAM addresses from each node. Storing and processing are the main differences between native and non-native graph databases [237]. Although improving traversal performance, native graph processing makes some non-traversal queries difficult or memory-intensive [241].

5.3. Data Acquisition and Preparation

This section details the data and the data sources used in this study to establish the domain knowledge required for archaeological site analysis. These sources include semantic records, which comprehensively describe archaeological sites and their contexts, and digital vector maps, which offer spatial representations of the sites and their surrounding landscape. Integrating these datasets ensures a multidimensional understanding of the archaeological and environmental context, bridging semantic knowledge with geospatial analysis.

5.3.1. Semantic Records

In this dissertation, we used two primary sources to gather information about the dolmens: the *Portal do Arqueólogo (PA)* [8] and the *Carta Arqueológica (CA)* [73]. The CA was published in 2012 as a physical book. It emerged following archaeological work conducted since the mid-1990s, supported by the local authority and research projects by specialists in the field. Mora Municipality and the Ministry of Culture supported these projects. In contrast, the PA is a digital platform intended for professionals and researchers in archaeology in Portugal — allowing researchers to access semi-structured information on Portugal's cultural heritage [8]. Recently, in 2021, it has also provided access to a geoportal with spatial information on the locations of these monuments. The platform is managed by the Direção-Geral do Património Cultural (DGPC). The description of the monuments in both data sources is presented in a semi-structured or unstructured format. Figure 5.3 illustrates the record for the dolmen Anta de Pavia PA.

FIGURE 5.3. Record of the dolmen Anta de Pavia in the PA database (Obtained on 11/09/2024) [8].

Visualization Keys	Dolmen aspects that impact how they are visualized			
	Complete chamber with capstone	Complete chamber without capstone	Incomplete chamber with capstone	With tumuli covering the chamber
Artificial earth mound (tumuli): - Material: Earth and rocks; - Height: $\geq$1m - Diameter: $\geq$5m				X
Chamber: - Material: Granite and schist (Color grey); - Diameter: 1.5m-5m; - Shape: polygonal, circular, semi-circular or quadrangular.	X	X	X	
Chambers orthostats: - tends to be inward-angled; - Some must be in situ or are identified through ground marks.		X	X	

TABLE 5.2. Key Dolmen aspects influencing satellite visualization and recognition, highlighting features that distinguish different configurations.

A more complete understanding of monuments was enabled by using multiple sources, since each may contain unique details. For both of the data sources, we extracted, standardised, and transformed the relevant information to be represented in a structured way. These sources address different versions of available data about dolmens. It is important to recognize and take into account that information about monuments can change over time. As monuments undergo alterations or are studied by different experts across various periods, their analyses and records may vary. Interpretations can differ as people approach the same monument with different perspectives. This variability means that the information found in the data sources analyzed may differ due to these evolving records and interpretations. The information provided by these sources provides context for explaining the characteristics of this type of monument. Figure 2.1 shows an example of a monument, highlighting the structural elements. In contrast, Table 5.2 outlines the key visual features of the monument, serving as classification criteria to guide the understanding of how its site appears in an aerial view, depending on its condition (e.g., complete chamber with capstone, or without capstone, incomplete chamber, or buried).

Monuments may be visible, but understanding them requires more than just observing them in an image. The context helps visualize what is visible and understand what is not.

This includes details on the monument's designation (some monuments have multiple designations), location (place name and its geographical coordinates), description (covering general concepts such as monument type, and specific details like (i) its component parts: chamber, presence of a corridor, or a burial mound; and (ii) the characteristics of these parts: shape, size, number of orthostats, construction material, and condition), as well as a description of the data source (metadata).

The data sources describe various aspects of the monuments, including their identification (Class, Designation(s), Period of construction), physical characteristics (shape, size, material, parts, and state of preservation), access (Localization), collected remains (Remains, Deposit) and more.

Defining Key Features

A dolmen structure can be represented as a whole or a collection of individual components. Different elements can represent each component based on its characteristics (for example, materials, dimensions, and conditions). A terminology was defined to describe the object's structure components. The terms were first analysed using a thesaurus. For example, dolmens are described as *"A megalithic tomb of Neolithic date comprising a **chamber** bounded by large upright **orthostats**, often only three or four in number, which supports a large **capstone**"* [232]. The terms used to describe the dolmens, such as **chamber**, **orthostats**, and **capstone**, were defined based on the thesaurus for the terms it included.

The thesaurus served as a foundation for defining terminology, but proved insufficient to capture information at a granular level. To address this limitation, we analyzed specialized articles and references commonly used by experts to determine the terminology used to describe the component parts of the monument (chamber, corridor, and burial mound). Furthermore, our analysis of the data sources informed the identification of specific physical characteristics to extract and represent for each component. Table 5.3 provides a representation of dolmen's structural elements and the terminology associated with the definition of the information.

Dolmen	Object Structure	Structure Information
Whole	Dolmen	Condition State; Material; Dimension;
Components	Chamber	Shape; Condition State; Dimension; Orthostat - (number and position); capstone (condition state);
	Corridor	Condition State; Dimension; Orthostat (Number and side)
	Burial Mound	Material; Condition State; Dimension;

TABLE 5.3. Terminology for describing dolmen structural elements and concepts: Representation of the object and its components

Condition State	Material	Measurement Type	Chamber Shape	Presence	Location Definition
Good	Granite	Diameter	Polygonal	Yes	in situ
Bad	Schizt	Lenght	Circular	No	nositu
Destructed	Quartz	Widht	Semi circular	Maybe	north
Vestigy		E-W direction	Quadrangular		south
Regular		N-S direction	Oblong		
		min height	Trapeizodal		
		max height	Rectangular		

TABLE 5.4. Terminology and attributes for describing dolmen structural elements.

Managing Data Sources

In what concerns site descriptions, both sources provided complementary yet varying perspectives. PA records 73 described monuments in Pavia, 51 of which are not mentioned in CA. On the other hand, CA documents 49 monuments, of which 21 are only described by this source. The analysis of both documents revealed a total of 94 known dolmens located within Pavia. It should be noted that these sources contain some inconsistent information, with terminology differing even within the same source.

The main differences between the two sources are information about assets' geographic locations and the terms they use to describe designations, conditions, and measurement types. For example, the coordinate formats were different: CA data used ESRI:102164 — Lisboa Hayford Gauss IgeoE system, while PA data were obtained through its geoportal using WGS 84 format. For consistency, all spatial references have been standardized. Although the spatial coordinates were standardized, we preserved all the coordinate information, with a focus on extracting the original coordinate information and having the data in either WGS 84 or the Cartesian coordinates system.

To ensure a detailed and systematic extraction of information from each document, we tried to maintain the terminology employed by the source for describing the object's attributes. Table 5.4 presents the defined terminology for describing the characteristics of each part of the monument. The information was originally in Portuguese, but we converted it into English using equivalent terms, such as "bom" to "good". All attributes that relate to the dolmen condition state (e.g., good, bad, destroyed, vestigial, and regular), material (granite or schist) and shape (polygonal, circular, semicircular, quadrangular, oblong, trapezoidal, and rectangular) were only extracted when clearly stated in the data sources.

About measurements, despite all dimensions being reported in meters, terminology varied among descriptions — some referred to "diameter," others to "minimum height," "maximum height," or directional measurements such as E-W and N-S. The measurements were maintained as stated in the original source, ensuring consistency. Measurements are

used to describe the components of the monument (chamber, corridor and burial mound), including specific features such as the capstone and orthostats.

In some cases, the descriptions were straightforward, such as stating that the monument was in a good or poor condition. But others required interpretation, as in statements like "the capstone is displaced or broken". For such cases, specific terms were defined to represent this information in a structured manner. The presence of features was recorded as a "yes", "no", or "maybe" for object structure components such as corridors, mounds, and capstone. Position of characteristic elements, like the "capstone", were described in terms of location (in situ, not in situ, North, or South), and the presence of orthostats for elements like "corridor" and "chamber" was also documented using numerical values to describe the number of orthostats that cast the monument.

As already mentioned, differences were observed between both sources' descriptions during the analysis and data extraction process. Due to discrepancies between data sources, where the IDs, and sometimes even the names of the monuments differ, we assigned Global IDs to each monument to ensure consistency and resolve conflicts. A monument's Global ID starts with a capital letter "D" followed by a numerical value (e.g., D11) and are created according to the monument's name and/or location. So, a monument with the same designation or location has the same ID, allowing us to track the monument information efficiently, even if it is documented under different designations. This attribution required manual analysis and understanding of the descriptions to assign the Global IDs. The manual work was necessary because the texts used different formats and terminology, with inconsistent references and descriptions that required careful interpretation.

Various monuments' levels of detail also differed between the two datasets. As an example, dolmen Lapeira 1 (D11) is recorded in CA and PA, but the descriptions differ: PA includes the condition of the monument, while CA does not, and the size measurements vary. In other cases, such as Alcarou de Baixo 3 (D10), only location information is provided, with no description in either source. There were also instances where no information was provided about a monument or specific details were missing in one or both sources, in which case *"Not Available (n/a)"* was added. The examples mentioned previously are shown in Table 5.6.

Global ID	**Source**	**Designation**	**Shape**	**Diameter**	**Condition**
D11	PA	Lapeira 1	Polygonal	3.4	Good
D11	CA	Lapeira 1	Polygonal	3.2	n/a
D10	PA/CA	Alcarou de Baixo 3	n/a	n/a	n/a

TABLE 5.6. Examples of data information provided by PA and CA datasets.

5.3.2. Vector Topographic Maps

In this dissertation we used various sources to obtain detailed information about the landscape of the AOI where the monuments are located. The aim was to spatially relate the monuments to elements in the landscape that may have influenced the choice of their construction in the past, such as natural watercourses, soil types, topography relief, and land use, as well as contribute to their recognition. By exploring these spatial relationships, we aim to understand how the proximity and interaction between the monuments and their surrounding landscape elements may have influenced the choice of their location in the past while also affecting their current visibility and preservation.

Each piece of consolidated information is generated or owned by different institutions, as shown in Table 5.7. Thus, vector maps were sourced from various institutions, namely the Direção-Geral do Território (DGT), the Direção-Geral de Agricultura e Desenvolvimento Rural (DGADR), the Sistema Nacional de Informação de Ambiente (SNIAmb) and the Municipality of Mora (MM). The following provides comprehensive details on each dataset. All data except the contour line map are in shapefile format obtained from the Sistema Nacional de Informação Geográfica (SNIG) portal [121] and licensed under CC-BY-4.0. The contour line map has been shared and allowed for use by the municipality of Mora.

Data	**Description**	**Owner**	**Format**
Borders	Official administrative map of Portugal	DGT	Polygon + In-line
Soil Use	The land use and occupation map of Portugal	DGT	Polygon
Soil Type	The soil type map of southern Portugal	DGADR	Polygon
Contour Line	Mora's contour lines map	MM	Line
Water Line	Surface water masses rivers of mainland Portugal	SNIAmb	Line

TABLE 5.7. Source and details of topographic vector maps used in the analysis.

Data on Borders

The Official Administrative Map of Portugal [2], produced by DGT, was officially endorsed by their order dated January 24, 2023, and was published as an official notice on February 3, 2023, according to Regulatory Decree no. 30/2012 of March 13. This dataset portrays Portugal's administrative boundaries, including districts, municipalities, and parishes. The dataset consists of eight shapefiles: seven in polygon format and one in line format (Catalog of Official Administrative Map of Portugal (CAOP) Entities n.d.). The locations for parish, municipality, and district were obtained from a cartography file (Cont_AAD_CAOP2022.shp). Table 5.8 lists all data in this shapefile, including columns for attribute names, their descriptions, and attribute value format types.

Attributes	Attributes Description	Format
DICOFRE	Unique identifier for parishes (e.g., Brotas, Pavia, and Mora)	Integer
Freguesia	The official designation by which the parish is known	Text
Concelho	Identification of the municipality to which the administrative area belongs	Text
Distrito	The designation by which the district is known	Text
TAA	Identification of the type of administrative area	Text
Area_T_ha	Total value of the parish area	Float
Area_EA_ha	Value of the Administrative Area	Float
Des_Simpli	Simplified designation of the parish name	Text

TABLE 5.8. Attributes, descriptions, and data types represented in the CAOP 2022 VTM.

Soil Use Data

The Land Use and Occupation Map, or *"Carta de Uso e Ocupação do Solo"* (COS), is a product of the Sistema de Monitorização da Ocupação de Solos (SMOS) initiative developed by the DGT. The Carta de Uso e Ocupação do Solo (COS) map, characterized by polygons representing homogeneous land use and occupation units, was published in 1995 [245], 2007 [246], 2010 [247], 2015 [248] and 2018 [87], and is currently being updated. We obtained the V2 versions for COS 1995, 2015, and 2018, and the V3 version for COS 2007. These updated versions, which include COS1995v2, COS2007v3, COS2010v2, COS2015v2, and COS2018v2, replace their earlier counterparts and are part of COS 2018 enhancements. The revised historical series marks the integration of COS into SMOS, which now also includes the Conjunctural Land Use Map (COSc). While COS provides structural information on land use, COSc focuses on land occupation. COS remains the primary national reference for land use mapping, which is why it was chosen for this study [249].

COS geospatial information divides the landscape into units representing homogeneous land use and occupation categories, excluding linear and point elements. Except for the 1995 COS, which featured 44 classes, all subsequent maps include up to 83 classes at their most detailed level – 4 levels. The nomenclature in each COS follows a hierarchical system of land use and occupation classes. Each polygon is assigned a code that corresponds to its classification within this hierarchy. Along the border with Spain, COS boundaries align with the CAOP of the corresponding year, while maritime boundaries are defined through photo interpretation. All COS maps use the reference system EPSG:3763:EPSG:4258. Table 5.9 lists all data in this shapefile, including columns for attribute names, their descriptions, and attribute value types (e.g., real numbers, doubles, integers).

Soil Type Data

The Land Type Cartography of Portugal, at a scale of 1:25,000, offers detailed information about the various soil types, represented on the map by distinct cartographic symbols (e.g.,

Attributes	Attributes Description	Level	Format
COS18n1_C	Class code of land use/occupation	1	Float
COS18n1_L	Label designation of the class code	1	Text
COS18n2_C	Class code of land use/occupation	2	Float
COS18n2_L	Label designation of the class code	2	Text
COS18n3_C	Class code of land use/occupation	3	Float
COS18n3_L	Label designation of the class code	3	Text
COS18n4_C	Class code of land use/occupation	4	Float
COS18n4_L	Label designation of the class code	4	Text
Area_ha	Value of the polygon's area		Float

TABLE 5.9. Attributes, descriptions, and data types represented in the vector map - COS.

Solos Litólicos, Não Húmicos Pouco Insaturados, Normais, de granitos (pg)). These series are subdivisions of families, defined as groups of soils that share similar horizons or layers, distributed uniformly along the surface and formed from the same organic material [85].

The explanation for the symbols used in the cartography is accessible online through the DGADR portal [1] and is not included within the map [1]. This note defines the codes and labels, explaining what each represents. The maps are divided into northern and southern parts of the country, and for this study, we analyzed only the southern part. Note that cartography uses reference systems EPSG:3763, EPSG:4258, and ESPG:4326. Table 5.10 presents the attributes, attribute descriptions, and attribute value types for this cartography.

Attributes	Attributes Description	Format
COD1_Solos	Soil code and phase	ENUM
COD2_Solos	Soil code and phase	ENUM
COD3_Solos	Soil code and phase	ENUM
WRB2014_1	Soil code.	ENUM
WRB2014_2	Soil code.	ENUM
WRB2014_3	Soil code.	ENUM
ADD_1	Soil phase.	ENUM
ADD_2	Soil phase.	ENUM
ADD_3	Soil phase.	ENUM

TABLE 5.10. Attributes, descriptions, and data types represented in the land type VTM

Contour Line Data

The municipality of Mora made available the approved cartography, at a scale of 1:10,000, with territorial data - approved up to 09/09/2022. Decree-Law n.º 130/2019 in article 2 defines the approved cartography as vector/topographic and hydrographic imaging and their respective thematic maps, which have been recognized by competent services as having met the defined criteria, norms and specifications. The DGT has specifications

[1]The explanation note can be accessed at: https://www.dgadr.gov.pt/nota-explicativa

for topographic cartography norms and technical specifications. The norms and technical specifications for large-scale vector and image topographic cartography were published through Notice n.º 11918-2019 in the Official Gazette on July 24th.

The data contains a map with contour lines that represent height measurement. The contour lines are imaginary lines intersecting the terrain where all points have the same elevation value relative to a specific altimetric datum. Table 5.11 presents the attributes, attribute descriptions, and attribute value types for this cartography.

Attributes	**Attributes Description**	**Format**
ALTITUDE	Numeric value indicating the terrain's altitude	Float
TIPO	Code preceded by a dash indicating contour line type (Master/Secondary/Auxiliary)	ENUM

TABLE 5.11. Attributes, descriptions, and data types represented in the contour line VTM.

Water Line Data

Although there are numerous sources of vector hydrographic data pertaining to the Portuguese territory, we chose to work with the *"Massas de água superficiais Rios de Portugal continental"* of SNIAmb, a geographic dataset at scale 1:25,000. The cartography refers to the water bodies of rivers reported to the European Commission under the *Diretiva Quadro da Água*, for the 2nd planning cycle 2016–2021. Regional Hydrographic Management Plans are instruments aimed at the management, protection, and environmental, social, and economic valorisation of waters at the regional hydrographic level.

The plans are drafted in planning cycles and revised and updated every six years. They encompass various subjects grouped into five major thematic areas: Water Bodies (surface and underground); Environmental Objectives; Measures Program; Economic Analysis; and Public Participation. These are developed in accordance with the Water Law and Dispatch n.º 11955/2018, 2nd series, dated December 12. All Regional Hydrographic Management Plans are reported to the European Commission, which, in partnership with the European Environment Agency, developed the platform WISE (Water Information System for Europe).

Table 5.12 presents the attributes, attribute descriptions, and attribute value types for this cartography.

Managing Data Sources

In order to relate entities to other semantic information, we converted these maps, composed of multiple layers, into a relational format.

Attributes	Attributes Description	Format
codigo	Code representing the water bodies	ENUM
Nome	Name of the water body	Text
Regiao_hid	Hydrographic Region	Text
Natur_fm_a	Classification of water bodies: Natural, Artificial, and Heavily Modified	Text
Transfront	Transboundary water bodies – Portugal / Spain (Yes / No)	Text
Est_pot_ec	Likely refers to the ecological potential status of the water bodies, especially for artificial or heavily modified ones (based in RH5 Final Environmental Report)	Text
Estado_qui	Chemical status of the water bodies (Good, Unknown, Insufficient)	Text
St_lenght	Length of the water bodies represented as lines	Float

TABLE 5.12. Attributes, descriptions, and data types represented in the water line VTM.

Data was extracted from each VTM using ArcGIS Pro Desktop[2]. For data cleaning and standardisation, all datasets were encoded consistently with ISO-8859-1 Unicode [251]. The standardization applied to spatial coordinates extracted from dolmen semantic records was also used for all VTMs. Afterwards, the geometric information was extracted and the coordinates were converted to Well-Known Text (WKT) format [252].

The extracted data was organized into a structured format to facilitate bulk loading into a database. The original descriptions from the source files were preserved to maintain data integrity and consistency. All details and terms used in the original documents are preserved in the database, allowing a seamless transition between raw data and the database.

Grid Implementation

The segmentation of the AOI was achieved through a structured grid that was created to standardize spatial data representation and facilitate its integration into the knowledge graph. The total area was divided into 150 individual cells with an area around 1 km^2. The resultant grid data was extracted and saved using the same descriptions as the VTMs.

Traditionally, linked data approaches represent spatial information by directly mapping geometries onto the surface of the Earth using technologies like GeoSPARQL. Although accurate, this method can lead to complex queries and long execution times, especially with overlapping geometries. Structured grid systems offer a significant advantage in addressing these challenges. In spatial data analysis, a regular grid of cells is used to divide the study area into discrete, manageable blocks. The grid-based approach

[2]ArcGIS is a geographic information system developed by Esri and designed for spatial data management, analysis, and visualization. It offers robust tools for capturing, analysing, and presenting geographic information, enabling users to work with a wide array of spatial data formats. Its user-friendly interface and extensive functionality facilitate the extraction and conversion of vector map data into various formats [250].

enables the examination of individual cells, the prediction of adjacent cell contents, and the rapid overview of spatially co-located features and regions.

Leveraging grid coordinates streamlines the search for spatial points, enhances data analysis and retrieval, reduces computational complexity, and improves query performance through pre-computed spatial relations. As Shimizu et al. (2021) [253] suggest, the use of a discrete global grid for integrating spatial data within a knowledge graph is a promising method, potentially enabling future knowledge graphs to perform integration more efficiently and rapidly [253].

5.4. Schema Definition

The schema model is based on CIDOC-Conceptual Reference Model (CRM), its extension CRMgeo, and GeoSPARQL. The CIDOC-CRM [59] ontology is an ISO standard for semantic interoperability among cultural institutions (ISO 21127:2006). This event-centric ontology provides a guide for modelling heterogeneous information. By providing a model for good conceptual modelling, the model guides the structuring of information and of the existing interrelationships without prescribing specific terminology or what should be documented [59].

The CRM (version 7.2.1) contains 81 hierarchically organized classes and 160 unique relationships (referred to as "properties" in the documentation) [59]. An identifier (number) and name are assigned to each declared class and relationship. Class names are prefixed with the letter 'E', while relationship identifiers begin with the letter 'P'. In both cases, the letter is followed by its respective identifier number. Properties are bidirectional: in the domain-to-range direction, they are written without parentheses; in the reverse range-to-domain direction, they are enclosed in parentheses [59].

The model can be implemented in various databases using any language. It is particularly well-suited to storing information as triples that emphasize entities rather than instances. There are, however, some limitations. This monolithic structure does not provide the modularity necessary to evaluate and represent the way cultural heritage objects are interpreted by different agents (e.g., researchers) and what new information is created as a result [96]. Further, CIDOC-CRM can handle some spatial relationships and temporal information but not detailed spatial data.

CRMgeo [57] is an extension of CIDOC-CRM designed to enhance the representation of geospatial information within cultural heritage. Its primary function is to integrate geoinformation available in Geographic Information Systems (GIS) format into CIDOC-CRM. Its primary purpose offers a schema that aligns with CIDOC-CRM, facilitating the integration of geoinformation through conceptualization, formal definitions, encoding standards, and topological relationships established by the Geospatial Consortium (OGC) in GeoSPARQL. A more comprehensive and contextually meaningful representation of spatial data can be achieved by linking cultural heritage data with precise site location and geometry information. The model integrates temporal entities with persistent items through a temporal-spatial scheme [94]. In 2015, CRMgeo 1.2 was released with

CLASSES	SOURCES
E1 CRM Entity	CIDOC-CRM
E3 Condition State	CIDOC-CRM
E12 Production	CIDOC-CRM
E13 Attribute Assignment	CIDOC-CRM
E17 Type Assignment	CIDOC-CRM
E18 Physical Thing	CIDOC-CRM
E19 Physical Object	CIDOC-CRM
E22 Human-Made Object	CIDOC-CRM
E28 Conceptual Object	CIDOC-CRM
E31 Document	CIDOC-CRM
E41 Appellation	CIDOC-CRM
E42 Identifier	CIDOC-CRM
E52 Time-Span	CIDOC-CRM
E53 Place	CIDOC-CRM
E54 Dimension	CIDOC-CRM
E55 Type	CIDOC-CRM
E57 Material	CIDOC-CRM
E58 Measurement Unit	CIDOC-CRM
E60 Number	CIDOC-CRM
E65 Creation	CIDOC-CRM
E74 Group	CIDOC-CRM
E93 Presence	CIDOC-CRM
E94 Space Primitive	CRMgeo
SP4 Spatial Coordinate Reference System	CRMgeo
SP6 Declarative Place	CRMgeo
SP15 Geometry	CRMgeo
geo:Feature	GeoSPARQL
geo:Geometry	GeoSPARQL
C1_Geometry_extracted from_maps	–

TABLE 5.13. Classes from CIDOC-CRM, CRMgeo, and GeoSPARQL used in the schema model.

13 classes and 19 properties. Based on the CRM model, each class and property is assigned a name and an identifier. For classes, the identifier consists of the prefix "SP" followed by a number, while for properties, it begins with the letter "Q." By combining CRM with specialized ontologies like CRMgeo, geospatial capabilities can be significantly enhanced [94, 57].

GeoSPARQL [252] is an OGC standard that facilitates geospatial data representation and querying. It provides a vocabulary for embedding geospatial data in RDF and extends SPARQL for geospatial querying. The ontology includes the class geo:SpatialObject, which denotes any entity with a spatial dimension. This class consists of two primary subclasses, namely the geo:Feature, which represents real-world entities such as rivers and houses, and geo:Geometry, which defines the spatial location of these features and can be

represented using literals and type hierarchies like WKT or Geography Markup Language (GML). Features are linked to their geometries via the geo:hasGeometry property [252].

