

Utilisation des LLM pour l'extraction automatique d'annotations conformes au modèle CIDOC-CRM

HARIRI Ali

ali.hariri@ensma.fr

<https://www.lias-lab.fr/members/alihariri>

Sommaire

1. Cas d'étude
2. Problématiques
3. CIDOC CRM
4. Modélisation et annotation
 - Annotation manuel
 - Annotation semi-automatique (LLM)
 - Expérimentation
 - Évaluation
 - Résultat
 - Impact environnemental
5. Perspectives de recherche
 - Sélection automatique des classes et propriétés
 - Fine-tuning

RAPPEL DE L'ÉTUDE DE CAS

Hypogée des Dunes

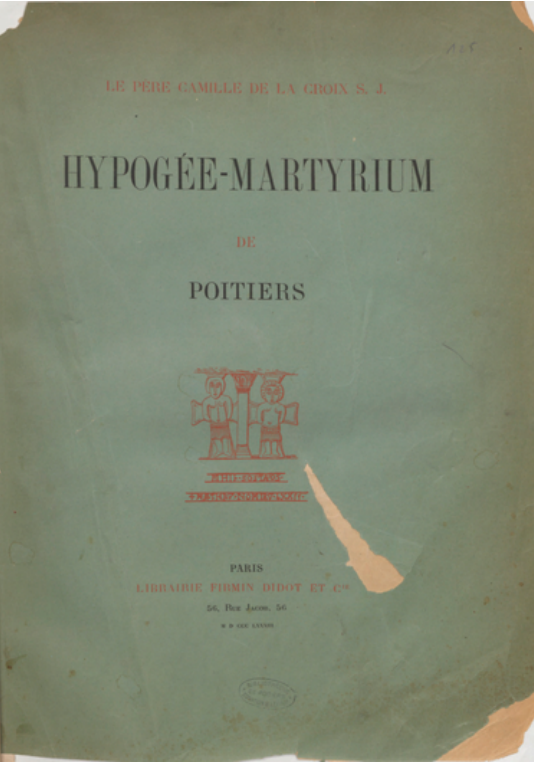
Hypogée des Dunes, Poitiers (VIe-VIIe siècles)

Un site funéraire mérovingien comprenant :

- Des chambres creusées dans la roche
- Des sépultures ornées de sculptures

Les documents

- Données non structurées
 - Articles scientifiques
 - Documents de travail
- Données structurées
 - 8 tableurs Excel
 - inventaires
 - résultats d'analyse
 - document Word (inventaire des outils de taille)

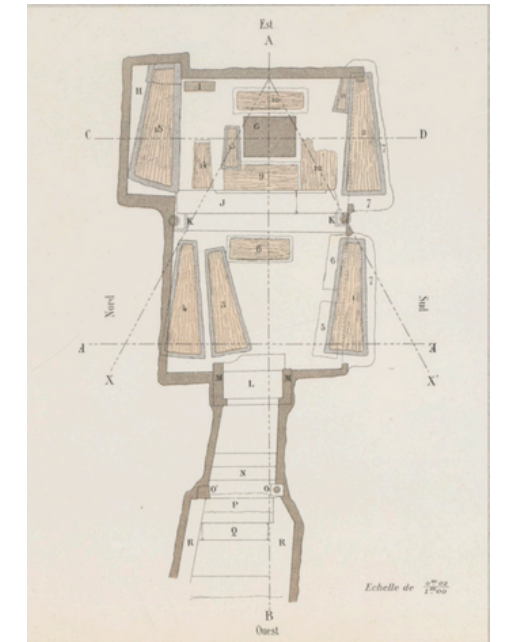
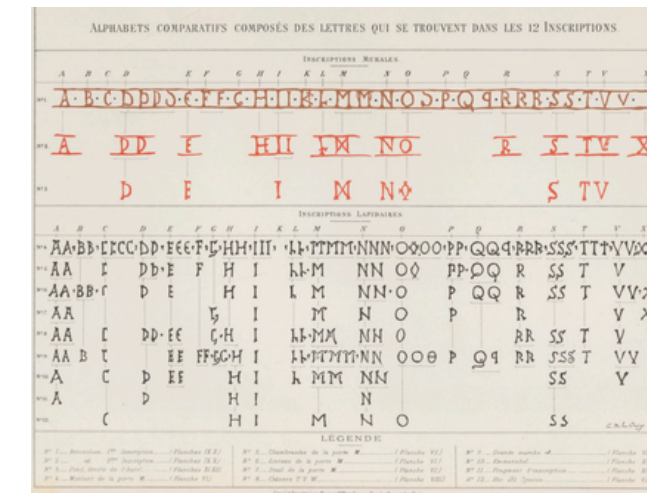