The LPG was implemented using classes from these three ontologies as labels and properties as relationships. To represent information extracted from VTMs and semantic data related to archaeological monuments, 23 classes from CIDOC-CRM ontology were used. Due to the lack of specific spatial classes in CRM, we incorporated three classes from the CRMgeo extension and two from GeoSPARQL. We created a new class *C1 Geometry extracted from maps* to store the data extracted from vector cartography — see Table 5.13 for the classes used.

Relationships establish connections between classes. Table 5.15 provides an overview of the relationships defined for our model, detailing how each class connects, from the domain class to the target/range class.

The schema model was defined to encompass all relevant information about monuments and their surroundings, organized into three main categories: (i) Metadata definitions, (ii) General and individual concepts, and (iii) Geometry. Metadata is defined as "data about dat" [254] and includes the source of information, ensuring that the origin of specific details can be traced, which is essential for referencing and citation. Whithin metadata, there are two types of concepts: general and specific. General concepts provide a framework for understanding various instances within their scope, such as "rivers" for types of bodies of water or "dolmen" for types of monuments. On the other hand, specific concepts refer to instances of general categories. For example, the Matalote River is an individual river (a specific river), while "Lapeira 1" is an individual dolmen (a specific monument). Each concept includes detailed attributes such as the length, shape, and chemical condition of the Matalote River or, for the monument, the size shape, as well as the condition of the dolmen. Finally, the model includes the exact location of each feature, as represented in the data source, and the geometry of each place.

For all data types, we used consistent classes to represent the geometry. Furthermore, for metadata definitions, each data source is treated as an event marking the creation and subsequent acquisition of information about an object. The fact that we were dealing with two separate concepts and types of data-archaeological monuments, for which information was extracted from semantic sources and landscape information derived from vector maps, made it necessary to create separate subgraphs. One subgraph was dedicated to the monuments, while each VTM was transformed into its respective subgraph, interconnected by their locations. By clearly differing between the data types and linking them via geographical coordinates, this approach enabled us to maintain clarity and organization.

5.4.1. Representing monuments

A schema for representing the dolmens can be seen in Figure 5.4. In this schema, records refer to events that represent the creation and acquisition of information about objects. To document assertions regarding the object's values, we use the *E13 Attribute Assignment* class. Descriptions are linked to an *E31 Document* class that represents the

Relations	Class-Domain	Class-Range
P1 is identified by (identifies)	E1CRMEntity	E55MonumentType
P1 is identified by (identifies)	E22HumanMadeObject	E41Appellation
P1 is identified by (identifies)	E28ConceptualObject	E41Appellation
P1 is identified by (identifies)	E41Appellation	E42Identifier
P2 has type (is type of)	E22HumanMadeObject	E55Type
P2 has type (is type of)	E53Place	E55Type
P2 has type (is type of)	E22HumanMadeObject	E55Type
P2 has type (is type of)	E58MeasurementUnit	E55Type
P2 has type (is type of)	E60Number	E55Type
P2 has type (is type of)	E52TimeSpan	E55Type
P2 has type (is type of)	E17TypeAssigment	E55Type
P4 has time-span (is time-span of)	E12Production	E52TimeSpan
P4 has time-span (is time-span of)	E13AttributeAssigment	E52TimeSpan
P4 has time-span (is time-span of)	E28ConceptualObject	E52TimeSpan
P14 carried out by (performed)	E65Creation	E74Group
P39 measured (was measured by)	E16Measurement	E22HumanMadeObject
P40 observed dimension (was observed in)	E16Measurement	E54Dimension
P41 classified (was classified by)	E17TypeAssigment	E28ConceptualObject
P42 assigned (was assigned by)	E17TypeAssigment	E55Type
P44 has condition (is condition of)	E19PhisicalComponent	E3ConditionState
P45 consists of (is incorporated in)	E22HumanMadeObject	E57Material
P46 is composed of (forms part of)	E22HumanMadeObject	E22HumanMadeObject
P46 is composed of (forms part of)	E22HumanMadeObject	E19PhisicalComponent
P48 has preferred identifier (is preferred identifier of)	E22HumanMadeObject	E42Identifier
P48 has preferred identifier (is preferred identifier of)	E55Type	E42Identifier
P48 has preferred identifier (is preferred identifier of)	SP6DeclarativePlace	E42Identifier
P55 has current location (currently holds)	E22HumanMadeObject	SP6DeclarativePlace
P57 has number of parts	E19PhisicalComponent	E60Number
P67 refers to (is referred to by)	SP6DeclarativePlace	E42Identifier
P70 documents (is documented in)	E31Document	E13AttributeAssigment
P89 falls within (contains)	SP6DeclarativePlace	E53Place
P89 falls within (contains)	E53Place	E53Place
P91 has unit (is unit of)	E54Dimension	E58MeasurementUnit
P94 has created (was created by)	E65Creation	E28ConceptualObject
P108 has produced (was produced by)	E12Production	E22HumanMadeObject
P127 has broader term (has narrower term)	E55Type	E55Type
P130 shows features of (features are also found on)	SP6DeclarativePlace	SP6DeclarativePlace
P140 assigned attribute to (was attributed by)	E13AttributeAssigment	E22HumanMadeObject
P168 place is defined by (defines place)	SP6DeclarativePlace	E94SpacePrimitive
P195 was a presence of (had presence)	E93Presence	E22HumanMadeObject
P195 was a presence of (had presence)	E93Presence	E19PhisicalComponent
Q9 is expressed in terms of	E94SpacePrimitive	SP4SpatialCoordinateReference
geo_hasGeometry	E94SpacPrimitive	SP15Geometry
GLP1 space primitive is defined by (defines space primitive)	E94SpacePrimitive	C1_Geometry_extracted from_maps
GLP1 space primitive is defined by (defines space primitive)	geo:Feature	C1_Geometry_extracted from_maps
geo_hasGeometry	geo:Feature	geo:Geometry
geo:sfWithin	geo:Feature	E94SpacePrimitive
geo:sfNearby	geo:Feature	E94SpacePrimitive

TABLE 5.15. An overview of each class's relationships regarding target and range within the knowledge graph.

source of information and is timestamped using an *E52 Time Span* class to identify when the data was recorded. This structure allows all pieces of information to be traced back to their source. Since the data in our system comes from a variety of sources and times, each record needs to be considered a distinct version of the object so that the provenance and timeframe of the information can be tracked.

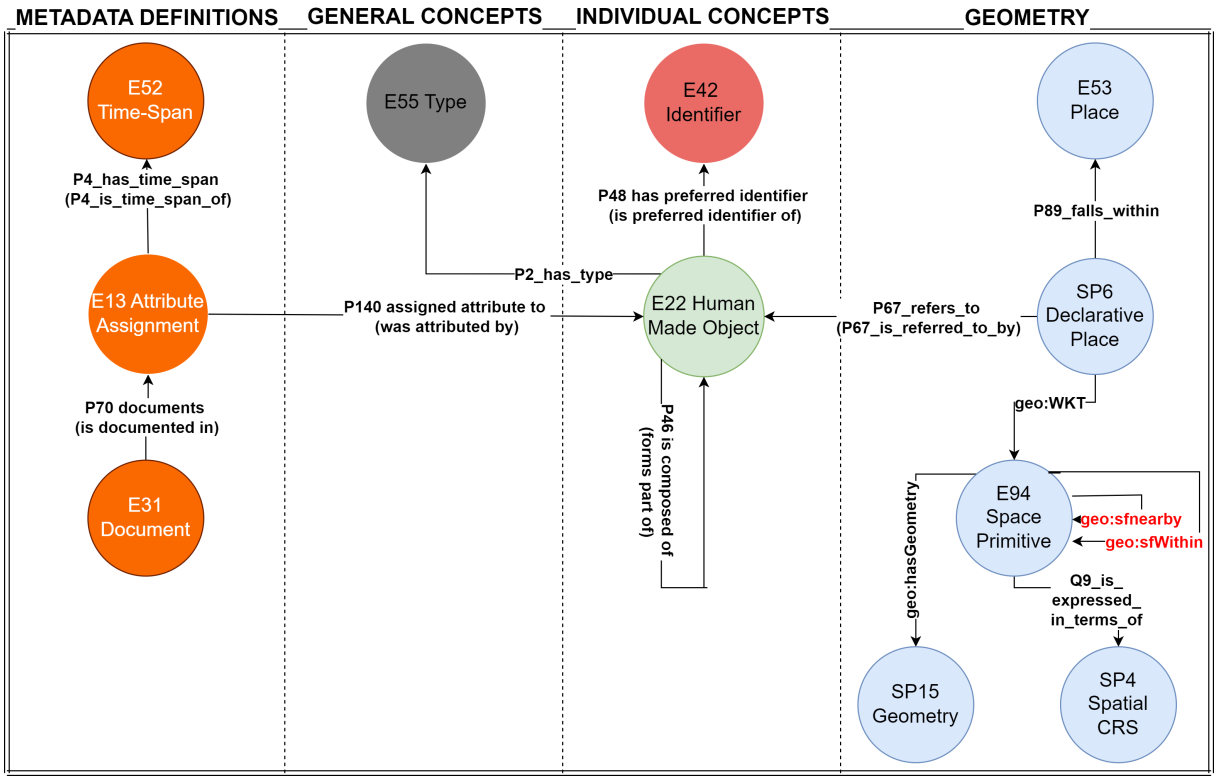


FIGURE 5.4. Concise schema for representing dolmens.

The *E22 Human-Made Object* class represents an individual archaeological monument concept. Dolmens, as enduring physical structures created by humans, fall under this class, the most specific of human-made items in the hierarchy. As described by [59], the *E22 Human-Made Object* class *"comprises all persistent physical items of any size that are purposely created by human activity."*. Each record about a dolmen results in an instance of the E22 node, regardless of the data source. In our case, with two distinct data sources, up to two versions of information can exist for a single monument, each linked through a global ID (*E42 Identifier* class), as shown in Figure 5.5. This approach allows multiple versions of information about the same monument to be represented.

Dolmens and their components are represented using the same *E22 Human-Made Object* class. Components within this class are treated as distinct instances. The hierarchical schema uses the relationship label *P46 is composed of* to link these different components and sub-components. This design transforms the dolmen node from a physical object into an abstract container defined by its components. This allows a granular description of the monument. The CRM framework aims to describe data structures at a high level, focusing on entities and relationships without specializing in structural or topological details. Existing CRM extensions, such as CRMba [203], handle topological relations of functional spaces. In our context, topological relations were not necessary for satellite image contextualization. However, the graph can be extended in the future to include such information if needed.

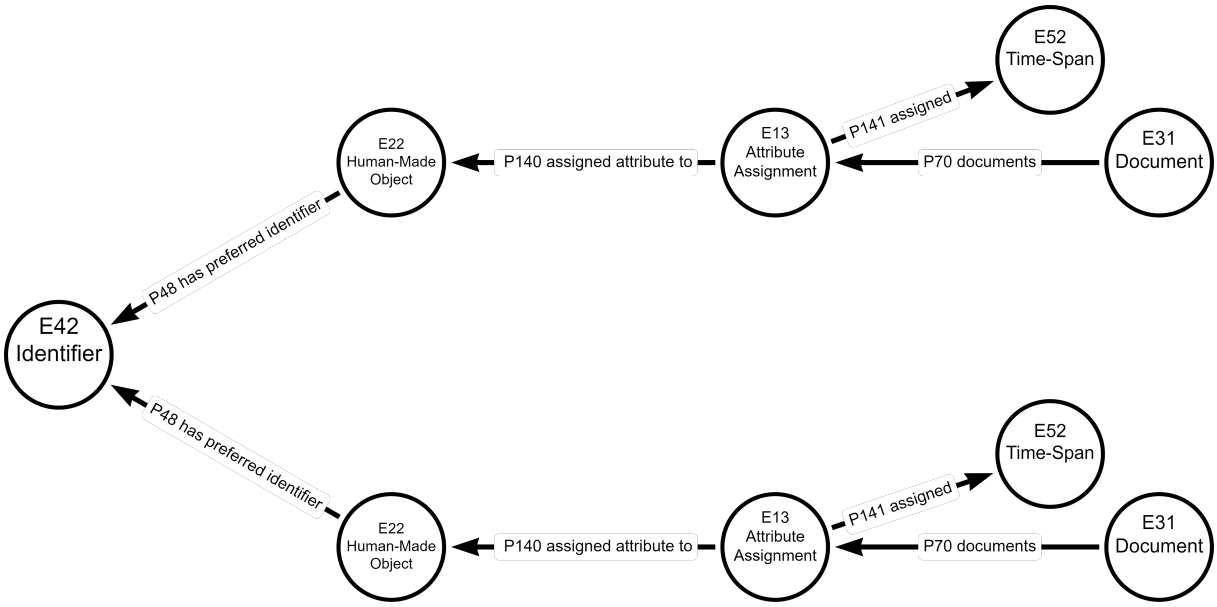


FIGURE 5.5. Multiple E22 instances linked to a global ID (E42 Identifier), demonstrating how multiple versions of information for a single dolmen can be represented and expanded to accommodate additional versions.

To describe the features of the dolmen components, we use several specific classes, such as *E57 Material* for construction materials, *E3 Condition State* for preservation state, *E54 Dimension* for measurements, and *E19 Physical Object* for features like capstones or orthostat numbers. Each *E22 Human-Made Object* is also linked to its geographical location represented in *SP6 Declarative Place* class. This establishes a clear connection between the monument and its spatial context. For a visual representation of the full schema model, please refer to the illustration provided in the Appendices 8.2.

5.4.2. Representing Landscape

The schema created to represent VTMs is shown in figure 5.6. This model employs the *E28 Conceptual Object* class from the CRM and showcases cartography as a conceptual object crafted to represent physical entities, which consist of human-produced data that have become objects of discourse. To represent who created the cartograph, we need to associate the event of creation between the object created and the person/group responsible for it. In this case, we use *E65 Creation* class to describe the cartograph creation event (E28) and relate them through *P94 has created*. The creation event (E65) is linked to the creation responsible through *P14 carried out by* and is described in class *E74 Group*. It is important to note that each cartograph represents the geographical objects present at the time it was developed. To determine a cartograph's date span, we use the class *E52 Time Span* and link it to the conceptual object represented (E28) through the relationship *P4 has time-span*.

To relate the conceptual framework of cartography (E28) to the physical elements described, we used the property P41 "classified (P41 was classified by)." This property links the general concepts represented in each vector map (E17 Type Assignment)—such

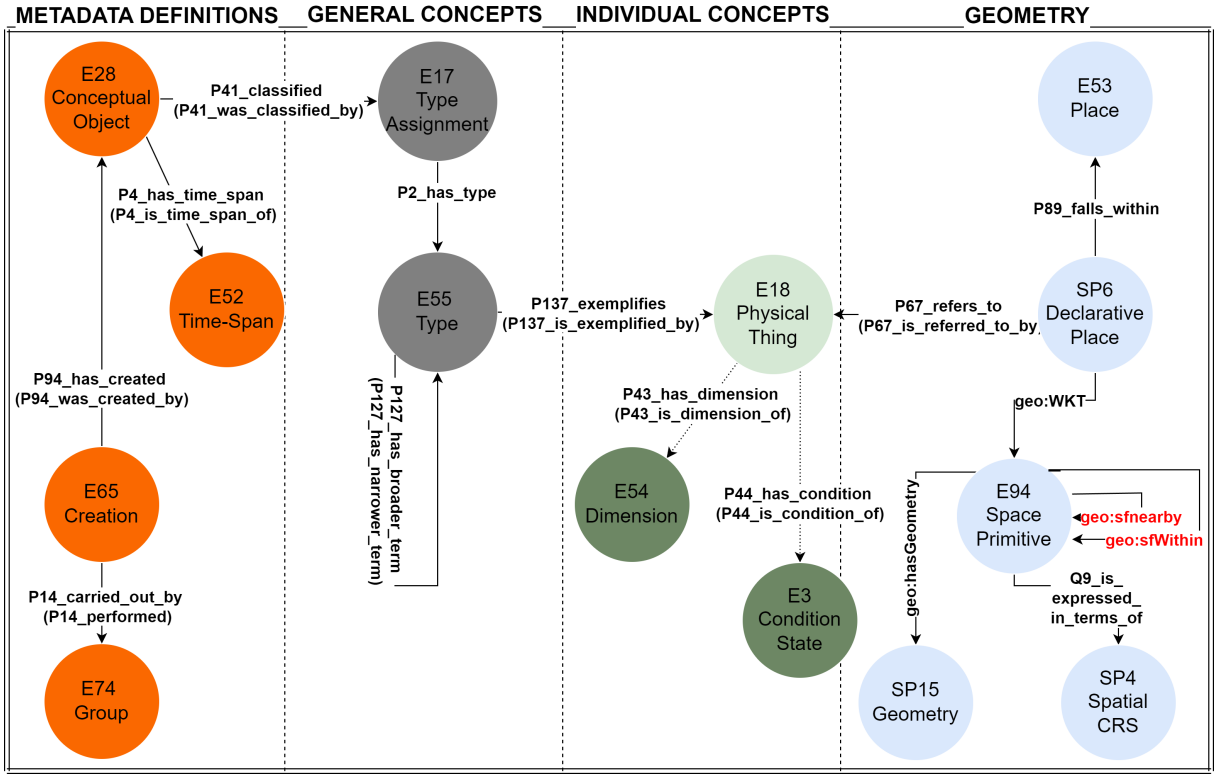


FIGURE 5.6. Concise schema for representing the vector topographic maps.

as "Hydrograph" or "Soil Use"—to the E55 Type class, which provides a more detailed definition of these concepts. For example, "Rivers" is used in hydrographic maps, while "Urban Areas" is included in land use maps. A hierarchical structure of concepts can be found in some maps, such as the COS 2018, which categorizes 85 types of land use (for example, forests and urban areas). We use the property P127 "has broader term (is a broader term of)" to link the hierarchical type (E55) to the general concepts depicted in the map (E17). This taxonomy facilitates the organization and classification of map concepts, offering a controlled vocabulary and detailed content information. To link the more specific concepts in the map (E55) with the physical objects they describe, we use the property P137 "exemplifies (is exemplified by)" to connect them with the E18 Physical Thing class, enabling a detailed articulation of individual features based on the classification.

In the E18 Physical Thing class, all instances representing natural elements can be related to other classes to capture their attributes. For example, a river is classified by its type—such as "natural" or "artificial" (E55 Type)—and its condition—such as "good" or "modified" (E3 Condition State). To describe the geometry and location of each element, instances of E18 are connected to the SP6 Declarative Place class using the property P67 "refers to" (or P67 "is referred to by"). This linkage allows for a precise representation of the spatial and locational attributes of each physical element, as illustrated in the cartographic data. This approach ensures that each element's characteristics and geographical context are represented and integrated into the overall analysis. For a visual

representation of the entire schema model, please refer to the illustration provided in the Appendices 8.1.

5.4.3. Representing Spatial Relationships

The schema model highlights the importance of spatial data in capturing instances and their relationships. In order to establish granular topological relationships between entities, specific relationships were defined to manage and link different spatial formats, such as lines, points, and polygons. Without that, we would be left with various subgraphs interconnected by their respective regions (E53 Place — e.g., Pavia); however, this representation lacked precision. Without it, we would be unable to determine the spatial relation between geographic objects represented in the model.

Spatial entities are intrinsically interconnected. GeoSPARQL provides a standardized way of representing these topological relationships. It implements the Simple Feature Access Common Architecture specification to describe spatial relations. This model is based on Egenhofer's extension of RCC8, a subset of Region Connection Calculus (RCC) that defines eight pairwise disjoint spatial relations. Egenhofer's work was further generalized in the Nine Intersection Model [252]. The Simple Features topological relations include equals, disjoint, intersects, touches, crosses, within, contains, overlaps, and relate as shown in Figure 5.7 [255].

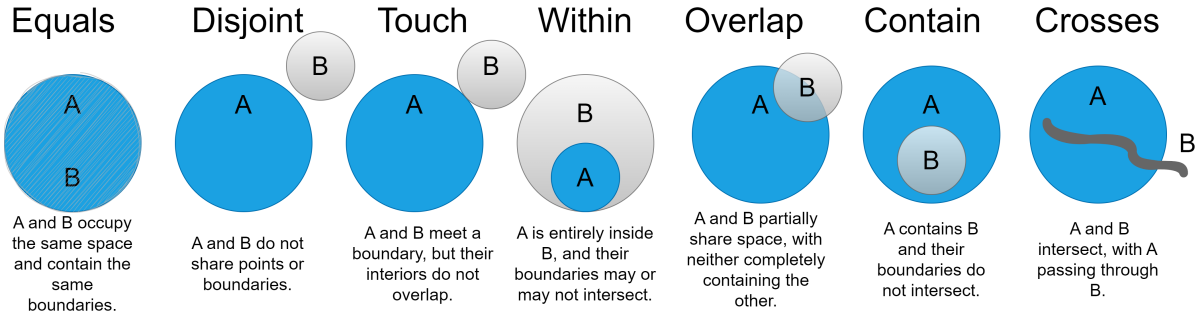


FIGURE 5.7. Illustration of spatial relationships between objects.

To enable a clear understanding of how entities relate to one another, we determined that specifying whether one element is within another and detailing the nearby between elements were sufficient for our use case. The GeoSPARQL's geo:sfWithin was strategically harnessed to express the topological relationship where one geometry is entirely contained within another. With this relationship, it is possible to infer, for example, when an *E22 Human-Made Object* is contained within a specific (E18 Physical Thing) by relating instances of the *E94 Space Primitive*.

While GeoSPARQL encompasses different topological relationships (Section **??**), it does not explicitly define an "Adjacent" relationship. The geo:sfTouches property can represent objects that share only a boundary, without any overlapping interior, potentially capturing the essence of adjacency in certain contexts. However, to address more nuanced spatial relationships and to capture specific instances of proximity with or without direct

boundary contact, there arises a need for a distinct relationship that could deal with the representation of proximity between monuments and landscapes.

In order to denote proximity, a custom property called "geo:sfNearby" was created. The bespoke relationship is then used to link E94 Space Primitive entities that are geographically close to one another (whether adjacent or not). In order to enrich this relationship, distances have been directly incorporated. In this way, proximity was established, as well as the exact value of distance between them. Table 5.16 summarizes spatial relationship properties. Following its creation, the property "sfNearby" can be enriched with specific distance metrics in meters to provide a comprehensive representation of physical proximity between entities. By embedding specific distance metrics directly as properties of these relationships rather than as attributes of nodes, a direct linkage of spatial relationships to entities is established. This model provides a clear distinction between raw data sourced from primary materials, such as reports and vector maps, and information derived from geospatial analyses.

Spatial Relationship Properties	
E94: sfWithin:E94	An sfWithin relationship links E94 classes that indicate when a point is entirely contained within a polygon. They indicate when monuments fall into "Land Use" or "Soil Type" — but can also be used to analyze when other E18s fall within other E18.
E94: sfNearby:E94	An sfNearby relationship connects E94 classes that indicate closeness. They are used to indicate the position of monuments in relation to other instances of E22 or E18. Here, they are used to specifically relate the monument's position to its "Hydrography", "Contour Lines", and "Dolmen". In this relationship, the distance between each domain-range class (E22 or E18) is embedded.

Table 5.16. Definition of spatial relationships.

These spatial relationships allow us to represent how different landscape components and archaeological features are interconnected. By defining these relationships, we enhance the graph's capability to model and interpret spatial dependencies. This is essential for understanding archaeological monument interactions within their environment and contextualizing scenes.

5.5. Model Implementation

To implement the LPG schema model we have used Neo4j graph database [3] and data was mapped from the files to the database. Neo4J is a schema-free, NGDB built on properties and characterized by its Cypher query language. The Cypher language relies on relatively complex patterns that, when used effectively, can provide insights that are not readily available from traditional database engines [256]. The high readability of Cypher,

[3]The Neo4j graph database can be downloaded at: https://neo4j.com/

coupled with the ability of Neo4j to present query results in a variety of formats, allows for enhanced flexibility and diverse interpretations of data [257]. Although implementation may be time-consuming, queries are typically less complex and execute faster than those in PostgreSQL, making this system a strong candidate for data storage and analysis [257]. In addition to its high functionality, it complies with ACID principles [237].

When compared with other graph databases Neo4J consumes less memory for processing [258], performs better through indexing techniques for queries retrieval performance, and gets the best results with traversal workloads [259, 260, 261]. A graph database like Neo4J is not designed to handle basic graph patterns and atomic lookups [236] or to handle search based on a limited number of relationships (low number of JOINs in SQL databases) [238]. In contrast, they are ideal for applications involving multiple levels of relationships between data [240, 238] — ideal for applications involving relationships between instances [239] and finding patterns [259].

Representations, searches, and retrievals of spatial geometry are supported via points linked to a specific Coordinate Reference System (CRS), whose values are represented in Cartesian coordinates or WGS-84. The schema-free nature of Neo4J allows labels, types, and properties to be applied to nodes and relationships in a flexible way, although this can sometimes lead to vague semantics. It is not only possible to represent archaeological sites in this structure but also to perform multi-relational searches for each one. The flexibility of Neo4J enables the construction of a model based on well-recognized ontologies (CIDOC-CRM, CRMgeo, and GeoSPARQL), which can represent semantic and spatial data in detail. The Neo4J desktop (V: 1.5.6) was the platform of choice for data integration.

To populate the LPG, the structured file derived from data standardization was used. In this mapping process, table columns were mapped to **property keys** and rows to **values**. Each column header corresponds to a predefined property of each class, while each row represents an instance of that class. Our focus was on data mapping and curation, linking entities, matching them to their respective instances, and ensuring that the data was correctly connected to facilitate querying and information retrieval. Table 5.17 lists the classes (node labels) used, and also includes the definitions of property key attributes used to represent this information, serving as keys for assigning values within the database.

Even though data sources contained information of a larger area, such as the entire country, only information about the AOI was retained. This targeted approach can lead to efficient resource allocation. Additionally, it helps to minimize data overload and ensures that the most pertinent information is readily accessible. The LPG contains 141,380 nodes and 370,177 relationships. The model can, however, be expanded to include more information if needed.

Data insertion was performed with and without indexing to assess how indexing affected the import process efficiency. Indexes were applied to enhance data retrieval speed and performance. The performance of searches with and without indexes was consistently

NODE LABELS	PROPERTY KEYS	DESCRIPTION
E1 CRM Entity	E1CRMEntity	Indicate the broader concept of E22 (e.g. Megalithic Monument).
E3 Condition State	E3ConditionState	Comprises the E22 condition state.
	E3ChemicalConditionState	Comprises the chemical condition of E18 - Rivers.
	E3EcologyConditionState	Comprises the condition of E18 - Rivers.
E12 Production	E12Production	Indicate the E22 production.
E13 Attribute Assignment	E13AttributeAssigment	Indicates actions for asserting E22 properties.
E17 Type Assignment	E17TerminologyDefinition	Indicates the broader concept of classified terminology for the E28.
E18 Physical Thing	E18SpecificResource	Comprises all instances of natural physical things (e.g. River X).
E19 Physical Object	E19Components	Indicates the components of the E22.
E22 Human-Made Object	E22Dolmen	Comprises instances of archaeological sites (e.g. dolmen X).
	E22DolmenComponent	Indicates the E22 parts.
E28 Conceptual Object	E28ConceptualObject	Indicate the source of the information for the E28.
E31 Document	E31Document	Indicate the source of information for the E22.
E41 Appellation	E41Designation	Comprises the designation of the E22.
	E41DesignationMaps	Indicate the designation that identifies the E28.
	E41AppelationOfE18	Comprises the designation of E18.
E42 Identifier	E42GlobalID	Indicate a global identifier to each unique E22.
	E42LocalID	Comprises the identifier of the monument as indicated by E31.
	E42AssignedTypeID	Comprises the identifier of the E18 as indicated by the E28.
	E42PolygonID	Comprises the identifiers for each geometry extracted from the E28.
E52 Time-Span	E52TimeSpanHumanMadeObject	Indicate when E22 was built.
	E52TimeSpanDocument	Indicate when E31 was released.
	E52TimeSpanAcquisition	Indicate when E31 information was extracted.
	E52TimeSpanMaps	Indicate when E28 was released.
E53 Place	E53Place	Indicate the place designation for E22 and E18.(e..g, municipality and parish).
E54 Dimension	E54DimensionValue	Indicate any dimension value type used to describe an E22 or E18 feature.
E55 Type	E1CRMEntityType	Indicate E22 in a narrower context (e.g. dolmen).
	E55TerminologyType	Comprises the terminology provided by E28 to represent E18.
	E55TypeObject	Indicate E28 representation format.
	E55ChamberShape	Comprises the shape of the E22.
	E55TypeOfDocument	Indicate the E31's data type (e.g. book).
	LocalizationDistribution	Comprises E19's position (e.g. in situ).
E57 Material	ChamberMaterial	Comprises the material of the E22.
E58 Measurement Unit	E58UnityOfValue	Indicate the value unity type used to represent measurement (e.g. meters).
E60 Number	ComponentsParts	Comprises the number of parts for E19 and the number value for E54.
E65 Creation	E65CreationEvents	Indicate the events leading to the creation of E28 and E13.
E74 Group	E74Group	Indicate who created the E28 and E31.
E93 Presence	PresenceDescription	Comprise information about the presence of parts E22 (e.g. maybe).
E94 Space Primitive	E94GeometryWKT	Comprises the coordinates of E18 and E22 in WKT.
SP4 Spatial Coordinate Reference System	SP4Datum	Indicate the type of the E94.
SP6 Declarative Place	SP6DeclarativePlace	Indicate the place defined by the E94 and relate it to E22 or E18.
SP15 Geometry	SP15Geometry	Indicate the type of geometry in E94.
geo:Feature	GridID	Comprises the grid implemented .
geo:Geometry	GeometryWKT	Location in WKT format
C1_Geometry_extracted from_maps	GLE1	Comprises all information extracted of E18 and geo:Feature that allows insert it into GIS.

TABLE 5.17. The node labels, corresponding Properties Key definitions used in the graph database, and descriptions of the information contained in each node.

better with indexes. As an example, in query 5.1, the query to retrieve the relation between E22 and E13 is provided. For this analysis, database access decreased from 284903 to 2395 (db hits), and the cost of run time decreased from 76ms to 13ms. As a result of this query, the information about E22 pertaining to E13 — each monument analyzed along with its source — is retrieved.

LISTING 5.1. Search example

```
MATCH (n:E22_Human_Made_Object)-[r:P140_was_attributed_by]-(m:E13_Attribute_Assignment)
RETURN n, r, m
```

In all our queries, performance gains were more apparent when searching for spatial relationships between multiple entities. In query 5.5, the database hits decreased from 1176955 in 198 ms to 646967 in 52 ms for the database with indexes. This query returns all monuments classified as destroyed, along with their soil type, last use (2018), the natural rivers closest and if they are in a relief area.

```
MATCH dolmen=(E31Document:E31_Document {E31Document})-[:P70_documents]->(:
    E13_Attribute_Assignment)-[:P140_assigned_attribute_to]->(Dolmen:
    E22_Human_Made_Object)-[:P55_has_current_location]->(:SP6_Declarative_Place)-[:
    P168_place_is_defined_by]->(E94PolygonWKT:E94_Space_Primitive),

(Dolmen)-[:P46_is_composed_of]-(chamber:E22_Human_Made_Object {E22DolmenComponent: ''
    Chamber''})-[P43DolmenDimension:P43_has_dimension]-(E54TypeOfDimension:E54_Dimension
    {E54TypeOfDimension: ''Diameter''})-[:P90_has_value]-(ChamberDiameter:E60_Number),
    (chamber)-[:P44_has_condition]-(Condition:E3_Condition_State),

WATER=(E17Water:E17_Type_Assigment {E17TerminologyDefinition: ''Hidrografia''})-[
    P2A1Water:P2A1_assigned_definition_types]->(E55Water:E55_Type)-[P137Water:
    P137_is_exemplified_by]->(WaterLine:E18_Phisical_Thing)-[P67Water:
    P67_is_referred_to_by]->(SP6Water:SP6_Declarative_Place)-[P168Water:
    P168_place_is_defined_by]->(E94Water:E94_Space_Primitive)<-[nearbyWater:geo_sfnearby
    ]-(E94PolygonWKT),

RELIEF=(E17Relief:E17_Type_Assigment {E17TerminologyDefinition: ''Hipsometria''})-[
    P2A1Relief:P2A1_assigned_definition_types]->(E55Relief:E55_Type)-[P137Relief:
    P137_is_exemplified_by]->(E18Relief:E18_Phisical_Thing)-[P67Relief:
    P67_is_referred_to_by]->(SP6Relief:SP6_Declarative_Place)-[P168Relief:
    P168_place_is_defined_by]->(E94Relief:E94_Space_Primitive)<-[nearbyRelief:
    geo_sfnearby]-(E94PolygonWKT),
    (E18Relief)-[P43Relief:P43_has_dimension]->(E534Relief:E54_Dimension)-[P90Relief:
        P90_has_value]->(E60Relief:E60_Number),

SOILTYPE=(E17SoilType:E17_Type_Assigment {E17TerminologyDefinition: ''SoilType''})-[
    P2A1SoilType:P2A1_assigned_definition_types]->(E55SoilType:E55_Type)-[P137SoilType:
    P137_is_exemplified_by]->(E18SoilType:E18_Phisical_Thing)-[P67SoilType:
    P67_is_referred_to_by]->(SP6SoilType:SP6_Declarative_Place)-[P168SoilType:
    P168_place_is_defined_by]->(E94SoilType:E94_Space_Primitive)<-[SoilTypeWithin:
    sfWithin]-(E94PolygonWKT),
```

```
SOILUSE=(E17SoilUse:E17_Type_Assigment {E17TerminologyDefinition: ''SoilUse''})-[
    P2A1SoilUse:P2A1_assigned_definition_types]->(E55SoilUseLevel1:E55_Type)-[
    P127SoilUse1:P127_has_narrower_term]->(E55SoilUseLevel2:E55_Type)-[P127SoilUse2:
    P127_has_narrower_term]->(E55SoilUseLevel3:E55_Type)-[P127SoilUse3:
    P127_has_narrower_term]->(E55SoilUseLevel4:E55_Type {b: ''COS2018''})-[P137SoilUse:
    P137_is_exemplified_by]->(E18SoilUse:E18_Phisical_Thing)-[P67SoilUse:
    P67_is_referred_to_by]->(SP6SoilUse:SP6_Declarative_Place)-[P168SoilUse:
    P168_place_is_defined_by]->(E94SoilUse:E94_Space_Primitive)<-[SoilUseWithin:sfWithin
    ]-(E94PolygonWKT)

WHERE Condition.E3ConditionState = ''destroyed''

RETURN *
```

In terms of hardware requirements, a CPU Core i7 with a memory of 16 GB is recommended. The test execution was carried out on a machine with the following configuration: AMD Ryzen 7 5800 8-Core Processor, 3401 Mhz, 8 Core(s), 16 Logical Processor(s) with 16.0 GB of RAM.

5.6. Information Retrieval and Discussion

The LPG model provides a robust basis for representing contextual and spatial information. The model implemented consists of several subgraphs derived from the different types of data sources used. With the LPG, we aimed to identify topological and feature multi-relationships between entities based on semantic and spatial relations. The Figure 5.8) illustrates how the different subgraphs are spatially related to each other, providing a visual representation of how geographic objects are interconnected.

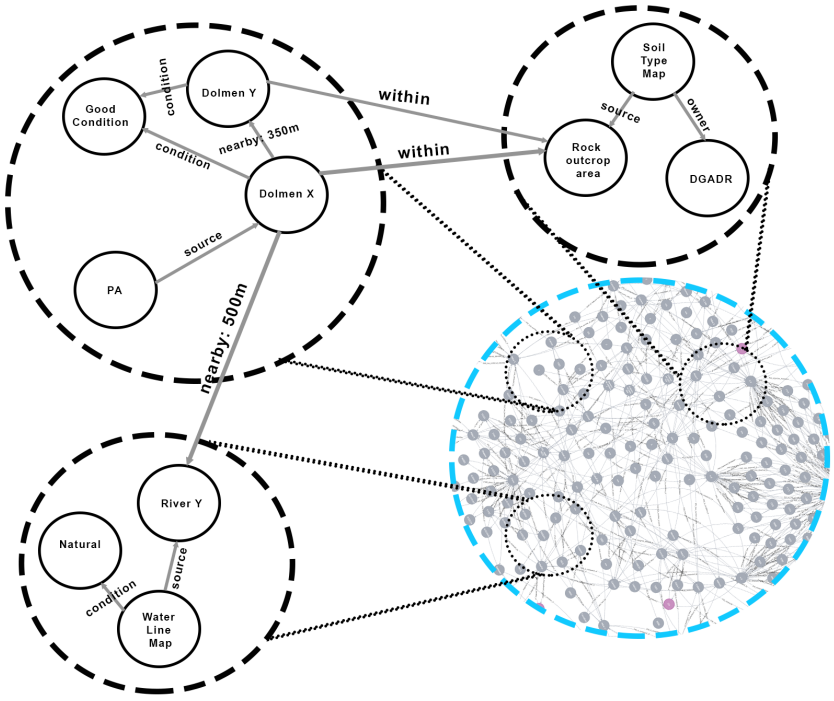


FIGURE 5.8. Visual illustration of the constructed spatial-LPG.

A spatial analysis of monuments and their surrounding landscape features is conducted at the instance level. To ensure that the spatial relationships between geographic objects were correctly interconnected and to verify this, we performed data retrieval. To achieve this, we conducted retrieval tests using Cypher queries to return all nodes and relationships related to the monuments and their connections to the landscape, verifying the results against the original source data. For all queries, whether involving one or multiple relationships, the results were returned in milliseconds. Additionally, spatial relationships were manually confirmed using GIS. While the retrieval process and the GIS verification are separate, both approaches confirm the accuracy of the spatial relationships. This suggests that the LPG accurately represents the entities and their spatial and topological relationships, which facilitates data retrieval.

We utilized the LPG to retrieve information that allowed us to confirm known truths about the spatial distribution of monuments within the area of interest. In this case, it is widely accepted that monuments in this AOI are typically located near watercourses, rocky outcrops, areas with gentle slopes, and soils with low agricultural capacity — I call these characteristics of Classification Keys that show the interplay of dolmens with the environmental context. Additionally, it is also well-established that monuments tend to be located in close proximity to one another.

Our analyses were conducted both at the level of collective (group analysis) to confirm known patterns and at the individual level to understand their specific spatial contexts. This aimed not only to confirm this known domain knowledge but also to uncover nuances and particularities that could further enrich our understanding of these monuments and their interactions with the landscape.

For example, group analysis for dolmens' proximity was performed. Within the studied region, the average distance between dolmens of the same type is approximately 300 meters. In fact, 84% of the dolmens (62 out of 67) were within less than 1000 meters from one another. This data demonstrates a clear tendency for these sites to cluster together.

In addition to group analysis, we performed analyses focused on individual instances to identify broader spatial patterns among the sites. For example, Figure 5.9 illustrates the proximity between the dolmen Anta Capela de S. Diniz (D11) and its closest monument, dolmen Ferragial de Nossa Senhora (D34), as returned by the LPG subgraph. The figures represent the data sources using a colour scheme: Spatial and geometric details are shown in blue, general concepts are shown in light green, specific concepts are shown in dark green, descriptions of concepts are shown in yellow, and metadata details are shown in beige. This information not only provides insights into the spatial relationships of the dolmens but also includes metadata – indicating the source of information (PA), specific attributes (conservation states: good for D11 and poor for D34; measurements: 4 meters in diameter for D11 and not applicable for D34, and geometric details (spatial coordinates).

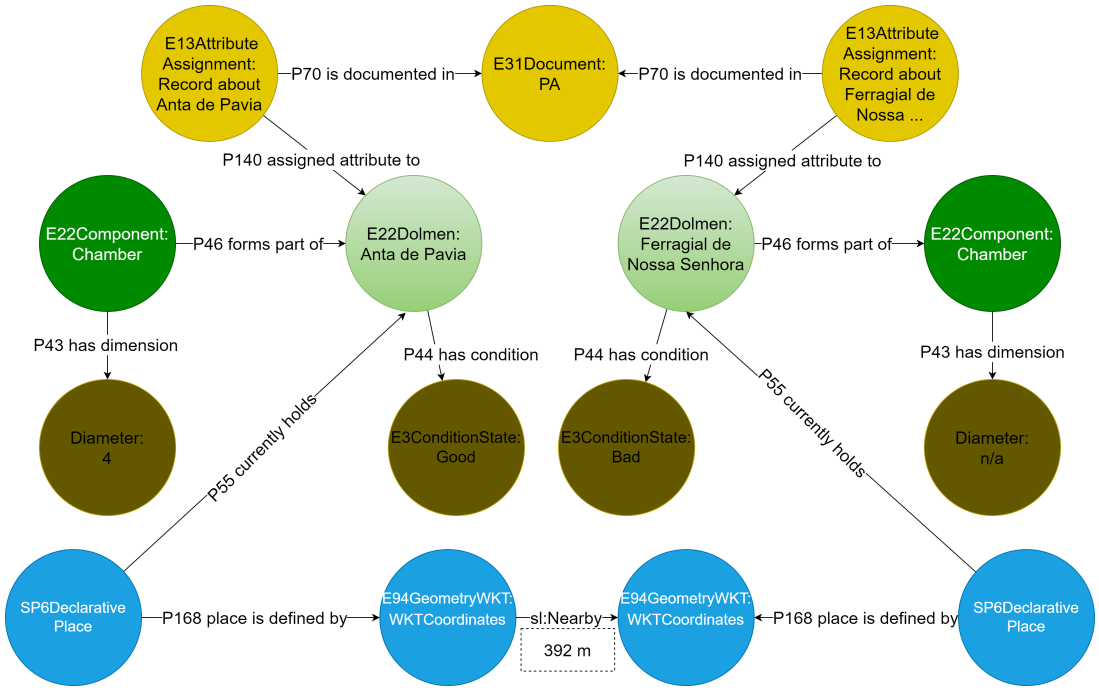


FIGURE 5.9. Proximity between dolmen Anta Capela de S. Diniz (D11 – pink node) and the nearest dolmen, Ferragial de Nossa Senhora (D34)

An example of information retrieval to analyze the interaction between specific dolmens and all geographic objects related to the landscape surroundings is shown in Figure 5.10, which shows the interaction between dolmen D11 and its environs. In the image, the data sources, soil type where the monument is located; the soil use, indicating the current land use in the area (urban area); the water lines, showing the distance between the dolmen and the nearest waterline (700m) and precisely which waterline is nearby (Ribeira da Tera); a hypsometry level indicating that the dolmen is situated at an elevation of 190 meters is shown. The figures represent the data sources using a colour scheme: Spatial and geometric details are shown in blue, general concepts are shown in light green, specific concepts are shown in dark green, descriptions of concepts are shown in yellow, and metadata details are shown in beige. Spatial relationships are depicted using "within" (in blue) and "nearby" (in red) to show how different entities are interconnected. It is possible to retrieve the interaction between all of these entities, regardless of whether they are adjacent or not. We prioritized the E18 Physical Thing class closest to the E22 Human-Made object class in this case. It is possible to analyse how each thing interacts with its surroundings based on any coordinate point.

This analysis is advantageous because it integrates multiple spatial relationships, links multiple subgraphs, and provides a comprehensive semantic understanding of the interconnected elements. Visualizing these connections highlights interactions among entities within a broader context. The search can be extended to retrieve information on all monuments, allowing queries that identify patterns or define specific criteria, such as finding structures with one or several similarities or differences. This flexibility enables users

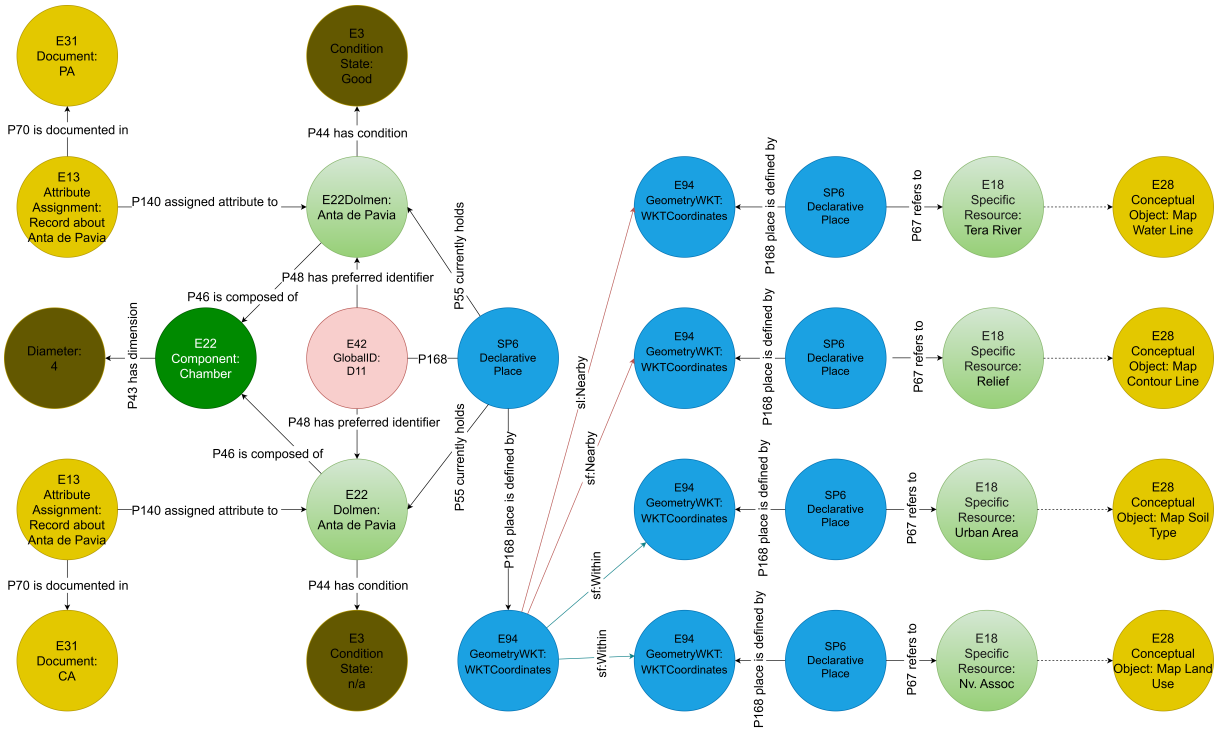


FIGURE 5.10. Information retrieval of the Anta Capela de S. Diniz and its spatial relation with the landscape.

Monument	Water Line	Topographic Relief	Soil Type
Tera 4	Tera River	165m	Pmg
Forca Velha	Tera River	180m	Pm
Monte das Figueiras	Tera River	170m	Pg
Tera 5	Tera River	170m	Pg
Adua 5	Tera River	175m	Pgn
Ferragial da Fonte	Tera River	175m	Pgn

TABLE 5.18. Retrieve information where monuments classified as "destroyed" share similar characteristics – water line, topographic relief and soil type.

to refine searches based on semantic relationships, uncovering insights about monuments that share common features or exhibit distinct characteristics. For example, which monuments share the same condition (e.g., 'destroyed'), water line proximity (Tera River), topographic relief areas and soil type and, in terms of spatial relations, are they in proximity of each other. Table 5.18 shows an example that highlights this retrieval of information.

Using queries, it was possible to perceive that, in spite of the AOI's predominantly flat terrain, which spans altitudes between 50 meters and 205 meters, revealed a marked preference for site placement on slightly elevated terrain. A total of 37 dolmens were found at altitudes exceeding 160 meters, 25 located between 100 meters and 159 meters,

Dolmen	Quantity	Soil Type	Condition
Outeiro da Forca	1	Pdg	destroyed
Têra 4	1	Pmg	destroyed
Entreáguas 4; Anta do Forno	1	Sr	destroyed
Forca Velha	1	Pm	destroyed
Folha da Anta	1	A	destroyed
Entreáguas	1	Vt	destroyed
Freixo	1	Pg, Arg	destroyed
Pucícaros 2, Pucícaros 1	2	Pdc, Pdg	destroyed
Monte das Figueiras, Têra 5, Madre de Deus 1, Casa Branca 2	4	Pg	destroyed
Oliveira 2; Anta do Cabeço da Anta, Adua 5, Gonçala 2, Ferragial da Fonte	4	Pgn	destroyed

TABLE 5.19. Retrieving data from the source "PA" about monuments described as "destructed" and the soil type on which they are situated.

and 11 below 100 meters. The majority of these sites are located along prominent terrain reliefs.

Diverse analyses can be performed to return information about spatial relationships between E18 Physical Thing and E22 Human-Made Object classes. For example, contour lines can be analyzed by examining whether steep slopes or reliefs follow local contour lines around monuments and how these interact with water lines or soil use. This allows for an assessment of whether the areas adjacent to the monument are higher or lower, which can provide insights into how these choices influence the monument's condition or help with theories of choices for its placement.

By overlaying dolmen locations with regional soil types, a clear relationship between monuments and specific soil characteristics emerged. The analysis confirmed that dolmens are predominantly situated on PG (Litholic Soils) and ARG (Rock Outcrops) soils. PG soils are shallow and rest atop granite bedrock, providing excellent drainage and limiting deep root growth. In contrast, ARG soils are located near exposed granite or quartz diorite rock, offering minimal soil cover. These soil types not only influence vegetation and land but can impact decisions related to settlement and construction.

Only a few sites appeared in the following soil types: *Argiluviados Pouco Insaturados* (Pgn, Sr, Pdc, Pdg, Pmg, Pag, Pm, Pac and Vcm), *solos incipientes* (A, At) and other types of *solos litólicos* (Vt and Par). No monuments were found in areas of black, brown, and reddish brown clay, *calcarios*, and *Hidromórficos* soils. *Barros* are evolved soils of profile A, B or C as well as solos *calcarios* and *podlizados* soils (Servico de Reconhecimento e de Ordenamento Agrário 1970). When soil's conservation status is correlated with monument destruction, it was observed that "destroyed" monuments were predominantly located in PGN and PG areas, as shown in table 5.19.

The information extracted from the soil type map contained information coded as symbols, so we mapped the explanatory legend of soil types into the LPG, enhancing understanding and accessibility by providing contextual information and enabling seamless connections between related data. Additionally, Appendix 8.1 includes a table with explanation notes for each code.

Today's landscape reflects generations of ecological shifts and human interventions. A number of factors can prevent the identification of the megalithic structures, including the landscape (such as in urban areas or areas with high vegetation that difficult the observation of the monuments). The landscape in the area is dominated by Holm Oak Agroforestry Systems (SAFs) and most monuments (43) in different conservation states are located in these areas. The monuments' concentration spots are followed by pastures (9 monuments), temporary crops (6 monuments), oak forests (4 monuments), eucalyptus plantations (3 monuments), olive groves (3 monuments), and SAF cork oak agroforestry systems (2 monuments). Other land uses all have one monument.

Besides the SAF zone, the improved pasture areas contain the most destroyed classified monuments (5), with no one monument classified as in good condition. For understanding soil evolution related to the monument position, the land use representation is structured to enable spatiotemporal retrieval (figure 5.11). It was found that land use changed very little between 1995 and 2018. As a result, we focus on the most recent land use data to examine how it relates to dolmen's positioning.

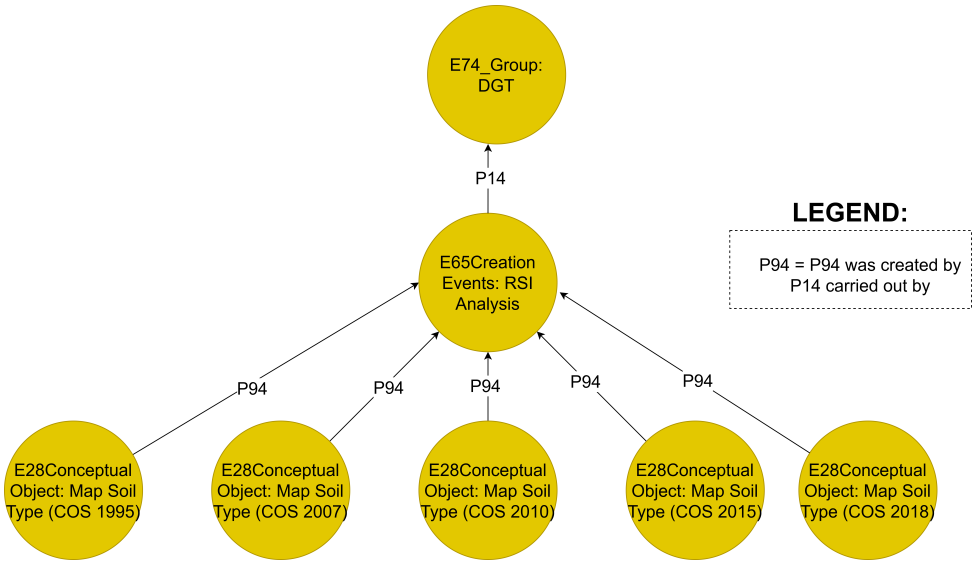


FIGURE 5.11. Visual representation of the knowledge graph for all COS. They can share similar classes, but each polygon is associated with a specific source of information.

Considering that land use patterns have changed since the construction of archaeological monuments such as those built during the Neolithic time, it is important to emphasize that land use patterns do not impact the placement of dolmens since they reflect modern uses. Instead, the current land use allows us to contextualize the dolmens within their

present landscape. It provides a better understanding of their location relative to contemporary land use patterns. It is possible that these land uses contributed to the monument's degradation, and this information can be used to understand this relationship.

Additionally, an analysis was conducted to determine the relationship between dolmen and water sources. Around 49% of the dolmen were located between 30 and 400 meters from a water source. This number increases to almost 90% when the area increases to 1 km. The maximum distance measured from a water source was less than 2 kilometres. This proximity to water suggests a correlation between dolmen locations and access to water. This could reflect preferences for placing monuments near vital resources.

The river near which most monuments are found is the Tera (33 monuments), followed by Matalote (8), Almadafe (5), Sorraia (3), Raia (3), Divor (2) and Freixo (1). In total, 7 of these are in good condition near the Tera River, 2 are in Almadafe, and 1 is in both Divor and Raia. Figure 5.12 provides a detailed view of the 2 monuments and their proximity to Ribeira da Tera and each other. The queries can be extended to understand the relationship between all things based on specific information about the monuments, including size, shape, material, or specific rivers such as names, composition and state of these and their geospatial relationship with each other. It is important to note that the hydrography data used for this analysis reflects the present-day situation. While the current proximity of dolmens to water sources offers valuable insights, it is crucial to recognize that waterlines and other geographical features may have changed over millennia. This is due to natural processes and environmental shifts.

FIGURE 5.12. A direct path to the final entity is shown — bypassing intermediate connections that connect other entities sequentially — showing how Remendo 2 and Tera 1 relate spatially to Tera River.

Based on the patterns identified through data retrieval, we confirm classification keys to use as domain rules since these features provide valuable insight into dolmen's spatial

relationships and contextual significance. These domain keys can be used as conceptual guidelines derived from previous observed patterns and correlations, which help in understanding the typical characteristics and spatial arrangements associated with dolmens and their surrounding landscape. Table 5.20 illustrates the main patterns observed in the placement of the dolmens relative to their surroundings.

Classification Keys: Interplay of dolmens with Environmental Contexts	
Proximity Among dolmens: ($\leq$ 1km distance)	**Site Density:** When a site is identified in a specific location, there's a high likelihood of finding at least two more dolmens within an immediate surrounding. **Average Proximity:** These sites tend to be located at an average distance of 300m – with 80% of these with $\leq$1Km from another. **Isolation Exceptions:** While many sites are clustered, some are exceptions and can be found isolated ($\geq$ 1km $\leq$3km).
Proximity with water lines: ($\leq$ 1km distance)	**Average Proximity:** On average, these sites are located within a distance of 417 meters from natural water sources. **Predominant Proximity:** A majority of these sites(approximately 80%) are situated within 1 km of a drinkable water source. **Cluster Indication:** Most of these sites are near water lines, especially along the Tera river in AOI.
Located near rock outcrops: ($\leq$ 1km distance)	**Geological Preference:** These sites are mainly located on PG and ARG soil types — granite soils. **Soil Depth and Rock Proximity:** Predominant attention to areas where the soil is either minimal or non-existent due to the underlying rock's proximity;
Located in relief areas: ($\geq$160m hight)	**Relief Preferences:** There are mainly sites above 160m altitude.

TABLE 5.20. The main interactions between dolmen and their surroundings that have been identified

These insights provide a deeper understanding of the relationships between the objects analysed and their surrounding landscape elements, allowing for a more comprehensive analysis of their context. The retrievals have validated theories regarding the placement of these monuments, showing their tendency to be situated on elevated terrains and near rocky outcrops and water lines [56, 88, 86].

CHAPTER 6

KG-ML Model Implementation

6.1. Introduction

Object detection approaches that leverage knowledge-based systems, where semantic information is used to provide context, rely heavily on a well-structured database. Such a database must contain detailed and interconnected information that can be used to enhance data-driven methods. By combining object detection with semantic enrichment and classification, the goal is to enhance both the precision and explainability of archaeological monument recognition. Ultimately creating the basis for a robust reviewer for object detection that helps deal with the false detections returned and decreases the necessity of extensive manual evaluation of incorrect outputs.

To explore the application of this concept in satellite image object detection for the recognition of archaeological monuments, we developed a new approach. Specifically, we created a Labelled Property Graph (LPG) to represent the monument and its surrounding context, employing this structured representation as the foundation for a Knowledge Graph (KG)-Machine Learning (ML) reviewer for object detection outputs. Developing such a hybrid approach requires a detection-based approach, a semantic model, and a combination of both. This chapter explains the methodology and results of our KG-ML model. It begins by a detailed explanation of the data acquisition process and describes how the collected data was prepared for use in Section 6.2. Next, it introduces the proposed algorithms in Section 6.3 and discusses the implementation process, focusing on the training strategies, model testing, and the metrics used to analyze the outputs in Section 6.4. Finally, the chapter concludes with a presentation of the results in Section 6.5, discussing the findings, their implications, and how they align with the study's objectives (Section 6.5).

In this approach, we aimed to illustrate a potential use case for the implemented LPG and assess whether, with the information structured in this way, the model could learn the contextual patterns that experts have identified as significant in the positioning of this type of monument.

6.2. Data Acquisition and Preparation

The data used in this phase comes from the object detection approach described in Chapter 4. The object detection data consists of the outputs obtained from detecting potential monuments in the area of interest. This data includes information on the spatial coordinates of the image, along with scores assigned to each detection within an image, referred to as Point of Interest (POI)s. Each POI is defined by the image's ID, its spatial

coordinates, and the detection score (e.g., FPC1: 38.913757, -7.993010, tensor(0.9950)). We used the POIs from Dataset 2, which comprises results from 16 images with dolmens and 70 images without dolmens. We also included POIs from a dataset referred to as the 'prediction set'. This dataset contains results from 100 images from the same area of interest, which had no monuments but returned 64 false detections. The POIs of the analyzed locations are shown in Figure 6.1.

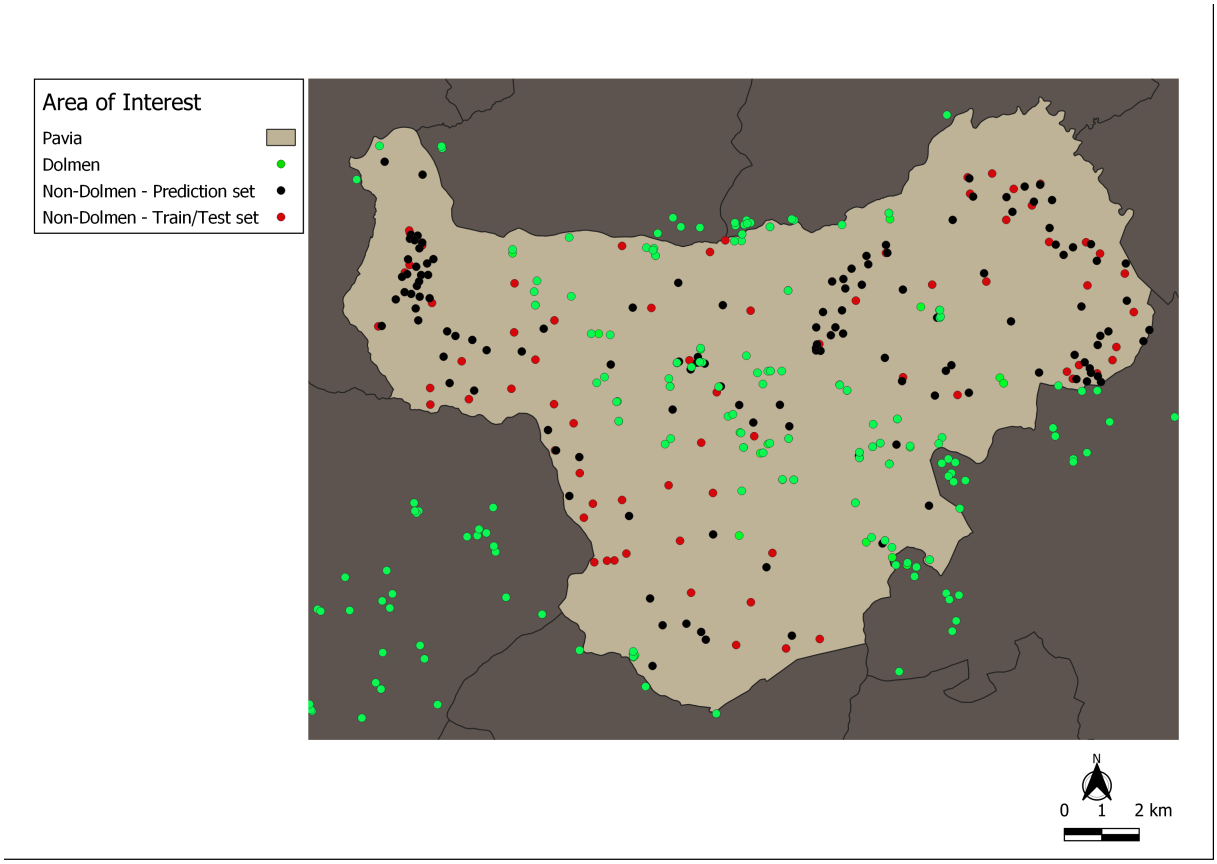


FIGURE 6.1. The map highlights the AOI, with green POIs representing sites used for training/testing and red POIs indicating non-sites. Black POIs show the predicted sites after the model has been trained and tested.

In the image analysis approach, only 16 monuments were considered across the available images, which limited the number of monuments that could be analyzed using object detection, consequently restricting the number of monument locations that could be used to train our KG-ML model. Out of these 16 POIs, outputs from 14 were consistently retained for the training and testing phases, while the remaining 2 were excluded and reserved exclusively for the prediction phase. However, the total number of monuments in the area of interest is 94, and these 16 detected monuments represent only a fraction of the full set. The other 78 monuments were not detected during photo-interpretation tasks due to various challenges, such as their conservation state (e.g., destroyed or in poor condition), coverage by modern layers, or other factors, and since they were not visible in the manual interpretation analysis, they were not analyzed by the automated object detection.

To overcome the limitation imposed by the visible monuments, we leveraged the full set of 94 monuments in our LPG to train our KG-ML model. By utilizing the known locations of all monuments in the Areas of Interest (AOI), even if they were not visible in the images, we could use them as POIs to capture the surrounding pattern information to enhance the model's performance. This allowed us to create a scenario where we could use the data from these monuments to augment our training and testing dataset. The scenarios were defined as follows:

(1) Test Scenario 1:
- Training/Testing Set: 14 Points of Interest (POIs) from Dataset 2.
- Prediction Set: 178 POIs, including 100 without dolmens and 78 with dolmens.

(2) Test Scenario 2:
- Training/Testing Set: 14 POIs from Dataset 2, plus the rest of known dolmens.
- Prediction Set: 103 POIs, consisting of 100 without dolmens and 3 with dolmens.

The data for analysis in each test scenario was divided into training and testing subsets. Outputs from all 70 POIs without monuments were consistently included across all scenarios. For images containing monuments, the dataset was adjusted to define two distinct configurations. Scenario 1 contains only the POIs returned by object detection during the test phase. In contrast, Scenario 2 incorporates not only these POIs but also the information about the other monuments in the LPG. The POIs in each scenario were distributed between the training and testing phases. The POIs not used in the training/testing were reserved for the prediction phase, where the model was evaluated with unseen data, referred to here as the "prediction scenario". The data used in the prediction scenario was derived from the object detection analysis explained in Section 4.6. Our dataset, referred to as the "prediction set," includes images of 100 POIs that did not contain monuments but were analyzed by the object detection model, resulting in 64 FPs.

6.3. Proposed Algorithms

In this section, we present our proposed approach, which combines object detection outputs with information extracted from our LPG. Our approach integrates the results from the object detection phase with a structured knowledge base, utilizing domain knowledge to improve these results. Figure 6.2 shows the flow of our proposed model.

In the first step, during pre-classification, the object detection model obtains scores for each identified monument. These scores, along with the coordinates of the analyzed images, form the basis for creating POIs, each representing a monument or a potential monument. In the second step, the POIs are used as inputs to query our LPG, which contains detailed information about the monuments and their surrounding landscape. The relevant landscape features, are defined in Table 5.20 in section 5.6. The features include

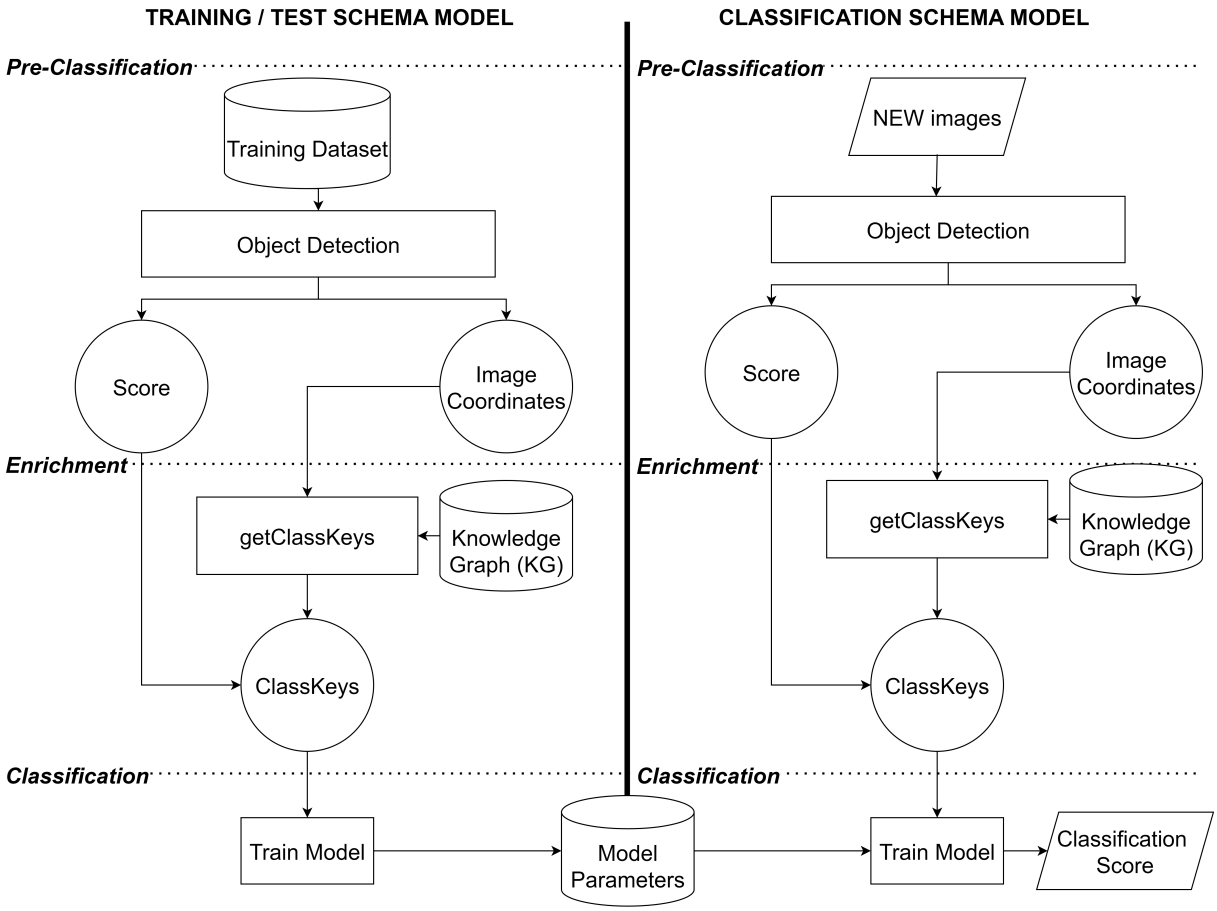


FIGURE 6.2. Pipeline architecture.

proximity to other monuments, water bodies, rock outcrops, and topographic relief areas. The information is retrieved from the KG to enrich each POI with domain knowledge about the AOI.

In the final step, the enriched data — combining the object detection outputs with the landscape features from the LPG — is used to train our KG-ML model. The training process leverages the proximity of monuments to key landscape elements, providing the model with the ability to use domain knowledge to classify monuments based on their environmental context. To train our predictive model, we tested different supervised machine-learning algorithms, implemented using Dataiku [262]. The algorithms used were K Nearest Neighbours (KNN), Logistic Regression (LR) and Least Absolute Shrinkage and Selection Operator (LASSO)-LARS, and decision tree models as Random Forest (RF) and Gradient Boosted Trees (GBT), Support Vector Machine (SVM), Single Layer Perceptron (SLP) and the Stochastic Gradient Descent (SGD) [144].

6.4. Implementation

In this section, we describe the KG-ML implementation performed in Dataiku [262]. This is a Data Science Studio platform that provides a workflow, enabling users to create scripts for data cleansing, normalization, and enrichment visually and interactively. This

approach simplifies the entire data science process, from data preparation to model development and deployment. Aside from providing an interface for manipulating data, it also provides users with the ability to integrate multiple data sources and handle large amounts of data efficiently [262]. Additionally, Dataiku offers AutoML algorithms that automate the process of model selection and tuning, further streamlining the workflow.

To retrieve the data from the KG, the POIs (defined in Section 6.2) were used as input to query and retrieve contextual knowledge from the LPG (described in Chapter 5). This task was accomplished by creating Python scripts to search and retrieve relevant information. These scripts focused on the spatial and semantic relationships between dolmens and landscape elements. The data is retrieved based on spatial relations, specifically "within" and "nearby".

The Python `Shapely` module was used to determine whether a point (each dolmen) is inside a polygon (representing land use or soil type) or near another point or line (e.g., proximity to other dolmens, water lines, relief features, or other land uses or soil types). The `distance()` method determines the shortest distance between two geometries [263]. As an example, the following code calculates the distance between POIs, which are potential detections, and nearby landscape elements, represented by points, lines, or polygons. In this case, the distance between POIs (represented as points) and rivers (represented as multilines) is being calculated.

```
point = Point(poi[0], poi[1]) # POIs latitude and longitude
multiline = shapely.wkt.loads(r[''polyg'']) # transform the WKT polygon in a shapely
    object
dist = point.distance(multiline) # compute the distance between the geometries
```

In general, the LPG is used to provide context to the scene by introducing unseen information that describes what is present in the AOI. For each POIs, the LPG provides structured and contextual data by identifying topographical relationships and returning key attributes such as soil type, land use, altitude, and water lines to the AOI.

Grid-based searches were used to optimize searches within the LPG. Rather than performing a full search across all nodes to identify the nearest one, we first determine which grid cell contains the target point based on its spatial coordinates. This approach significantly narrows down the search space, allowing us to focus only on the relevant subset of data. Once the grid is identified, we can efficiently retrieve information from the surrounding area, reducing the computational cost and time associated with searching large datasets. This method enhances the scalability of the system and supports the rapid retrieval of contextually relevant information, which is crucial for tasks such as monument detection and spatial analysis within the LPG.

When information is retrieved from the LPG, domain-specific knowledge is used to create new features for the waterlines and contour lines, which consist of various numerical values. For example, when the distance to a water line is less than 1000 meters, it is classified as 1, indicating proximity. The distance is 0 when it is 1000 meters or more away.

Similarly, for contour lines, elevation zones were categorized using interval binning into four ranges: Category A for elevations below 50 meters, Category B for elevations between 50 and 100 meters, Category C for elevations between 100 and 150 meters, and Category D for elevations above 150 meters.

This approach was not applied to semantic information, particularly when dealing with a set of specific concepts such as soil type and land use descriptions. Although it is non-ordinal data, its values represent categorical information. Therefore, the semantic data was kept in its original form, preserving its contextual significance for more advanced analysis.

6.4.1. Training and Testing

To analyse all POIs and their respective context that were retrieved from the AOI, the dataset was divided into training and testing portions — 80% for training and 20% for testing — for each of the training and testing scenarios described earlier in the Section 6.2. Like previously referred, our prediction task is a binary classification problem to distinguish images with dolmen(s) and without. To address the existing class imbalance, class rebalancing with an approximate ratio was performed by adjusting class weights to create a more balanced distribution of classes and improve model performance by mitigating bias toward any particular class. A 5-fold cross-validation strategy was used in training.

The training was carried out using the various algorithms already described. The Dataiku platform simplifies the process by offering predefined algorithm presets, some of which can be customized, while also suggesting optimal values based on the dataset. Some parameters were tailored for each algorithm to optimize performance, as indicated in the Table 6.1.

The hyperparameters for all algorithms use the grid search strategy for optimization. This approach performed better in our tests than the random search option, although the differences were minimal. Features handling was defined in the model to take into account the information about soil type, soil use, topographic relief, water line and the score to identify the class. Feature engineering is the process of constructing new features from existing ones. The goal is to derive new combinations and representations of our data that might be useful to the machine learning model [262]. For our case study, we defined explicit features to take into account in this process focusing on the interaction of soil type with relief and water line.

The training parameters described in Table 6.1 were applied to all scenario datasets using the same algorithms. This allowed for a comparison of the model's performance in handling varying data and conditions. After the model was trained, the respective prediction set was evaluated corresponding to the remaining hold-out set of data, which was not used in the training and testing phase.

Algorithm	Parameters / Options
KNN	K = 5 (number of neighbors); P = 2 (Euclidean distance);
LR	L2 regularization; Penalty parameters = 0.01, 0.1, 1, 10, 100;
LASSO	All features included; Number of tests: 100;
RF	Number of trees = 100; Maximun depht = 12;
GBT	Number of trees = 100; Maximun depht = 3; Loss = Exponential (AdaBoost);
SVM	Kernel = Sigmoid; Gamma = 1 / (number of features * variance); Stop Criteria = 0.001;
SLP	Hidden layer size = 1; Activation function = ReLU; Stop Criteria = 0.001; ADAM solver
SGD	Loss Function = Modified Huber; Stop Criteria = 0.001; L2 regularization; Alpha = 0.001

TABLE 6.1. Hyperparameters and options used in the machine learning algorithms for model training in Dataiku.

6.4.2. Feature Importance Metrics

When evaluating the feature importance in a knowledge-based machine learning approach, several metrics can be used to measure the impact of input features on the model. Shapley values identify each feature's role. This value can be understood as a weighted average of contributions to every possible subset of features [264]. This method estimates the average impact on the prediction by switching a feature's value from the one in a random sample (x) to the one in the sample to be explained (y) while also considering all possible combinations of feature switches. It computes predictions before and after the switch, repeats the process, and averages the results to determine the feature's impact (i) [262]. It considers all possible combinations of features, whether they are directly or indirectly impacting the model's output. Kumar et al. (2020) highlight computational complexity and feature selection as limitations for this conditional value feature. Calculating the exact Shapley value is difficult since it requires knowledge of multiple distributions (n combinations), which is computationally intensive and often requires approximations. Additionally, because influence is determined based on any set of features, selecting which features to include is crucial, as it impacts the explanations [264].

Shapley and ICE (Individual Conditional Expectation) can be used to compute the individual prediction explanations [262]. In contrast to Shapley, ICE focuses on a single

feature, estimating functional relationship for each observation [265]. It explains the impact of a feature by switching only its value in a sample (x) to different values, computing predictions (y), and comparing the actual prediction for (x) with the average of these predictions (y). ICE doesn't account for interactions between features as Shapley does but provides a simpler, more localized view of a feature's effect on predictions [262].

Feature importance metrics, play a crucial role in determining the model's performance by providing detailed information on each feature's importance. By using techniques like Shapley values, we can assess how each feature metric contributes to the model's precision, with each feature being assigned a weight that reflects its impact. The model's performance metrics include precision for the training phase and a cost matrix to analyse False Positives (FP), True Positives (TP), False Negatives (FN), and True Negatives (TN) (explained in Chapter 4.5.2).

6.5. Results and Discussion

This section presents the results obtained from our experiments and provides an analysis of the findings. As shown in Table 6.2, Scenario 1 yielded the worst performance overall. This may be attributed to the limited number of POIs available for testing, with only 14 data points. For instance, some algorithms, such as KNN and RF, achieved exceptionally high precision scores of 1.00, while others, like LASSO, performed poorly with a score of 0.26. In contrast, Scenario 2 demonstrated consistently strong performance across all tests, as also shown in Table 6.2. This improvement can likely be explained by adding more POIs for training the model, providing richer information about the AOI and enabling pattern identification across the area analyzed. Consequently, all subsequent analyses presented in this study are based on the results obtained from Scenario 2.

Algorithm	Test Scenario 1: Precision	Test Scenario 2: Precision
RF	1000	0,91
GBT	0,8	0,8
LR	0,8	0,81
SVM	0,44	0,81
SGD	0,66	0,83
KNN	1000	0,9
LASSO	0,26	0,91
SLP	0,44	0,91

TABLE 6.2. Performance metrics for tested algorithms in Scenario 1 and Scenario 2.

Most algorithms performed consistently, with several achieving notable results, as shown in Table 6.3. All models trained using SGD, Lasso Path, SLP, and RF achieved an average precision of around 80% or higher. These models also perform better when tested with different parameters in this case study. The SLP algorithm had better performance, with an average precision of 85%.

Algorithm	Base	Average Precision	Average F1-score
RF	Decision Tree	**0.81**	0.84
GBT	Decision Tree	0.80	0.80
LR	Linear Model	0.77	0.86
SVM	Kernel Function	0.78	0.80
SGD	Optimization for Linear Model	**0.80**	0.81
KNN	Distance Metric	0.79	0.77
LASSO	Linear Model	**0.80**	0.85
SLP	Single-Layer Neural Network	**0.85**	0.87

TABLE 6.3. Average precision and F1-score of KG-ML

In all tests, the top models achieved precision rates from 70 to 91 %. A number of factors contributed to the improvement in precision, including scoring of data derived from the previously used object detection approach and feature generation techniques that analyse nonlinear relationships among landscape elements, as described in the model training section. The techniques used were able to maintain or enhance the performance of most models, each contributing to overall precision.

Continuing our evaluation, the SLP-based model demonstrated strong performance across key metrics. This model achieved an average precision of 85%, a recall of 85% and an AUC (Area Under the Curve) of 89%. These metrics indicate a balanced performance, with high precision emphasizing the model's effectiveness in minimizing false positives, while a high recall reflects its ability to correctly identify true positives. The confusion matrix (as shown in Figure 6.3) further illustrates the model's robust performance, highlighting its capability to accurately use contextual information to classify the presence or absence of dolmens across the dataset.

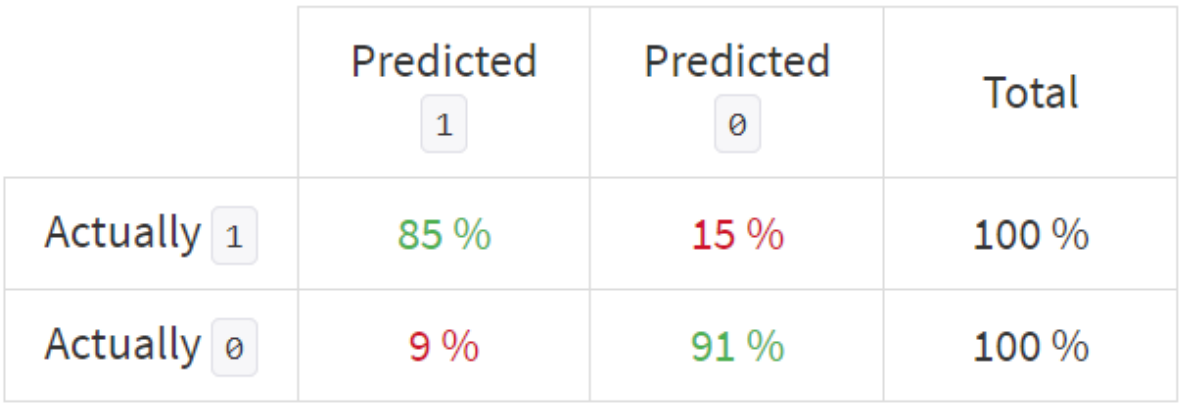


FIGURE 6.3. Confusion Matrix.

Most of the high-performing models prioritized the spatial relationships of the monument with water lines and soil type as the most influential features, followed by topography relief, land use, and finally, the detection score output from the pre-classification phase

as shown in Figure 6.4. For SLP algorithms, as demonstrated in the image, the spatial relationship with soil type, water lines and topographic relief areas was identified as the most critical factor in the model's accuracy. Land use played a secondary role, while object detection scores, although useful, had a lesser impact than the primary features. Even though the scores were not particularly solid, their inclusion in model training improved performance. In this way, scores can be used to capture patterns that, although subtle, may still be significant in certain circumstances and ultimately help guide the model. Additionally, the model can learn to distinguish between detections with high and low confidence by using the scores as a secondary element. By implementing a hybrid approach, in which scores are considered secondary information, overall performance was optimized.

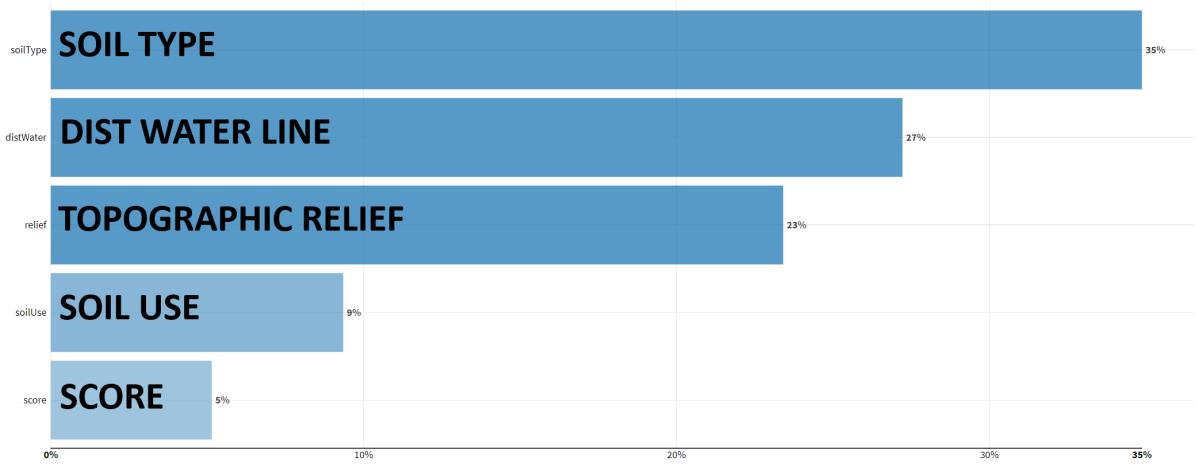


FIGURE 6.4. Feature importance — explainability metric

The feature "soilType" has the highest impact on predictions, with a feature impact of 0.92 and an information gain of 0.69. These help to explain the impact of features on the predictions of the model. In this case, for example, the presence of soil type Pg is linked to higher predictions, while soil types Vt are associated with lower predictions (Figure 6.5).

The second most impactful feature is "distWater", with a feature impact of 0.72 and a correlation of 0.71. Higher distance-to-water values are associated with higher predictions.

The feature "topographic relief" has an impact of 0.62 and an information gain of 0.68. Higher predictions are associated with Type D, while lower predictions are associated with Type C. Since there are no cases of prediction of 1 in train sections for A and B, the model is unaware of this modality (unrepresented in train). The partial dependence for relief features is shown in Figure 6.6.

The model provides insights into why each object was classified as Dolmen or non-Dolmen. The model accurately predicted 91% of the non-dolmens and 85% of the dolmens, highlighting its effectiveness in separating the two classes. Figure 6.7 shows various cases where false positives returned by object detection were correctly identified by the KG-ML model. The KG-ML model factors distance from water, topographic relief, soil type, and

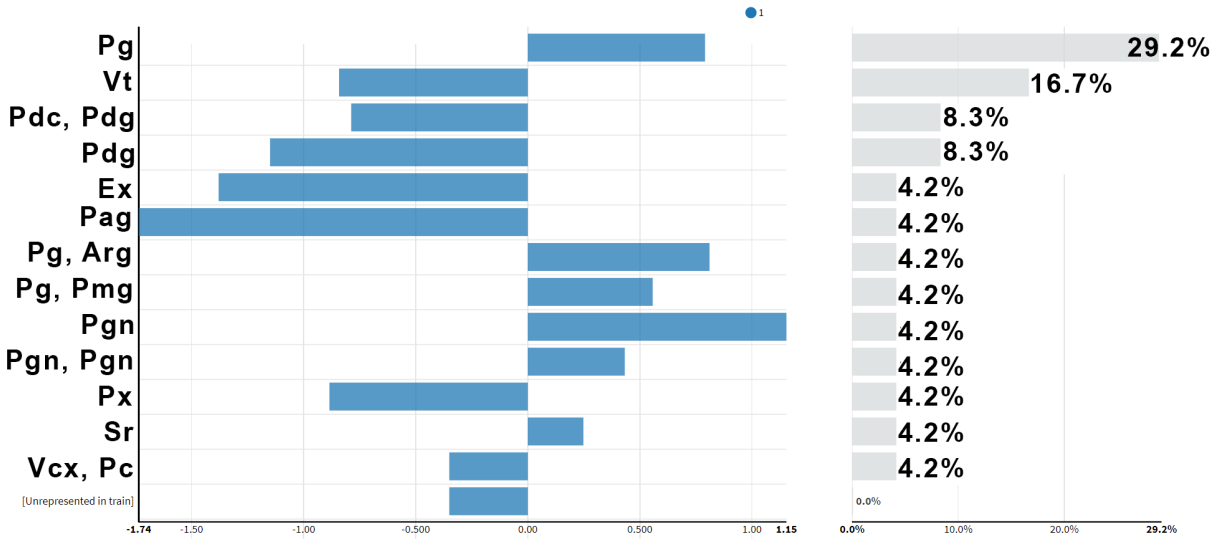


FIGURE 6.5. Partial dependence for soil type — explainability metric

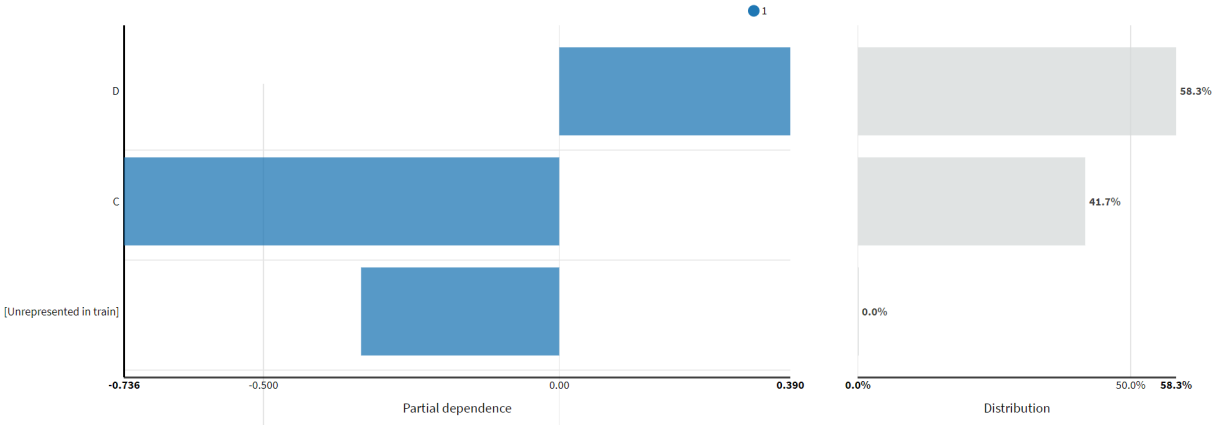


FIGURE 6.6. Partial dependence for topographic relief — explainability metric

land use. If a detection is made spatially over 1 km from a water line, in low-indication soil types (e.g., Vcx and Pc), and in a low-level relief area, and even if the detection output score is high or low, the KG-ML can correctly classify the point as not being a monument.

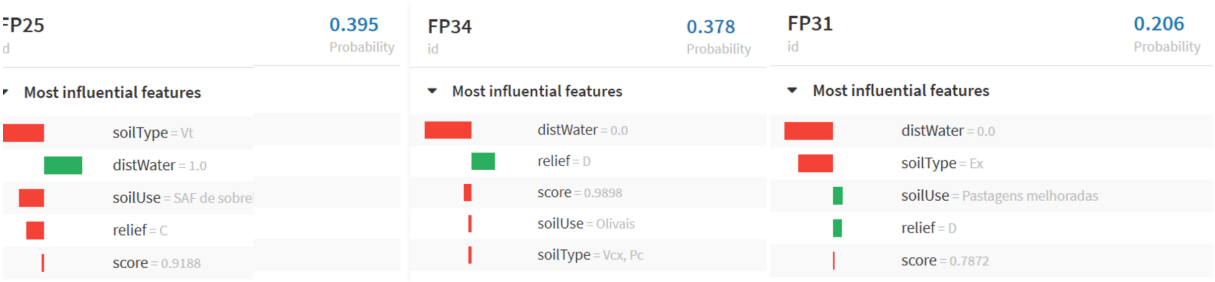


FIGURE 6.7. Individual explanations for an FP POI.

These results are particularly noteworthy given the limited data available for training and testing. Despite the small dataset, several algorithms performed remarkably well. This suggests that landscape context information effectively aids in recognizing patterns and predicting AOIs for dolmens. It underscores the potential of these models to generalize and make accurate predictions even with constrained data. This robustness across multiple algorithms highlights the value of landscape context in enhancing model performance. It reinforces the reliability of these approaches in class prediction under data-limited scenarios. Next is shown the use of the model for new POIs

Using the configuration described earlier, we tested the predictive model's performance with unviewed POIs. Using this KG-ML approach, FPs are reduced by 84%. In the appendices, Table 8.2 shows the POIs for false detections returned by the object detector that were given new scores by KG-ML. Of the 64 incorrectly classified images, only 10 remained. Additionally, the table contains the top 3 elements that influenced the model's decision and the weight assigned to each, derived from the combination of relevant relationships. Despite additional classification datasets and a wide range of POIs across the 185 km^2 area of interest, the model maintained its high performance. This indicates that the model can adapt well to new POIs in the AOI and that its pattern recognition capabilities are effective for identifying AOIs. Through the use of environmental information, the model enhances detection scores significantly, even though it doesn't directly detect monuments. In addition to improving detection scores, it provides valuable insights into relevant patterns and features associated with dolmens in diverse locations.

In analyzing the results of a test set, using POIs with known dolmens that are not visible in images, the model was capable of precisely identifying them. This highlights the model's performance not only in identifying AOIs but also in improving monument location detection precision. These results suggest that the model performs well in enhancing detection precision and providing valuable insights into classifying true monument locations.

Overall, these results highlight the model's effectiveness in enhancing detection precision and minimizing false positives while also suggesting that further optimization could help capture all true positives and reduce missed detections, even when new coordinates and datasets are introduced.

CHAPTER 7

Conclusions

In this final chapter, we present an overall view of the work reported in this dissertation. The following sections begin with a detailed discussion of the key components of the project in Section 7.1, followed by an outline of the contributions of this work in Section 7.2, and conclude with insights and possibilities for future developments (Section 7.3).

7.1. Global Considerations

The literature review shows that object detection approaches contribute to the recognition of archaeological sites from satellite images, allowing archaeologists to manage vast datasets more efficiently. These approaches facilitate faster identification of heritage sites across extensive geographical areas. However, there are still limitations. While promising, these methods often generate false positives, especially when identifying small features. Also, these approaches tend to focus on simpler forms with distinct geometric features or specific spectral behaviours, such as material reflectance in agricultural fields that contrast with the surrounding landscape. Not surprisingly, in the Areas of Interest (AOI) we worked on (Pavia, Portugal), where both the dolmens, our target object, and the surrounding terrain share the same spectral behaviour (granite for both soil and monument) and where the monuments are too small to be easily visualized (≤ 4 meters in diameter), detection proved be particularly challenging — returning many false detections.

The False Positives (FP)s returned by automated approaches to archaeological site detection are traditionally analyzed manually by specialists. In these interpretations, experts bring their domain knowledge to the table, enabling them to better understand and interpret the context. By incorporating domain knowledge into automated approaches, this process could be streamlined, with fewer data points that will require immediate attention. In fields like remote sensing, Knowledge Graph (KG)s are already seen as the future for contextualizing scenes and improving object detection. The implementation of KGs, representing domain knowledge, for both cultural heritage representation and landscape analysis is gaining popularity, however they are addressed separately at the same time that info to derive this domain knowledge is not interoperable, as the literature demonstrates.

By implementing a model that interrelates information about the landscape and the monuments, we create a Labelled Property Graph (LPG) model that can be used to provide context for what surrounds a site. Using multiple Vector Topographic Maps (VTM)s, we integrated isolated layers of spatial data and incorporated relationships based

on their interactions into a unified model. By linking it with heritage information sourced from various textual references, we create a semantic model that holds the info about what exists in a space — topological features that matter for the dolmen recognition. This approach provided a more comprehensive and interconnected view of the data — enhancing its depth and context that can be explored through sophisticated queries, facilitating a deeper understanding of spatial interactions and heritage contexts. By utilizing the well-established ontologies — CIDOC-CRM and GeoSPARQL — to create our schema model, and by incorporating all metadata, concepts, and geometry provided by each source, our implementation of the LPG is structured to be findable, accessible, interoperable and reusable.

Using the LPG as a basis to train an ML model with outputs returned from traditional object detection, we created what I called the KG-ML model in order to predict AOIs where monuments are located to be used to improve these outputs. Our tests demonstrated that integrating a semantic model enriched with landscape information significantly improved the automated object detection outputs and provided explainability for each detection.

7.2. Contribution and Implications

Our proposed approach in this thesis leveraged existing domain knowledge—traditionally used by experts to recognize dolmens—and transformed it into an interoperable format that machines could utilize, allowing us to address RQ2, RQ3 and finally the main reserach question (MRQ).

In our approach to object detection, semantic information was used differently. As opposed to merging object detection outputs with knowledge graph data in a unified embedding as most previous research has done, we standardized and linked VTMs with textual information into a LPG to map a scene. Instead of labelling every object within an image, we labelled only our target objects and retained the spatial coordinates of images with detected objects. The spatial coordinates from object detection outputs — indicating where detections occurred — are used to query the graph to uncover relevant contextual information about the AOI. Context was then provided through the LPG that details what exists in the area, regardless of whether it is visible in the image. For instance, even if a river is not visible, whether due to scale, being dry, or being covered, its presence is derived semantically from the maps. Thus, our context relies not on visual features but on semantic information linked through the graph. This approach removes the need to label all scene objects, like rivers or soil types, and accounts for aspects that are difficult to label, such as land use.

By using information — which is not visible in most images, this approach can provide information such as monuments destroyed or not visible and correctly identify monument localizations as positives even if they are not visible. It appears that by relationally considering environmental factors, the machine can identify interest zones. By providing

detailed knowledge, researchers can retrieve, compare, and analyze monument components more precisely by querying and exploring cultural heritage data. Integrating various sources of information into one unified model makes it easier to contextualize and interpret non-visible features, making it an effective tool for future research regarding spatial relationships and image identification.

There is a growing interest in utilizing KGs to represent contextual and spatial knowledge to meet the challenges of heterogeneity and interoperability. However, KGs as a tool for contextualizing scenes in satellite images is still in its infancy. As far as we know, this is the first study linking geospatial data with contextual information about archaeological monuments and landscape elements derived from textual and VTMs data sources into a KG and employ it as a reviewer to improve object detection outputs. As opposed to existing KGs, our model integrates both landscape and archaeological information and uses this information to train an Machine Learning (ML) model to identify patterns and predict AOIs where archaeological monuments can be found — thus providing context to the representation of real-world features in each location.

Using machine learning to automatically analyse the relationship of an AOI to detect patterns offers several significant benefits. It helps reduce FPs typically returned by data-driven approaches by incorporating contextual and semantic information, leading to more precise results. This reduces the workload for specialists who would need to confirm detections manually. Additionally, the ability to identify patterns across an entire region can propel future research by providing insights into broader trends and relationships within the landscape. This has been demonstrated in cases involving false negatives, where the model successfully identified patterns and provided context to correctly analyze previously missed detections.

Given the lack of research in this domain for the archaeological field, this contribution highlights the innovative use of KGs as a tool to train an interoperable model that can be used to leverage automated data-based approaches.

7.3. Limitations and Future Work

Despite the promising results achieved with our KG-ML approach, several limitations remain. One major challenge is the scarcity of cultural heritage data available for training, a common issue also encountered in traditional object detection. It is critical to note that we have trained the model with all available information about known monuments for the AOI. Therefore, if specific scenarios are not covered, it indicates that the limitations are not due to a lack of data on the machine's part. Many details about the monuments were not present in the sources, and some monuments mentioned by archaeologists in the literature may be missing or no longer exist. Future work should focus on integrating additional information into the KG — in response to new data becoming available.

Additionally, the KG offers extensive possibilities for analyzing spatial relationships between entities, for example, assessing monument conservation status and understanding how landscape changes, such as land use, may affect their preservation. It provides

a framework for answering questions about how the surrounding environment influences monuments, offering valuable insights into these interactions. It is possible to use this LPG in any field that seeks to understand the landscape in the future. Leveraging insights from the KG-ML model can be particularly valuable for fields like urban planning and environmental management, enabling them to identify areas of archaeological interest early on. By doing so, potential disruptions, such as delays or changes required when discovering archaeological monuments in modern construction zones, can be avoided, saving time and resources.

Further analyses also could deepen our understanding of how landscape elements—such as rivers, topography, soil types, land use, and elevation—interact with the target monument. Our current focus has been on how these individual factors can help in the identification of areas of interest, as dolmens are often found in areas with specific topography features. However, there is significant potential to expand this by considering more factors. For example, by integrating data on how soil reacts to precipitation with existing information on soil type, land use, local hydrology, and elevation, we can gain a better understanding of how environmental conditions, such as heavy rainfall or floods, affect the place — since certain soil types may be more vulnerable to erosion under excessive moisture, potentially threatening the monument's stability. This approach could enhance understanding of how environmental factors affect monument preservation and predict risks. However, integrating domain knowledge of landscape change would require input from experts in other fields. While my focus has been on archaeological knowledge to identify monument locations, the LPG is designed for expansion to incorporate additional data as it becomes available.

As the model is designed for expansion, future work could focus on enriching the KG with additional regional data and a broader range of archaeological monuments, using it to predict AOIs across different sites. Also, the KG-ML model is agnostic to specific object detection methods. It can use coordinates and detection scores from any previous recognition approach, making it a flexible tool for reviewing diverse detection outputs.

Finally, the model integrates data from various sources, with information gathered manually. This approach has provided valuable insights, but a future direction would be to incorporate Natural Language Processing (NLP) to automate extracting relevant text information, improving the model's scalability and efficiency. Additionally, while the model currently serves as a reviewer, it could be further leveraged to guide the object detection process directly

References

[1] DGADR. Direção-geral de agricultura e desenvolvimento rural. `https://www.dgadr.gov.pt/`, 2024. Accessed: 20 Sep. 2024.

[2] Direção-Geral do Território. Carta administrativa oficial de portugal - caop2022 (continente), 2023. Criado em 3 de fevereiro de 2023, atualizado em 21 de setembro de 2023. Creative Commons Attribution 4.0 - CC BY 4.0.

[3] Direção-Geral de Agricultura e Desenvolvimento Rural. Cartas de solos e de capacidade de uso do solo, 1999. Accessed: 30-May-2023.

[4] Agência Portuguesa do Ambiente, I.P. Massas de água superficiais rios de portugal continental: conjunto de dados geográfico sniamb, April 2018. Dados abertos.

[5] Município de Mora. Cartografia vetorial à escala 1:10 000 do concelho de mora, num total de cerca de 44 395 ha, 2022.

[6] Luigi Magnini and Cinzia Bettineschi. Theory and practice for an object-based approach in archaeological remote sensing. *Journal of Archaeological Science*, 107:10–22, 2019.

[7] Google. Google earth versions. `https://www.google.com/earth/about/versions/`, 2024. Accessed: 20 Sep. 2024.

[8] Direção-Geral do Património Cultural. Portal do arqueólogo, 2024. Accessed: 2024-09-19.

[9] John A Richards and Xiuping Jia. *Remote sensing digital image analysis*, volume 3. Springer, 1999.

[10] Sarah H Parcak. *Satellite remote sensing for archaeology*. Routledge, 2009.

[11] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[12] Filippos Gouidis, Alexandros Vassiliades, Theodore Patkos, Antonis Argyros, Nick Bassiliades, and Dimitris Plexousakis. A review on intelligent object perception methods combining knowledge-based reasoning and machine learning. *arXiv preprint arXiv:1912.11861*, 2019.

[13] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[14] Arianna Traviglia, Dave Cowley, Karsten Lambers, et al. Finding common ground: Human and computer vision in archaeological prospection. *AARGnews*, 53:11–24, 2016.

[15] Lieven Verdonck, Alessandro Launaro, Martin Millett, Frank Vermeulen, and Giovanna Bellini. Semi-automated object detection in gpr data using morphological

filtering. In *AP 2017: 12th International Conference of Archaeological Prospection*, pages 261–263. Archaeopress Publishing Ltd, 2017.

[16] Mehrnoush Soroush, Alireza Mehrtash, Emad Khazraee, and Jason A Ur. Deep learning in archaeological remote sensing: Automated qanat detection in the kurdistan region of iraq. *Remote Sensing*, 12(3):500, 2020.

[17] Peter M Atkinson, A Stein, and C Jeganathan. Spatial sampling, data models, spatial scale and ontologies: Interpreting spatial statistics and machine learning applied to satellite optical remote sensing. *Spatial Statistics*, 50:100646, 2022.

[18] Ariele Câmara. A fotointerpretação como recurso de prospeção arqueológica. chaves para a identificação e interpretação de monumentos megalíticos no alentejo: aplicação nos concelhos de mora e arraiolos. Master's thesis, Universidade de Évora, Portugal, 2017.

[19] Øivind Due Trier, Jarle Hamar Reksten, and Kristian Løseth. Automated mapping of cultural heritage in norway from airborne lidar data using faster r-cnn. *International Journal of Applied Earth Observation and Geoinformation*, 95:102241, 2021.

[20] Enrique Cerrillo-Cuenca. An approach to the automatic surveying of prehistoric barrows through lidar. *Quaternary International*, 435:135–145, 2017.

[21] Enrique Cerrillo-Cuenca and Primitiva Bueno-Ramírez. Counting with the invisible record? the role of lidar in the interpretation of megalithic landscapes in southwestern iberia (extremadura, alentejo and beira baixa). *Archaeological Prospection*, 26(3):251–264, 2019.

[22] A de Boer. Using pattern recognition to search lidar data for archeological sites. *in: Figueiredo, A. and G. Leite Velho (eds.) The world is in your eyes. CAA2005. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 33rd Conference, Tomar, March 2005.*, 2007.

[23] Alexandre Guyot, Laurence Hubert-Moy, and Thierry Lorho. Detecting neolithic burial mounds from lidar-derived elevation data using a multi-scale approach and machine learning techniques. *Remote sensing*, 10(2):225, 2018.

[24] Karsten Lambers, Wouter B Verschoof-van der Vaart, and Quentin PJ Bourgeois. Integrating remote sensing, machine learning, and citizen science in dutch archaeological prospection. *Remote Sensing*, 11(7):794, 2019.

[25] Øivind Due Trier and Lars Holger Pilø. Automatic detection of pit structures in airborne laser scanning data. *Archaeological Prospection*, 19(2):103–121, 2012.

[26] Øivind Due Trier, Maciel Zortea, and Christer Tonning. Automatic detection of mound structures in airborne laser scanning data. *Journal of Archaeological Science: Reports*, 2:69–79, 2015.

[27] Øivind Due Trier, Arnt-Børre Salberg, and Lars Holger Pilø. Semi-automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In *CAA2016: Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology*, pages 219–231. Archaeopress

Oxford, 2016.

[28] Øivind Due Trier, David C Cowley, and Anders Ueland Waldeland. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on arran, scotland. *Archaeological Prospection*, 26(2):165–175, 2019.

[29] Wouter B Verschoof-van der Vaart, Karsten Lambers, Wojtek Kowalczyk, and Quentin PJ Bourgeois. Combining deep learning and location-based ranking for large-scale archaeological prospection of lidar data from the netherlands. *ISPRS International Journal of Geo-Information*, 9(5):293, 2020.

[30] Wouter B Verschoof-van der Vaart and Juergen Landauer. Using carcassonnet to automatically detect and trace hollow roads in lidar data from the netherlands. *Journal of Cultural Heritage*, 47:143–154, 2021.

[31] Dylan S Davis, Carl P Lipo, and Matthew C Sanger. A comparison of automated object extraction methods for mound and shell-ring identification in coastal south carolina. *Journal of Archaeological Science: Reports*, 23:166–177, 2019.

[32] Rosa Lasaponara, Nicodemo Abate, and Nicola Masini. On the use of google earth engine and sentinel data to detect "lost" sections of ancient roads. the case of via appia. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[33] Dylan S Davis, Vanillah Andriankaja, Tahirisoa Lorine Carnat, Zafy Maharesy Chrisostome, Christophe Colombe, Felicia Fenomanana, Laurence Hubertine, Ricky Justome, François Lahiniriko, Harson Léonce, et al. Satellite-based remote sensing rapidly reveals extensive record of holocene coastal settlement on madagascar. *Journal of Archaeological Science*, 115:105097, 2020.

[34] Dylan S Davis and Kristina Douglass. Remote sensing reveals lasting legacies of land-use by small-scale foraging communities in the southwestern indian ocean. *Frontiers in Ecology and Evolution*, 9, 2021.

[35] Nicodemo Abate, Abdelaziz Elfadaly, Nicola Masini, and Rosa Lasaponara. Multitemporal 2016-2018 sentinel-2 data enhancement for landscape archaeology: The case study of the foggia province, southern italy. *Remote Sensing 2020*, 12:1309, 4 2020.

[36] Elijah FW Bowen, Brett B Tofel, Sarah Parcak, and Richard Granger. Algorithmic identification of looted archaeological sites from space. *Frontiers in ICT*, 4:247381, 2017.

[37] David Galvão Chambel Caçador. Automatic recognition of megalithic objects in areas of interest in satellite imagery. Master's thesis, ISCTE, 2020.

[38] Véronique De Laet, Etienne Paulissen, and Marc Waelkens. Methods for the extraction of archaeological features from very high-resolution ikonos-2 remote sensing imagery, hisar (southwest turkey). *Journal of Archaeological Science*, 34(5):830–841, 2007.

[39] Rosa Lasaponara, Giovanni Leucci, Nicola Masini, and Raffaele Persico. Investigating archaeological looting using satellite images and georadar: the experience in lambayeque in north peru. *Journal of Archaeological Science*, 42:216–230, 2014.

[40] Rosa Lasaponara, Giovanni Leucci, Nicola Masini, Raffaele Persico, and Giuseppe Scardozzi. Towards an operative use of remote sensing for exploring the past using satellite data: The case study of hierapolis (turkey). *Remote sensing of Environment*, 174:148–164, 2016.

[41] Carla Klehm, Adam Barnes, Forrest Follett, Katie Simon, Christopher Kiahtipes, and Sarah Mothulatshipi. Toward archaeological predictive modeling in the bosutswe region of botswana: Utilizing multispectral satellite imagery to conceptualize ancient landscapes. *Journal of Anthropological Archaeology*, 54:68–83, 2019.

[42] Hector A Orengo, Francesc C Conesa, Arnau Garcia-Molsosa, Agustín Lobo, Adam S Green, Marco Madella, and Cameron A Petrie. Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *Proceedings of the National Academy of Sciences*, 117(31):18240–18250, 2020.

[43] Olaotse Lokwalo Thabeng, Stefania Merlo, and Elhadi Adam. High-resolution remote sensing and advanced classification techniques for the prospection of archaeological sites' markers: The case of dung deposits in the shashi-limpopo confluence area (southern africa). *Journal of Archaeological Science*, 102:48–60, 2019.

[44] Olaotse L Thabeng, Stefania Merlo, and Elhadi Elhadi. From the bottom up: Assessing the spectral ability of common multispectral sensors to detect surface archaeological deposits using field spectrometry and advanced classifiers in the shashi-limpopo confluence area. In *Spatial Approaches in African Archaeology*, pages 25–49. Springer, 2020.

[45] Øivind Due Trier, Siri Øyen Larsen, and Rune Solberg. Automatic detection of circular structures in high-resolution satellite images of agricultural land. *Archaeological Prospection*, 16(1):1–15, 2009.

[46] Igor Zingman, Dietmar Saupe, Otavio AB Penatti, and Karsten Lambers. Detection of fragmented rectangular enclosures in very high resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4580–4593, 2016.

[47] Lei Luo, Xinyuan Wang, Huadong Guo, Chuansheng Liu, Jie Liu, Li Li, Xiaocui Du, and Guoquan Qian. Automated extraction of the archaeological tops of qanat shafts from vhr imagery in google earth. *Remote Sensing*, 6(12):11956–11976, 2014.

[48] Lei Luo, Xinyuan Wang, Huadong Guo, Rosa Lasaponara, Pilong Shi, Nabil Bachagha, Li Li, Ya Yao, Nicola Masini, Fulong Chen, et al. Google earth as a powerful tool for archaeological and cultural heritage applications: A review. *Remote Sensing*, 10(10):1558, 2018.

[49] Gino Caspari and Pablo Crespo. Convolutional neural networks for archaeological site detection–finding "princely" tombs. *Journal of Archaeological Science*, 110,

2019.

[50] Gonzalo J Linares Matás and Jonathan S Lim. Monumental funerary landscapes of dhar tagant (south-eastern mauritania): Towards ethical satellite remote sensing in the west african sahel. *Archaeological Prospection*, 28(3):357–378, 2021.

[51] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*, 2016.

[52] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020.

[53] Lieven Verdonck, Philippe De Smedt, and Jeroen Verhegge. Making sense of anomalies: Practices and challenges in the archaeological interpretation of geophysical data. In *Innovation in Near-Surface Geophysics*, pages 151–194. Elsevier, 2019.

[54] Wenwen Li and Chia-Yu Hsu. Geoai for large-scale image analysis and machine vision: recent progress of artificial intelligence in geography. *ISPRS International Journal of Geo-Information*, 11(7):385, 2022.

[55] Laure Nuninger, Rachel Opitz, Philip Verhagen, Thérèse Libourel, Clément Laplaige, Samuel Leturcq, Nathanael Le Voguer, Catherine Fruchart, Ziga Kokalj, and Xavier Rodier. Developing fair ontological pathways: linking evidence of movement in lidar to models of human behaviour. *Journal of Computer Applications in Archaeology*, 3(1):63–75, 2020.

[56] Leonor RochA. Povoamento megalítico de pavia. contributo para o conhecimento da pré-história regional, 1998.

[57] Gerald Hiebel, Martin Doerr, and Øyvind Eide. Crmgeo: A spatiotemporal extension of cidoc-crm. *International Journal on Digital Libraries*, 18:271–279, 2017.

[58] Nicholas J Car and Timo Homburg. Geosparql 1.1: Motivations, details and applications of the decadal update to the most important geospatial lod standard. *ISPRS International Journal of Geo-Information*, 11(2):117, 2022.

[59] Chryssoula Bekiari, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios, editors. *Volume A: Definition of the CIDOC Conceptual Reference Model*, volume 7.2. CIDOC, 2021.

[60] Giuseppe Futia and Antonio Vetrò. On the integration of knowledge graphs into deep learning models for a more comprehensible ai—three challenges for future research. *Information*, 11(2):122, 2020.

[61] Paulo Alvim. De nascente para poente: reflexões sobre a sintaxe da arquitectura megalítica no alentejo. In *Arqueologia de Transição: O Mundo Funerário. Actas do II Congresso Internacional Sobre Arqueologia de Transição*, pages 211–216, 2013.

[62] Auditorium du Musée de Bretagne. Megalithic architectures: intentions and construction, styles and techniques. In *IV Meeting of the European Megalithic Studies Group*, 2012.

[63] Primitiva Bueno Ramírez, Rosa Barroso Bermejo, and Rodrigo de Balbín Behrmann. Between east and west: megaliths in the centre of the iberian peninsula. *The megalithic architectures of Europe*, page 157, 2015.

[64] F Russel Cortez. Contributo para o estudo do neolítico de portugal. *Trabalhos de Antropologia e Etnologia*, 13(3-4), 2020.

[65] Leonor Rocha and Manuel Calado. *Megalitismo de Mora. Nas Fronteiras do Alentejo Central*. Apenas Livros, Lda, 2006.

[66] Leonor Rocha. Where were the dead buried in recent prehistory? the problem of architectures versus chronologies in central alentejo (portugal). In Florian Cousseau and Luc Laporte, editors, *Pre and Protohistoric. Stone Architectures. Comparisons of the social and techinical contexts associated to theis building*, pages 86–94. Archaeopress Publishing Ltd, Oxford, 2020.

[67] Leonor Rocha. Megalitismo, destruições e classificações: perspetivas e problemáticas sobre o estado da questão no alentejo. *Vol. 1: IV Jornadas de Arqueologia do Norte Alentejano*, 2022.

[68] E Thurlow Leeds. Viii.—the dolmens and megalithic tombs of spain and portugal. *Archaeologia*, 70:201–232, 1920.

[69] Leonor Oliveira, Catarinaand Rocha and Cândido da Silva. O megalitismo funerário no alentejo central–arquitectura e orientações: o estado da questão em montemor-o-novo. *CHAIA*, 2007.

[70] Georg Leisner. *Antas dos arredores de Évora*. A cidade de Évora, 1948.

[71] V Correia. El neolítico de pavía, comisión de investigaciones paleontológicas y prehistóricas, 1921.

[72] Direção-Geral do Património Cultural. Official gazette no. 39/2022. abertura do procedimento de classificação do megalitismo alentejano. Diário da República, 2.ª série.

[73] Manuel Calado, Leonor Rocha, and Pedro Alvim. *O Tempo das Pedras. Carta Arqueológica de Mora*. Câmara Municipal, Mora, 2012.

[74] Ariele Câmara and Teresa Batista. Photo interpretation and geographic information systems for dolmen identification in portugal: The case study of mora and arraiolos. In *2017 12th Iberian conference on information systems and technologies (CISTI)*, pages 1–6. IEEE, 2017.

[75] Leonor RochA. Aspectos do megalitismo da área de pavia, mora (portugal). *Revista Portuguesa de Arqueologia*, 1999.

[76] Jorge de Oliveira. O" eterno descanso" no neolítico do alentejo norte. In *Actas do II Congresso Internacional Sobre Arqueologia de Transição*, 2015.

[77] G. Zbyszewski, O. Da Veiga Ferreira, and A. Barros e Carvalhosa. *Carta Geológica de Portugal na escala 1/50.000, Folha 35-B (Mora) e respectiva notícia explicativa*. Direcção Geral de Geologia e Minas Serviços Geológicos de Portugal, Lisboa, 1979.

[78] A. A. Araújo, J. P. Almeida, J. Borrego, J. Pedro, and T. Oliveira. As regiões central e sul da zona de ossa-morena. In Rui Dias, Alexandre Araújo, Pedro Terrinha, and José Carlos Kullberg, editors, *Geologia de Portugal, Vol. I: Geologia Pré-mesozóica de Portugal*, pages 509–549. Livraria Escolar Editora, Lisboa, 2013.

[79] António Pereira Jorge Ferreira. *Dados geoquímicos de base de sedimentos fluviais de amostragem de baixa densidade de Portugal Continental: Estudo de factores de variação regional*. PhD thesis, Universidade de Aveiro (Portugal), 2000.

[80] Catarina Ramos. Condições geomorfológicas e climáticas das cheias de ribeira de tera e do rio maior (bacia hidrográfica do tejo), 2019.

[81] Carlos Miranda Rodrigues, Rita Cabral Guimarães, and Madalena Moreira. Apontamentos para as aulas de hidrologia. *Departamento de Engenharia Rural. Universidade de Évora.*, 2011.

[82] Carvalho Cardoso. *Os solos de Portugal: sua classificação, caracterização e génese*. Secretaria de Estado da Agricultura - Direcao Geral dos Servicos Agricolas, 1965.

[83] Carlos Alexandre, Alberto Azevedo Gomes, Maria João Rosado, Sandra Mesquita, Irene Cadima, and Francisco Rego. Caracterização preliminar dos solos ocupados com montados de sobro e azinho nos concelhos de avis, mora e arraiolos. *MED*, 2002.

[84] Carlos Marques and Mário Carvalho. A agricultura e os sistemas de produção da região alentejo de portugal: Evolução, situação atual e pespectivas. *Revista de Economia e Agronegócio - REA*, 2017.

[85] Serviço de Reconhecimento e de Ordenamento Agrário. *Carta dos Solos de Portugal: Classificação e Caracterização Morfológica dos Solos*, volume 1. Secretaria do Estado da Agricultura, 6ª ed. edition, 1970.

[86] Leonor Rocha. *O Neolítico no Alentejo: novas reflexões*. Associação dos Arqueólogos Portugueses, 2016.

[87] Direção-Geral do Território. Carta de uso e ocupação do solo - cos. `https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/metadata/b498e89c-1093-4793-ad22-63516062891b`, 2018. Versão 2018.

[88] Leonor RochA. *As origens do megalitismo funerário no Alentejo Central: a contribuição de Manuel Heleno*. PhD thesis, FLL Universidade de Lisboa, 2005.

[89] Edward C Harris. *Principles of archaeological stratigraphy*. Academic Press, London, 1989.

[90] Laure Nuninger, Philip Verhagen, Thérèse Libourel, Rachel Opitz, Xavier Rodier, Clément Laplaige, Catherine Fruchart, Samuel Leturcq, and Nathanael Levoguer. Linking theories, past practices, and archaeological remains of movement through ontological reasoning. *Information*, 11(6):338, 2020.

[91] Kenneth Brophy and David Cowley. *From the air: understanding aerial archaeology*. Tempus Stroud, 2005.

[92] Gerald Hiebel, Gert Goldenberg, Caroline Grutsch, Klaus Hanke, and Markus Staudt. Fair data for prehistoric mining archaeology. *International Journal on Digital Libraries*, 22(3):267–277, 2021.

[93] Dylan S Davis. Defining what we study: The contribution of machine automation in archaeological research. *Digital Applications in Archaeology and Cultural Heritage*, 18:e00152, 2020.

[94] Martin Doerr, Athina Kritsotaki, and Katerina Boutsika. Factual argumentation—a core model for assertions making. *Journal on Computing and Cultural Heritage (JOCCH)*, 3(3):1–34, 2011.

[95] Richard Gartner and Richard Gartner. *Metadata*. Springer, 2016.

[96] Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. Pattern-based design applied to cultural heritage knowledge graphs. *Semantic Web*, 12(2):313–357, 2021.

[97] Dylan S Davis. Geographic disparity in machine intelligence approaches for archaeological remote sensing research. *Remote Sensing*, 12(6):921, 2020.

[98] Peter McKeague, Anthony Corns, Åsa Larsson, Anne Moreau, Axel Posluschny, Koen Van Daele, and Tim Evans. One archaeology: A manifesto for the systematic and effective use of mapped data from archaeological fieldwork and research. *Information*, 11(4):222, 2020.

[99] Fredrik Gunnarsson. *Archaeological challenges, digital possibilities: Digital knowledge development and communication in contract archaeology*. PhD thesis, Linnaeus University Press, 2018.

[100] Marcia Lei Zeng. Semantic enrichment for enhancing lam data and supporting digital humanities. review article. *El profesional de la información*, 28(1), 2019.

[101] Anwen Cooper and Chris Green. Embracing the complexities of 'big data'in archaeology: the case of the english landscape and identities project. *Journal of Archaeological Method and Theory*, 23:271–304, 2016.

[102] Carlos Henrique Marcondes. Integrated classification schemas to interlink cultural heritage collections over the web using lod technologies. *International Journal of Metadata, Semantics and Ontologies*, 15(3):170–177, 2021.

[103] J. Bugalhão, A. Lucena, F. Bragança, F. Neto, M. J. Sousa, S. Gomes, and T. Fraga. Endovélico. sistema de gestão e informação arqueológica. *Revista Portuguesa de Arqueologia*, 5(1):277–283, 2002.

[104] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T Evelo, et al. Fair principles: interpretations and implementation considerations, 2020.

[105] Sara Encarnação. Interpretação da dimensão geográfica do objeto em detecção remota.

[106] Paulo Roberto Meneses and T de Almeida. *Introdução ao processamento de imagens de sensoriamento remoto*. Universidade de Brasília, Brasília, 2012.

[107] E. P. M. Sampaio. Noções básicas de detecção remota, fotogrametria e fotointerpretação em pedologia, 2007.

[108] Karsten Lambers. Airborne and spaceborne remote sensing and digital image analysis in archaeology. *Digital Geoarchaeology: New Techniques for Interdisciplinary Human-Environmental Research*, pages 109–122, 2018.

[109] Geert J Verhoeven. Are we there yet? a review and assessment of archaeological passive airborne optical imaging approaches in the light of landscape archaeology. *Geosciences*, 7(3):86, 2017.

[110] Divino Figueiredo. Conceitos básicos de sensoriamento remoto. *São Paulo*, 2005.

[111] Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

[112] Anabela Carvalho. A fotografia aérea na arqueologia. *Ângulo, Tomar*, 1(1):57–67, 2002.

[113] PwC. Copernicus services in support to cultural heritage. Technical report, Publications Office of the European Union, Luxembourg, 2018. Published in 2019.

[114] NASA. Nasa worldview. `https://worldview.earthdata.nasa.gov/`, 2024. Accessed: 20 Sep. 2024.

[115] United States Geological Survey. Earth explorer. `https://earthexplorer.usgs.gov/`, 2024. Accessed: 20 Sep. 2024.

[116] Basel Shbita, Craig A Knoblock, Weiwei Duan, Yao-Yi Chiang, Johannes H Uhl, and Stefan Leyk. Building spatio-temporal knowledge graphs from vectorized topographic historical maps. *Semantic Web*, 14(3):527–549, 2023.

[117] Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. A review of location encoding for geoai: methods and applications. *International Journal of Geographical Information Science*, 36(4):639–673, 2022.

[118] SNIAmb. Sistema nacional de informação de ambiente. `https://sniamb.apambiente.pt/`, 2024. Accessed: 20 Sep. 2024.

[119] SNIRH Sistema Nacional de Informação de Recursos Hídricos. Sistema nacional de informação de recursos hídricos. `https://snirh.apambiente.pt/`, 2024. Accessed: 20 Sep. 2024.

[120] DGT. Direção-geral do território. `https://www.dgterritorio.gov.pt/`, 2024. Accessed: 20 Sep. 2024.

[121] SNIG. Sistema nacional de informação geográfica. `https://snig.dgterritorio.gov.pt/`, 2024. Accessed: 20 Sep. 2024.

[122] SMOS. Sistema de monitorização da ocupação de solos. `https://smos.dgterritorio.gov.pt/visualizadores`, 2024. Accessed: 20 Sep. 2024.

[123] K. Tempfli, G. C. Huurneman, W. H. Bakker, L. L. F. Janssen, W. F. Feringa, A. S. M. Gieske, K. A. Grabmaier, C. A. Hecker, J. A. Horn, N. Kerle, F. D.

van der Meer, G. N. Parodi, C. Pohl, C. V. Reeves, F. J. A. van Ruitenbeek, E. M. Schetselaar, M. J. C. Weir, E. Westinga, and T. Woldai. *Principles of Remote Sensing: An Introductory Textbook*, volume 2 of *ITC Educational Textbook Series*. International Institute for Geo-Information Science and Earth Observation, 2009.

[124] Peter Drewett. *Field archaeology: an introduction*. Routledge, 2011.

[125] Damien Arvor, Laurent Durieux, Samuel Andrés, and Marie-Angélique Laporte. Advances in geographic object-based image analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing*, 82:125–137, 2013.

[126] Dave Cowley, Łukasz Banaszek, George Geddes, Angela Gannon, Mike Middleton, and Kirsty Millican. Making light work of large area survey? developing approaches to rapid archaeological mapping and the creation of systematic national-scaled heritage data. *Journal of Computer Applications in Archaeology*, 3(1):109–121, 2020.

[127] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. Object detection meets knowledge graphs. International Joint Conferences on Artificial Intelligence, 2017.

[128] Damien Arvor, Mariana Belgiu, Zoe Falomir, Isabelle Mougenot, and Laurent Durieux. Ontologies to interpret remote sensing images: why do we need them? *GIScience & remote sensing*, 56(6):911–939, 2019.

[129] Yong Ge, Xining Zhang, Peter M Atkinson, Alfred Stein, and Lianfa Li. Geoscience-aware deep learning: A new paradigm for remote sensing. *Science of Remote Sensing*, 5:100047, 2022.

[130] R Girshick. Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[131] Bjoern H Menze, Jason A Ur, and Andrew G Sherratt. Detection of ancient settlement mounds. *Photogrammetric Engineering & Remote Sensing*, 72(3):321–327, 2006.

[132] Melda Küçükdemirci and Apostolos Sarris. Deep learning based automated analysis of archaeo-geophysical images. *Archaeological Prospection*, 27(2):107–118, 2020.

[133] Zhao Gun and Jianyu Chen. Novel knowledge graph-and knowledge reasoning-based classification prototype for obia using high resolution remote sensing imagery. *Remote Sensing*, 15(2):321, 2023.

[134] Athos Agapiou. Enhancement of archaeological proxies at non-homogenous environments in remotely sensed imagery. *Sustainability*, 11:3339, 6 2019.

[135] Rutherford V Platt and Lauren Rapoza. An evaluation of an object-oriented paradigm for land use/land cover classification. *The Professional Geographer*, 60(1):87–100, 2008.

[136] Gang Chen, Qihao Weng, Geoffrey J Hay, and Yinan He. Geographic object-based image analysis (geobia): Emerging trends and future opportunities. *GIScience & Remote Sensing*, 55(2):159–182, 2018.

[137] Rosa Lasaponara and Nicola Masini. Satellite remote sensing in archaeology: Past, present and future perspectives. *Journal of Archaeological Science*, 38(9):1995–2002, 2011.

[138] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117:11–28, 2016.

[139] Stefan Lang, Geoffrey J Hay, Andrea Baraldi, Dirk Tiede, and Thomas Blaschke. Geobia achievements and spatial opportunities in the era of big earth observation data. *ISPRS International Journal of Geo-Information*, 8(11):474, 2019.

[140] Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108, 2020.

[141] Suryakanthi Tangirala. Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2):612–619, 2020.

[142] Christopher Sevara, Michael Pregesbauer, Michael Doneus, Geert Verhoeven, and Immo Trinks. Pixel versus object—a comparison of strategies for the semi-automated mapping of archaeological features using airborne laser scanning data. *Journal of Archaeological Science: Reports*, 5:485–498, 2016.

[143] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[144] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009.

[145] Petru Soviany and Radu Tudor Ionescu. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 209–214. IEEE, 2018.

[146] Arsalan Tahir, Hafiz Suliman Munawar, Junaid Akram, Muhammad Adil, Shehryar Ali, Abbas Z Kouzani, and MA Parvez Mahmud. Automatic target detection from satellite imagery using machine learning. *Sensors*, 22(3):1147, 2022.

[147] Shaoqing Ren. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[148] Victoria Eyharabide, Imad Eddine Ibrahim Bekkouch, and Nicolae Dragoș Constantin. Knowledge graph embedding-based domain adaptation for musical instrument recognition. *Computers*, 10(8):94, 2021.

[149] Amirah Hanani Jamil, Fitri Yakub, Azizul Azizan, Shairatul Akma Roslan, Sheikh Ahmad Zaki, and Syafiq Asyraff Ahmad. A review on deep learning application for detection of archaeological structures. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 26(1):7–14, 2022.

[150] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.

Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[151] Khitem Amiri and Mohamed Farah. Graph of concepts for semantic annotation of remotely sensed images based on direct neighbors in rag. *Canadian Journal of Remote Sensing*, 44:551–574, 11 2018.

[152] Yansheng Li, Song Ouyang, and Yongjun Zhang. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. *Knowledge-based systems*, 243:108469, 2022.

[153] Till F Sonnemann, Douglas C Comer, Jesse L Patsolic, William P Megarry, Eduardo Herrera Malatesta, and Corinne L Hofman. Semi-automatic detection of indigenous settlement features on hispaniola through remote sensing data. *Geosciences*, 7(4):127, 2017.

[154] Alishiba Dsouza, Nicolas Tempelmeier, Ran Yu, Simon Gottschalk, and Elena Demidova. Worldkg: A world-scale geographic knowledge graph. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4475–4484, 2021.

[155] Gilles-Antoine Nys, Muriel Van Ruymbeke, and Roland Billen. Spatio-temporal reasoning in cidoc crm: an hybrid ontology with geosparql and owl-time. In *CEUR Workshop Proceedings*, volume 2230. RWTH Aachen University, Aachen, Germany, 2018.

[156] Luciano Serafini, Ivan Donadello, and Artur d'Avila Garcez. Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proceedings of the Symposium on Applied Computing*, pages 125–130, 2017.

[157] Ivan Donadello and Luciano Serafini. Compensating supervision incompleteness with prior knowledge in semantic image interpretation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[158] Gilles-Antoine Nys, Jean-Paul Kasprzyk, Pierre Hallot, and Roland Billen. Towards an ontology for the structuring of remote sensing operations shared by different processing chains. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(4), 2018.

[159] Jiaxin Du, Shaohua Wang, Xinyue Ye, Diana S Sinton, and Karen Kemp. Gis-kg: building a large-scale hierarchical knowledge graph for geographic information science. *International Journal of Geographical Information Science*, 36(5):873–897, 2022.

[160] Marjan Alirezaie, Martin Längkvist, Michael Sioutis, and Amy Loutfi. Semantic referee: A neural-symbolic framework for enhancing geospatial semantic segmentation. *Semantic Web*, 10:863–880, 1 2019.

[161] Yansheng Li, Deyu Kong, Yongjun Zhang, Yihua Tan, and Ling Chen. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. *ISPRS Journal of*

Photogrammetry and Remote Sensing, 179:145–158, 2021.

[162] Adekanmi Adeyinka Adegun, Jean Vincent Fonou-Dombeu, Serestina Viriri, and John Odindi. Ontology-based deep learning model for object detection and image classification in smart city concepts. *Smart Cities*, 7:2182–2207, 8 2024.

[163] Samuel Andrés, Damien Arvor, Isabelle Mougenot, Thérèse Libourel, and Laurent Durieux. Ontology-based classification of remote sensing images using spectral rules. *Computers Geosciences*, 102:158–166, 5 2017.

[164] Caixia Rong and Wenxue Fu. A comprehensive review of land use and land cover change based on knowledge graph and bibliometric analyses. *Land*, 12(8):1573, 2023.

[165] Sebastian Monka, Lavdim Halilaj, and Achim Rettinger. A survey on visual transfer learning using knowledge graphs. *Semantic Web*, 13(3):477–510, 2022.

[166] Dylan Davis. Theoretical repositioning of automated remote sensing archaeology: Shifting from features to ephemeral landscapes. *Journal of Computer Applications in Archaeology*, 4(1):94–109, 2021.

[167] Rachel Opitz and Jason Herrmann. Recent trends and long-standing problems in archaeological remote sensing. *Journal of Computer Applications in Archaeology*, 1(1):19–41, 2018.

[168] Thomas Schleider, Raphael Troncy, Thibault Ehrhart, Mareike Dorozynski, Franz Rottensteiner, Jorge Sebastián Lozano, and Georgia Lo Cicero. Searching silk fabrics by images leveraging on knowledge graph and domain expert rules. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 41–49, 2021.

[169] Babak Ranjgar, Abolghasem Sadeghi-Niaraki, Maryam Shakeri, and Soo-Mi Choi. An ontological data model for points of interest (poi) in a cultural heritage site. *Heritage Science*, 10(1):13, 2022.

[170] Gengchen Mai, Yingjie Hu, Song Gao, Ling Cai, Bruno Martins, Johannes Scholz, Jing Gao, and Krzysztof Janowicz. Symbolic and subsymbolic geoai: Geospatial knowledge graphs and spatially explicit machine learning. *Trans. GIS*, 26(8):3118–3124, 2022.

[171] Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2(2):131–143, 2018.

[172] GeoNames. Geonames geographical database, 2023. Accessed: 2023-09-22.

[173] Wenwen Li, Sizhe Wang, Xiao Chen, Yuanyuan Tian, Zhining Gu, Anna Lopez-Carr, Andrew Schroeder, Kitty Currier, Mark Schildhauer, and Rui Zhu. Geographvis: a knowledge graph and geovisualization empowered cyberinfrastructure to support disaster response and humanitarian aid. *ISPRS International Journal of Geo-Information*, 12(3):112, 2023.

[174] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby Fisher, Ling Cai, Gengchen Mai, et al. Know,

know where, knowwheregraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. *AI Magazine*, 43(1):30–39, 2022.

[175] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[176] Nikolaos Karalis, Georgios Mandilaras, and Manolis Koubarakis. Extending the yago2 knowledge graph with precise geospatial knowledge. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*, pages 181–197. Springer, 2019.

[177] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing wikidata to the linked data web. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*, pages 50–65. Springer, 2014.

[178] Simon Gottschalk and Elena Demidova. Eventkg–the hub of event knowledge on the web–and biographical timeline generation. *Semantic Web*, 10(6):1039–1070, 2019.

[179] Shu Wang, Xueying Zhang, Peng Ye, Mi Du, Yanxu Lu, and Haonan Xue. Geographic knowledge graph (geokg): A formalized geographic knowledge representation. *ISPRS International Journal of Geo-Information*, 8(4):184, 2019.

[180] Gengchen Mai, Weiming Huang, Ling Cai, Rui Zhu, and Ni Lao. Narrative cartography with knowledge graphs. *Journal of Geovisualization and Spatial Analysis*, 6(1):4, 2022.

[181] Sören Auer, Jens Lehmann, and Sebastian Hellmann. Linkedgeodata: Adding a spatial dimension to the web of data. In *The Semantic Web-ISWC 2009: 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings 8*, pages 731–746. Springer, 2009.

[182] Xueying Zhang, Yi Huang, Chunju Zhang, and Peng Ye. Geoscience knowledge graph (geokg): Development, construction and challenges. *Transactions in GIS*, 26(6):2480–2494, 2022.

[183] Luis M Vilches-Blázquez and Jhonny Saavedra. A graph-based representation of knowledge for managing land administration data from distributed agencies–a case study of colombia. *Geo-spatial Information Science*, 25(2):259–277, 2022.

[184] Diego Calvanese, Avigdor Gal, Davide Lanti, Marco Montali, Alessandro Mosca, and Roee Shraga. Mapping patterns for virtual knowledge graphs. *arXiv preprint arXiv:2012.01917*, 2020.

[185] Jesse Casana. Global-scale archaeological prospection using corona satellite imagery: Automated, crowd-sourced, and expert-led approaches. *Journal of Field Archaeology*, 45(1):S89–S100, 2020.

[186] Nikos Partarakis, Voula Doulgeraki, Effie Karuzaki, George Galanakis, Xenophon Zabulis, Carlo Meghini, Valentina Bartalesi, and Daniele Metilli. A web-based platform for traditional craft documentation. *Multimodal Technologies and Interaction*, 6(5):37, 2022.

[187] Inês Koch, Nuno Freitas, Cristina Ribeiro, Carla Teixeira Lopes, and João Rocha da Silva. Knowledge graph implementation of archival descriptions through cidoc-crm. In *International conference on theory and practice of digital libraries*, pages 99–106. Springer, 2019.

[188] Lázaro Costa, Nuno Freitas, and João Rocha da Silva. An evaluation of graph databases and object-graph mappers in cidoc crm-compliant digital archives. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3):1–18, 2022.

[189] Philip K Carlisle, Ioannis Avramides, Alison Dalgity, and David Myers. The arches heritage inventory and management system: a standards-based approach to the management of cultural heritage information. In *CIDOC (International Committee for Documentation of the International Council of Museums) conference: access and understanding–networking in the digital era, Dresden, Germany*, pages 6–11, 2014.

[190] Martin Doerr, Maria Theodoridou, Edeltraud Aspöck, Anja Masur, and ÖAW-Österreichische Akademie der Wissenschaften. Mapping archaeological databases to cidoc-crm. In *CAA2015 Keep the revolution going–Proceedings of 43rd Annual Conference of Computer Applications and Quantitative Methods in Archaeology, Archaeopress Archaeology, Oxford*, pages 443–451, 2016.

[191] Ceri Binding, Keith May, Renato Souza, Douglas Tudhope, and Andreas Vlachidis. Semantic technologies for archaeology resources: results from the star project. In *Th Annual Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2010), Granada: Archaeopress*, pages 555–561. CAA, 2010.

[192] Olivier Marlet, Thomas Francart, Béatrice Markhoff, and Xavier Rodier. Openarchaeo for usable semantic interoperability. In *ODOCH 2019@ CAiSE 2019*, 2019.

[193] Henrik Jarl Hansen and Kate Fernie. Carare: connecting archaeology and architecture in europeana. In *Digital Heritage: Third International Conference, EuroMed 2010, Lemessos, Cyprus, November 8-13, 2010. Proceedings 3*, pages 450–462. Springer, 2010.

[194] Inês Koch, Cristina Ribeiro, and Carla Teixeira Lopes. Archonto, a cidoc-crm-based linked data model for the portuguese archives. In *International Conference on Theory and Practice of Digital Libraries*, pages 133–146. Springer, 2020.

[195] Achille Felicetti Martin Doerr Paola Ronzino, Franco Niccolucci et al. Definition of the crmba: An extension of cidoc crm to support buildings archaeology documentation. Technical report, CIDOC-CRM, December 2016. Proposal for approval by CIDOC CRM-SIG, Version 1.4, Current Document Type, Under Revision since 3/12/2016.

[196] Ghazal Faraj and András Micsik. Representing and validating cultural heritage knowledge graphs in cidoc-crm ontology. *Future Internet*, 13(11):277, 2021.

[197] Aline JE Deicke. Modeling as a scholarly process: The impact of modeling decisions on data-driven research practices. In *In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.):Graph Technologies in the Humanities - Proceedings 2020, published at http://ceur-ws.org*, 2020.

[198] Laura Willot, D Vodislav, Livio De Luca, and Valérie Gouet-Brunet. Automatic structuring of photographic collections for spatio-temporal monitoring of restoration sites: Problem statement and challenges. In *ISPRS WG II/8 9th International Workshop 3D-ARCH" 3D Virtual Reconstruction and Visualization of Complex Architectures"*, volume 46, pages 521–528, 2022.

[199] Y Yuexin Huang, S Suihuai Yu, J Jianjie Chu, H Hao Fan, and B Bin Du. Using knowledge graphs and deep learning algorithms to enhance digital cultural heritage management. *Heritage Science*, 11(1):204, 2023.

[200] Manolis Gergatsoulis, Georgios Papaioannou, Eleftherios Kalogeros, Ioannis Mpismpikopoulos, Katerina Tsiouprou, and Robert Carter. Modelling archaeological buildings using cidoc-crm and its extensions: the case of fuwairit, qatar. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23*, pages 357–372. Springer, 2021.

[201] Ivo Santos, Renata Vieira, Cassia Trojahn, Leonor Rocha, and Enrique Cerrillo Cuenca. Megalithism representation in cidoc-crm. In *European Conference on Advances in Databases and Information Systems*, pages 550–558. Springer, 2022.

[202] Paola Ronzino, Nicola Amico, Achille Felicetti, and Franco Niccolucci. European standards for the documentation of historic buildings and their relationship with cidoc-crm. In *CRMEX@ TPDL*, pages 70–79, 2013.

[203] Paola Ronzino, Franco Niccolucci, Achille Felicetti, and Martin Doerr. Crm ba a crm extension for the documentation of standing buildings. *International Journal on Digital Libraries*, 17:71–78, 2016.

[204] Su Yang and Miaole Hou. Knowledge graph representation method for semantic 3d modeling of chinese grottoes. *Heritage Science*, 11(1):266, 2023.

[205] Jonas Bruschke, Cindy Kröber, Ronja Utescher, and Florian Niebling. Towards querying multimodal annotations using graphs. In *Workshop on Research and Education in Urban History in the Age of Digital Libraries*, pages 65–87. Springer, 2023.

[206] Yannis Tzitzikas, Michalis Mountantonakis, Pavlos Fafalios, and Yannis Marketakis. Cidoc-crm and machine learning: a survey and future research. *Heritage*, 5(3):1612–1636, 2022.

[207] R Garozzo, F Murabito, C Santagati, C Pino, and C Spampinato. Culto: An ontology-based annotation tool for data curation in cultural heritage. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:267–274, 2017.

[208] Luigi Asprino, Luana Bulla, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. A large visual question answering dataset for cultural heritage. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 193–197. Springer, 2021.

[209] Waldo Tobler. On the first law of geography: A reply. *Annals of the association of American geographers*, 94(2):304–310, 2004.

[210] Ceri Binding and Douglas Tudhope. Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries*, 17:5–21, 2016.

[211] Raissa Garozzo, Cettina Santagati, Concetto Spampinato, and Giuseppe Vecchio. Knowledge-based generative adversarial networks for scene understanding in cultural heritage. *Journal of Archaeological Science: Reports*, 35:102736, 2021.

[212] Abhishek V Potnis, Surya S Durbha, and Rajat C Shinde. Semantics-driven remote sensing scene understanding framework for grounded spatio-contextual scene descriptions. *ISPRS International Journal of Geo-Information*, 10(1):32, 2021.

[213] Daniel André Barbosa Marçal. Object detection of megalithic dolmens in google satellite imagery. Master's thesis, ISCTE, Lisbon, Portugal, 2023.

[214] Francisco Gerson Amorim de Meneses and Lanna Letícia Goes Silva Oliveira. Estudo prático sobre a geolocalização de sítios arqueológicos no google earth. *Revista Arqueologia Pública*, 8(1 [9]):35–46, 2014.

[215] Abdelaziz Elfadaly, Nicodemo Abate, Nicola Masini, and Rosa Lasaponara. Sar sentinel 1 imaging and detection of palaeo-landscape features in the mediterranean area. *Remote Sensing*, 12(16):2611, 2020.

[216] Roboflow. Roboflow. `https://roboflow.com/`, 2022. Accessed: 1 June 2024.

[217] Ronald J Brachman. *Knowledge Representation and Reasoning*. Morgan Kaufman/Elsevier, 2004.

[218] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.

[219] Ian Horrocks, Peter F Patel-Schneider, and Frank Van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Journal of web semantics*, 1(1):7–26, 2003.

[220] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? *Handbook on ontologies*, pages 1–17, 2009.

[221] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[222] Federico Bianchi, Gaetano Rossiello, Luca Costabello, Matteo Palmonari, and Pasquale Minervini. Knowledge graph embeddings and explainable ai. In *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges*, pages 49–72. IOS Press, 2020.

[223] Edison Andrade Martins Morais and Ana Paula L Ambrósio. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens. *Relatório Técnico–RT-INF-001/07*, 2007.

[224] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

[225] Lars Vogt. Fair data representation in times of escience: a comparison of instance-based and class-based semantic representations of empirical data using phenotype descriptions as example. *Journal of Biomedical Semantics*, 12:1–25, 2021.

[226] Jesús Barrasa and Jim Webber. *Building Knowledge Graphs.* " O'Reilly Media, Inc.", 2023.

[227] Francis Harvey and Robert G Raskin. Spatial cyberinfrastructure: building new pathways for geospatial semantics on existing infrastructures. In *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications*, pages 87–96. Springer, 2011.

[228] Kristin Stock, Gobe Hobona, Carlos Granell, and Mike Jackson. Ontology-based geospatial approaches for semantic awareness in earth observation systems. In *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications*, pages 97–118. Springer, 2011.

[229] T. Liebig. Neo4j: A reasonable rdf graph database reasoning engine. `/`. Accessed: 01.09.2020.

[230] Ester Alba, Mar Gaitán, Arabella León, Dunia Mladenić, and Janez Brank. Weaving words for textile museums: the development of the linked silknow thesaurus. *Heritage Science*, 10:1–14, 12 2022.

[231] Gonçalo Melo da Silva, Ana Celeste Glória, Ângela Sofia Salgueiro, Bruno Almeida, Daniel Monteiro, Marco Roque de Freitas, and Nuno Freire. Rossio infrastructure: A digital humanities platform to explore the portuguese cultural heritage. *Information*, 13(2):50, 2022.

[232] FISH. Forum on information standards in heritage. terminology working group. `http://www.heritagedata.org/blog/about-heritage-data/fish/`, n.d. Accessed: 30-May-2023.

[233] A. Singhal. Introducing the knowledge graph: Things, not strings. `https://www.blog.google/products/search/introducing-knowledge-graph-things-not/`, 2012. Accessed: 2024-12-05.

[234] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102, 2023.

[235] Stefano Ferilli. Integration strategy and tool between formal ontology and graph database technology. *Electronics*, 10(21):2616, 2021.

[236] Daniel Hernández, Aidan Hogan, Cristian Riveros, Carlos Rojas, and Enzo Zerega. Querying wikidata: Comparing sparql, relational and graph databases. In *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15*, pages 88–103. Springer, 2016.

[237] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.", 2015.

[238] Liana Stanescu. A comparison between a relational and a graph database in the context of a recommendation system. In *FedCSIS (Position Papers)*, pages 133–139, 2021.

[239] Petri Kotiranta, Marko Junkkari, and Jyrki Nummenmaa. Performance of graph and relational databases in complex queries. *Applied sciences*, 12(13):6490, 2022.

[240] Chaimaa Messaoudi, Rachida Fissoune, and Hassan Badir. A performance evaluation of nosql databases to manage proteomics data. *International Journal of Data Mining and Bioinformatics*, 21(1):70–89, 2018.

[241] Steve Ataky Tsham Mpinda, Lucas Cesar Ferreira, Marcela Xavier Ribeiro, and Marilde Terezinha Prado Santos. Evaluation of graph databases performance through indexing techniques. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 6(5):87–98, 2015.

[242] Martin Macak, Matus Stovcik, Barbora Buhnova, and Michal Merjavy. How well a multi-model database performs against its single-model variants: Benchmarking orientdb with neo4j and mongodb. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 463–470. IEEE, 2020.

[243] Justin J Miller. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324, pages 141–147, 2013.

[244] Daniel Marçal, Ariele Câmara, João Oliveira, and Ana de Almeida. Evaluating r-cnn and yolo v8 for megalithic monument detection in satellite images. In *International Conference on Computational Science*, pages 162–170. Springer, 2024.

[245] Direção-Geral do Território. Carta de uso e ocupação do solo - cos. `https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/metadata/e2049ec4-0ea8-4211-bf62-64a9135777cd?tab=responsible`, 1995. Versão 1995.

[246] Direção-Geral do Território. Carta de uso e ocupação do solo - cos. `https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/metadata/0839b251-8d9b-4eee-9316-5d92e2bff5ff`, 2007. Versão 2007.

[247] Direção-Geral do Território. Carta de uso e ocupação do solo - cos. https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/metadata/e581be49-7202-4160-9190-d07a43c6e8a6, 2010. Versão 2010.

[248] Direção-Geral do Território. Carta de uso e ocupação do solo - cos. https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/metadata/94dab05c-513d-4a40-b9e3-9ddd775f8543, 2015. Versão 2015.

[249] Direção-Geral do Território. Especificações técnicas da carta de uso e ocupação do solo (cos) de portugal continental para 2018. Technical report, Direção-Geral do Território, Lisboa, 2019. Eds: Mário Caetano e Filipe Marcelino.

[250] Bob Booth, Andy Mitchell, et al. Getting started with arcgis, 2001.

[251] ISO/IEC. Iso/iec 8859-1: Latin-1, iso 8859-1: 1987. International Standard, 1987. Available at: https://www.iso.org/standard/28245.html.

[252] Robert Battle and Dave Kolas. Geosparql: enabling a geospatial semantic web. *Semantic Web Journal*, 3(4):355–370, 2011.

[253] Cogan Shimizu, Rui Zhu, Gengchen Mai, Colby Fisher, Ling Cai, Mark Schildhauer, Krzysztof Janowicz, Pascal Hitzler, Lu Zhou, and Shirly Stephen. A pattern for features on a hierarchical spatial grid. In *Proceedings of the 10th International Joint Conference on Knowledge Graphs*, pages 108–114, 2021.

[254] Jenn Riley. Understanding metadata. *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)*, 23:7–10, 2017.

[255] Open Geospatial Consortium. Opengis r implementation standard for geographic information - simple feature access - part 1: Common architecture. Technical Report 06-103r4, Open Geospatial Consortium, 2011.

[256] Norbert Zaniewicz and Andrzej Salamończyk. Comparison of mongodb, neo4j and arangodb databases using the developed data generator for nosql databases. *Studia Informatica. System and information technology*, 26(1):61–72, 2022.

[257] Jessica AM Stothers and Andrew Nguyen. Can neo4j replace postgresql in healthcare? *AMIA Summits on Translational Science Proceedings*, 2020:646, 2020.

[258] Robert Campbell McColl, David Ediger, Jason Poovey, Dan Campbell, and David A Bader. A performance evaluation of open source graph databases. In *Proceedings of the first workshop on Parallel programming for analytics applications*, pages 11–18, 2014.

[259] Salim Jouili and Valentin Vansteenberghe. An empirical comparison of graph databases. In *2013 International Conference on Social Computing*, pages 708–715. IEEE, 2013.

[260] Sotirios Beis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Benchmarking graph databases on the problem of community detection. In *New Trends in Database and Information Systems II: Selected papers of the 18th East European Conference on Advances in Databases and Information Systems and Associated Satellite Events,*

ADBIS 2014 Ohrid, Macedonia, September 7-10, 2014 Proceedings II, pages 3–14. Springer, 2015.

[261] Naw Thiri Wai Khin, Nyo Nyo Yee, and Aung Aung Hein. Performance comparison between neo4j-based and owl-based query classification process. *International Journal of Scientific and Research Publications*, 2019.

[262] DaTAIKU. Dataiku documentation. `https://doc.dataiku.com/dss/latest/concepts/index.html`, 2024. Acessado em: 25 set. 2024.

[263] Sean Gillies. The shapely user manual. *URL https://pypi. org/project/Shapely*, 2013.

[264] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pages 5491–5500. PMLR, 2020.

[265] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

CHAPTER 8

Appendices

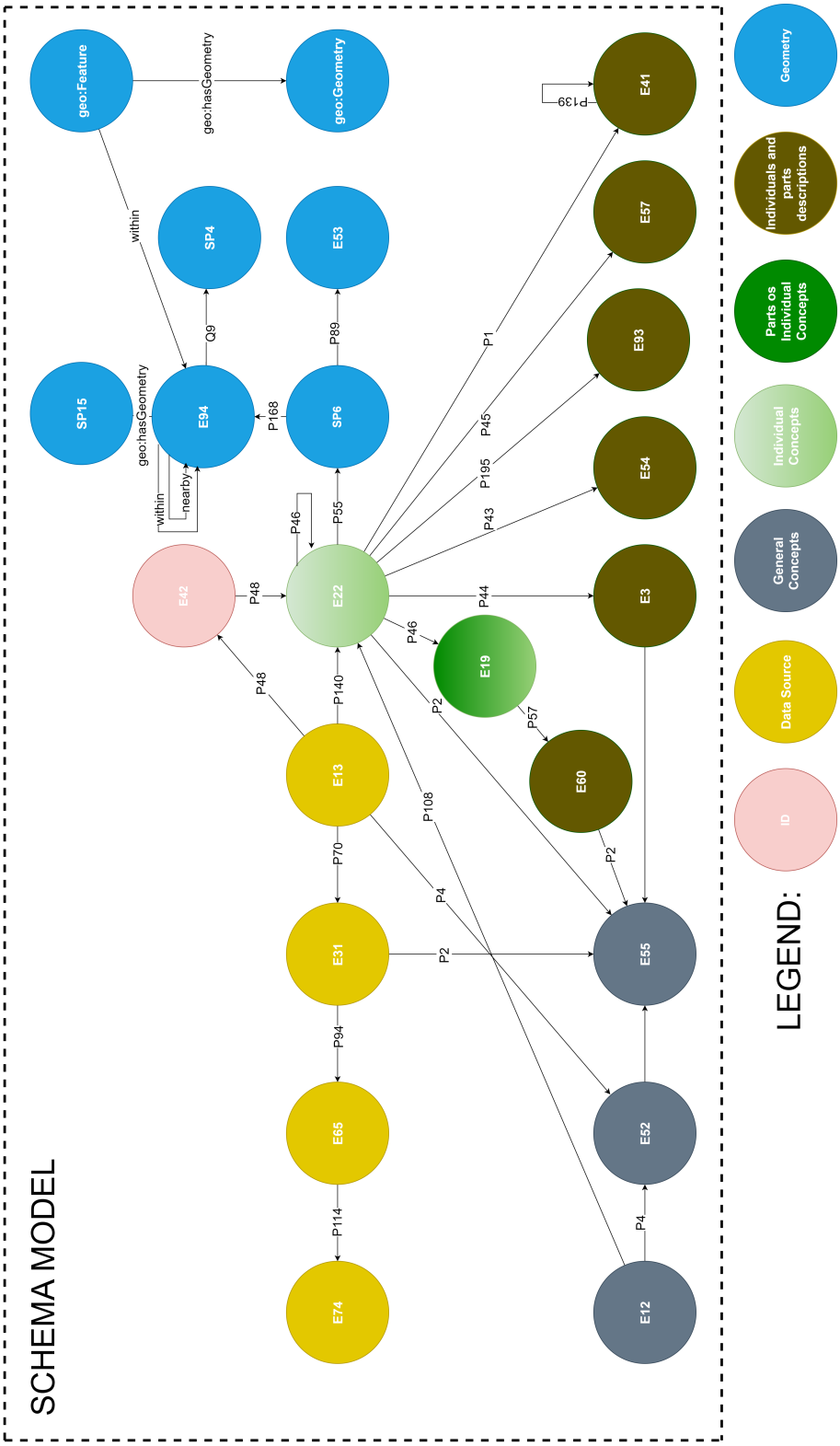


FIGURE 8.1. Model schema of the LPG used to represent monuments.

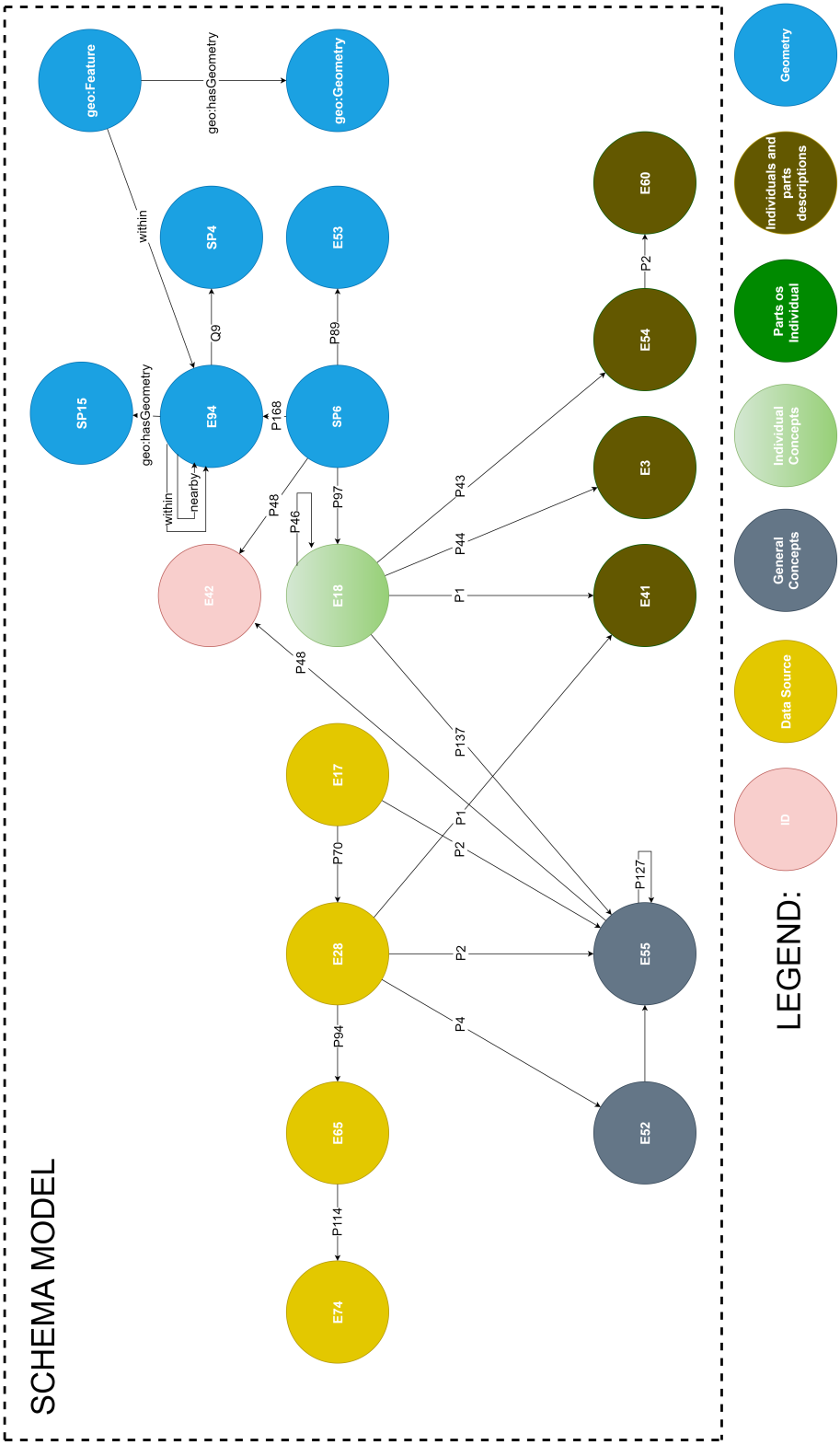


FIGURE 8.2. Model schema of the LPG used to represent the VTMs.

Type	Code	Description
Solos Incipientes	A	Solos Incipientes - Aluviossolos Modernos, Não Calcários, de textura mediana
	Al	Solos Incipientes - Aluviossolos Modernos, Não Calcários, de textura ligeira
	At	Solos Incipientes - Aluviossolos Antigos, Não Calcários, de textura mediana
	Atl	Solos Incipientes - Aluviossolos Antigos, Não Calcários, de textura ligeira
	Eg	Solos Incipientes - Litossolos dos Climas de Regime Xérico, de granitos ou quartzodioritos
	Egn	Solos Incipientes - Litossolos dos Climas de Regime Xérico, de gnaisses ou rochas afins
	Ex	Solos Incipientes - Litossolos dos Climas de Regime Xérico, de xistos ou grauvaques
	Rg	Solos Incipientes - Regossolos Psamíticos, Normais, não húmidos
	Sb	Solos Incipientes - Solos de Baixas (Coluviossolos), Não Calcários, de textura mediana
	Sbl	Solos Incipientes - Solos de Baixas (Coluviossolos), Não Calcários, de textura ligeira
Barros Castanhos	Bvc	Barros Castanho-Avermelhados, Calcários, Muito Descarbonatados, de dioritos ou gabros ou rochas cristalofílicas básicas associados a calcário friável
Barros Castanho-Avermelhados	Cpv	Barros Castanho-Avermelhados, Calcários, Pouco Descarbonatados, de rochas eruptivas ou cristalofílicas básicas associadas a calcário friável, ou de grés argilosos calcários, ou margas
Barros Pretos	Cp	Barros Pretos, Calcários, Pouco Descarbonatados, de rochas eruptivas ou cristalofílicas básicas associadas a calcário friável, ou de grés argilosos calcários ou margas
Solos Hidromórficos	Ca	Solos Hidromórficos, Sem Horizonte Eluvial, Para-Aluviossolos (ou Para-Coluviossolos), de aluviões ou coluviais de textura mediana
	Cac	Solos Hidromórficos, Sem Horizonte Eluvial, Para-Aluviossolos (ou Para-Coluviossolos), de aluviões ou coluviais de textura mediana, calcários
	Cal	Solos Hidromórficos, Sem Horizonte Eluvial, Para-Aluviossolos (ou Para-Coluviossolos), de aluviões ou coluviais de textura ligeira
	Ps	Solos Hidromórficos, Com Horizonte Eluvial, Planossolos, de arenitos ou conglomerados argilosos ou argilas

Solos Argiluviados Pouco Insaturados	Pac	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos de Materiais Calcários, Para-Barros, de margas ou calcários margosos ou de calcários não compactos associados com xistos, grés argilosos, argilitos ou argilas ou de grés argilosos calcários (de textura franca a franco-argilosa)
	Pag	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Não Calcários, Para-Solos Hidromórficos, de arenitos ou conglomerados argilosos ou argilas (de textura arenosa ou franco-arenosa)
	Pbc	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Calcários, Para-Barros, de calcários margosos associados a arcoses ou rochas afins
	Pdc	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Calcários, Para-Solos Hidromórficos, de arcoses ou rochas afins associadas a depósitos calcários
	Pdg	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Não Calcários, Para-Solos Hidromórficos, de arcoses ou rochas afins
	Pgn	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Não Calcários, Normais, de gnaisses ou rochas afins
	Pm	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Não Calcários, Para-Barros, de dioritos ou quartzodioritos ou rochas microfaneríticas ou cristalofílicas afins
	Pmg	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Não Calcários, Normais, de quartzodioritos
	Pv	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Não Calcários, Normais, de rochas cristalofílicas
	Px	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Pardos, de Materiais Não Calcários, Normais, de xistos ou grauvaques
	Sr	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Não Calcários, Normais, de de "rañas" ou depósitos afins
	Vcc	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Calcários, Normais, de calcários cristalinos ou mármores ou rochas cristalofílicas cálcio-siliciosas
	Vcd	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Calcários, Normais, de calcários compactos ou dolomias
	Vcm	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Calcários, Para-Barros, de margas ou calcários margosos
	Vgn	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Não Calcários, Normais, de gnaisses ou rochas afins
	Vx	Solos Argiluviados Pouco Insaturados - Solos Mediterrâneos, Vermelhos ou Amarelos, de Materiais Não Calcários, Normais, de xistos ou grauvaques

Solos Litólicos	Par	Solos Litólicos, Não Húmicos Pouco Insaturados, Normais, de materiais arenáceos pouco consolidados (de textura arenosa a franco-arenosa)
	Pg	Solos Litólicos, Não Húmicos Pouco Insaturados, Normais, de granitos
	Ppg	Solos Litólicos, Não Húmicos, Pouco Insaturados, Normais, de rochas microfíricas claras
	Vt	Litólicos, Não Húmicos, Pouco Insaturados Normais, de arenitos grosseiros
Solos Calcários	Pc	Solos Calcários, Pardos dos Climas de Regime Xérico, Normais, de calcários não compactos
	Pcd	Solos Calcários, Pardos dos Climas de Regime Xérico, Para-Litossolos, de calcários compactos (travertinos)
	Pcs	Solos Calcários, Pardos dos Climas de Regime Xérico, Normais, de margas ou materiais afins
	Vc	Solos Calcários, Vermelhos dos Climas de Regime Xérico, Normais, de calcários
	Vct	Solos Calcários, Vermelhos dos Climas de Regime Xérico, Normais, de arenitos grosseiros associados a depósitos calcários
	Vcx	Solos Calcários, Vermelhos dos Climas de Regime Xérico, Normais, de xistos ou grauvaques associados a depósitos calcários
Solos Podzolizados	Ppt	Solos Podzolizados - Podzóis, (Não Hidromórficos), Com Surraipa, com A2 incipiente, de ou sobre arenitos
Afloramentos Rochosos	Arg	Afloramento Rochoso de granitos ou quartzodioritos

TABLE 8.1. Detailed explanation of the Portuguese Solos Charter [1]for Pavia (46 types of solos)

POIs ID	Object Detection Score 1	KG-ML Score 1	KG-ML Prediction	Explanations
FPC1	995	449.618	0	”soilUse”: 0.22989976704085627, ”soilType”: 0.7867751618777192, ”distWater”: -0.2568191903075665
FPC3	9.951	449.625	0	”soilUse”: 0.22989828934496748, ”soilType”: 0.7867839642088739, ”distWater”: -0.2568118279666314
FPC4	8.367	202.995	0	”soilUse”: 0.21598968510329564, ”soilType”: -0.23505485741077892, ”distWater”: -0.3171490607871281
FPC5	7.481	203.163	0	”soilType”: -0.0797594191946489, ”relief”: -0.21837644064467843, ”distWater”: -0.4304651105910129
FPC6	9.824	271.296	0	”relief”: -0.2951803095207308, ”soilUse”: 0.3479574057208671, ”distWater”: -0.4596893656854788
FPC7	9.939	214.446	0	”score”: 0.052397982736154924, ”relief”: -0.2215257557524073, ”distWater”: -0.4123684765717135
FPC8	9.946	214.479	0	”score”: 0.05259252363584377, ”relief”: -0.22153472450983358, ”distWater”: -0.4123169401851655
FPC9	9.867	236.889	0	”soilUse”: 0.16515961915586708, ”relief”: -0.30687900051610206, ”distWater”: -0.41076054138310014
FPC10	9.974	197.002	0	”soilType”: 0.06635292406596927, ”soilUse”: -0.10891082602957392, ”distWater”: -0.12591324017501515
FPC11	7.105	154.029	0	”relief”: -0.2281958404741382, ”soilUse”: -0.30260694584258085, ”distWater”: -0.43899024698930655
FPC12	9.934	62.082	1	”relief”: -0.04608784164083085, ”soilType”: 0.4823567331516542, ”distWater”: -0.1978373970604736
FPC14	8.236	131.561	0	”soilUse”: -0.23904444019384652, ”soilType”: -0.4061352741933566, ”distWater”: -0.532798317096719
FPC18	9.408	780.785	1	”soilUse”: 0.38765264850442216, ”soilType”: 0.9263562952328672, ”distWater”: 1.4144990595952684
FPC19	8.094	452.394	0	”soilUse”: 0.2238042688467437, ”soilType”: -0.5559506167553067, ”distWater”: 0.923290822543101
FPC20	9.569	237.051	0	”soilUse”: -0.2720753950748639, ”soilType”: 0.30498802758261623, ”distWater”: -0.3052567801138163
FPC23	955	693.321	1	”soilUse”: -0.13549604404726634, ”relief”: 0.15174257470090136, ”soilType”: 2.072831554701101
FPC24	9.966	181.612	0	”soilUse”: 0.16231976422546057, ”soilType”: -0.4418911857299319, ”distWater”: -0.49684082392054285
FPC26	9.217	395.091	0	”soilType”: -0.21152978424319252, ”soilUse”: -0.22968706316359744, ”distWater”: 0.9569971064522595

FPC29	9.271	39.529	0	”soilType”: -0.2116233227798077, ”soilUse”: -0.23011514734918134, ”distWater”: 0.9569724173709525
FPC31	9.857	150.708	0	”soilType”: -0.24329342840259272, ”soilUse”: -0.2575094099047539, ”distWater”: -0.35628994990610807
FPC32	9.965	531.124	1	”soilType”: -0.22679124110608442, ”soilUse”: 0.303821024958959, ”distWater”: 0.9437254752148441
FPC37	9.848	261.744	0	”soilUse”: -0.14771504442364647, ”soilType”: 0.3000267003074555, ”distWater”: -0.3456844113886606
FPC38	9.827	53.014	1	”soilType”: -0.23084297298194523, ”soilUse”: 0.30296140754882606, ”distWater”: 0.942464041087625
FPC39	8.402	446.831	0	”score”: 0.013995627770624769, ”soilType”: -0.20259393561750405, ”distWater”: 1.0032998424054353
FPC40	8.627	169.834	0	”soilUse”: -0.09738086620000508, ”soilType”: -0.23790276166658209, ”distWater”: -0.3745260111999562
FPC44	9.848	172.622	0	”soilUse”: -0.09547580422616031, ”soilType”: -0.23020949649824574, ”distWater”: -0.3793194470232917
FPC45	9.914	458.278	0	”score”: 0.06020179537892126, ”soilType”: -0.16902143079653847, ”distWater”: 1.0192387035379016
FPC46	9.602	455.912	0	”score”: 0.050667189364511256, ”soilType”: -0.17620596646112063, ”distWater”: 1.0159497321931072
FPC47	9.905	159.408	0	”soilUse”: -0.13372615688872402, ”soilType”: -0.5236805796854314, ”distWater”: -0.41502354403808206
FPC63	9.913	565.123	1	”soilType”: 0.2748698030911674, ”relief”: -0.30283846170627726, ”distWater”: 1.4536728225467135
FPC69	9.915	333.855	0	”soilUse”: 0.2046164362765962, ”relief”: 0.7751260757354288, ”distWater”: -0.5791549491059756
FPC53	8.299	122.796	0	”soilUse”: -0.22209652909766842, ”soilType”: -0.458853235768393, ”distWater”: -0.3784599517637699
FPC76	9.954	213.303	0	”soilUse”: 0.3207653460105917, ”soilType”: -0.6599848157119274, ”distWater”: -0.6140612119523329
FPC61	9.889	210.124	0	”soilUse”: 0.3082214790295075, ”soilType”: -0.33403215410865317, ”distWater”: -0.5350962154722311
FPC91	9.718	126.854	0	”soilUse”: -0.15877216092447655, ”soilType”: -0.4412252133049974, ”distWater”: -0.452385588935053

FPC86	8.204	664.835	1	”soilType”: 0.15566747672817915, ”relief”: 0.33728850465244054, ”distWater”: 0.8820125327624436
FPC92	9.883	145.438	0	”relief”: -0.02510068304171864, ”soilType”: -0.5191036902962307, ”distWater”: -0.4549890018735021
FPC70	8.576	168.923	0	”soilUse”: -0.1661750358440297, ”soilType”: -0.2704605296854432, ”distWater”: -0.44641918713813866
FPC60	8.481	131.992	0	”soilUse”: -0.2374479860732306, ”soilType”: -0.3785400108480166, ”distWater”: -0.5423812392794907
FPC80	9.903	212.509	0	”soilUse”: 0.1991387036724046, ”soilType”: -0.5503984598633342, ”distWater”: -0.51902957301917
FPC95	8.878	566.798	1	”soilUse”: 0.3430223952423181, ”soilType”: -0.6188856081351858, ”distWater”: 1.4252805998111926
FPC59	8.885	206.683	0	”soilUse”: 0.28378342013561686, ”soilType”: -0.336480811124501, ”distWater”: -0.44333116461400535
FPC73	9.217	380.477	0	”soilUse”: 0.26606029655794183, ”soilType”: 0.2949779638014228, ”distWater”: -0.31399746181000293
FPC83	9.823	335.607	0	”distWater”: -0.1903691786014884, ”relief”: 0.428153701518481, ”soilUse”: 0.3184515387294621
FPC90	8.623	334.425	0	”relief”: -0.17234590808016892, ”soilType”: -1.0471590565266917, ”distWater”: 0.4357049171897306
FPC96	7.741	237.032	0	”soilType”: -0.321135382134933, ”soilUse”: 0.476646856269491, ”distWater”: -0.3855345455814201
FPC99	9.982	66.153	1	”soilType”: 0.16796795485411975, ”relief”: 0.33844459492315515, ”distWater”: 0.8225747639308485
FPC64	9.551	332.221	0	”soilUse”: 0.2031619734357949, ”relief”: 0.7740113574972235, ”distWater”: -0.5884939865763333
FPC98	9.349	16.462	0	”soilUse”: 0.24307414722905563, ”soilType”: -0.623894309762101, ”distWater”: -0.5644393405260395
FPC75	984	156.225	0	”relief”: -0.1256545782340266, ”soilType”: -0.6286243693109843, ”distWater”: -0.6083947875566464
FPC77	9.743	156.273	0	”relief”: -0.12435035168676056, ”soilType”: -0.6265322959938129, ”distWater”: -0.6082788202223652
FPC62	9.949	15.027	0	”relief”: -0.26714013571758155, ”soilType”: -0.6302935803893088, ”distWater”: -0.5411848023544508
FPC57	9.882	261.939	0	”soilUse”: -0.14764065671950655, ”soilType”: 0.3006557258603837, ”distWater”: -0.34546383045535056

FPC72	7.433	157.417	0	"relief": -0.09329093597349725, "soilType": -0.5785101074281611, "distWater": -0.6055171239111241
FPC52	9.784	192.899	0	"relief": -0.08758339408485427, "soilUse": -0.25853395922569744, "distWater": -0.4274822523042683
FPC88	7.214	228.808	0	"soilUse": -0.2595799423624008, "soilType": 0.27100069841984764, "distWater": -0.31417097961641516
FPC94	9.936	628.758	1	"soilType": 0.09999747829620387, "relief": 0.45120011595025333, "distWater": 1.6327552801839245
FPC55	9.455	275.996	0	"soilType": 0.14789957570369205, "soilUse": 0.21447669079084974, "distWater": -0.35472000585124397
FPC97	8.736	227.507	0	"soilType": -0.09632491032410417, "relief": -0.10290479757633597, "distWater": -0.09660099847912473
FPC58	9.944	172.843	0	"soilUse": -0.09532602048374517, "soilType": -0.22972543524378164, "distWater": -0.37969632649834506
FPC78	975	212.582	0	"soilUse": 0.3171789949779449, "soilType": -0.6590190315049408, "distWater": -0.6136435114895642
FPC100	9.895	386.497	0	"soilUse": 0.2704533217919689, "soilType": 0.2976845228106447, "distWater": -0.30720924262131544
FPC56	7.285	136.365	0	"soilUse": -0.09317811549562327, "soilType": -0.4856207944029247, "distWater": -0.40328706095403644

TABLE 8.2. KG-ML results from analyzing POIs determined incorrectly as TP by pre-classification (object detection), now with updated scores. Additionally, the image explains what factors influenced KG-ML's score.