PROBLÉMATIQUES

Les problématiques rencontrées

1. Présence d'études provenant de différents domaines :

- Étude des mortiers et des enduits
- Étude géologique
- Étude des élément sculptés et inscrits



2. Hétérogénéité des données (2D/3D/texte)

“La pierre aux larrons”

Texte



2D



3D

3. Variété du vocabulaire et des synonymes utilisés

Moellon

Pierre de taille

Bloc de pierre



Hypogée des Dunes
Hypogée de Mellebaude
Tombe souterraine



CIDOC CRM

Avantages :

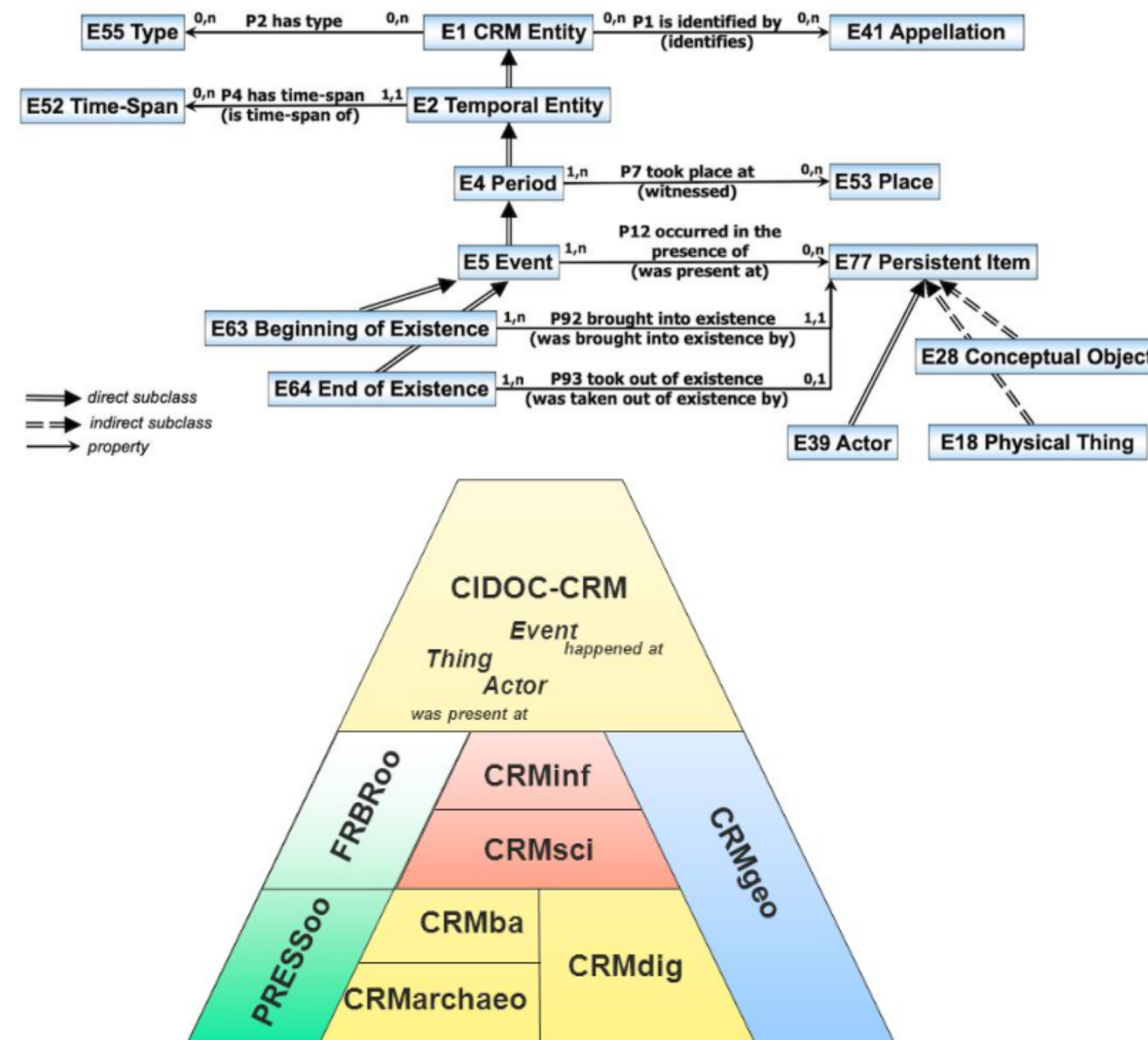
- Normes nationales et internationales
- Domaine d'archéologie et de conservation du patrimoine
- Flexibilité et extensibilité

Concepts fondamentaux :

- Relations avec les événements
- Relations spatiales
- Relations temporelles

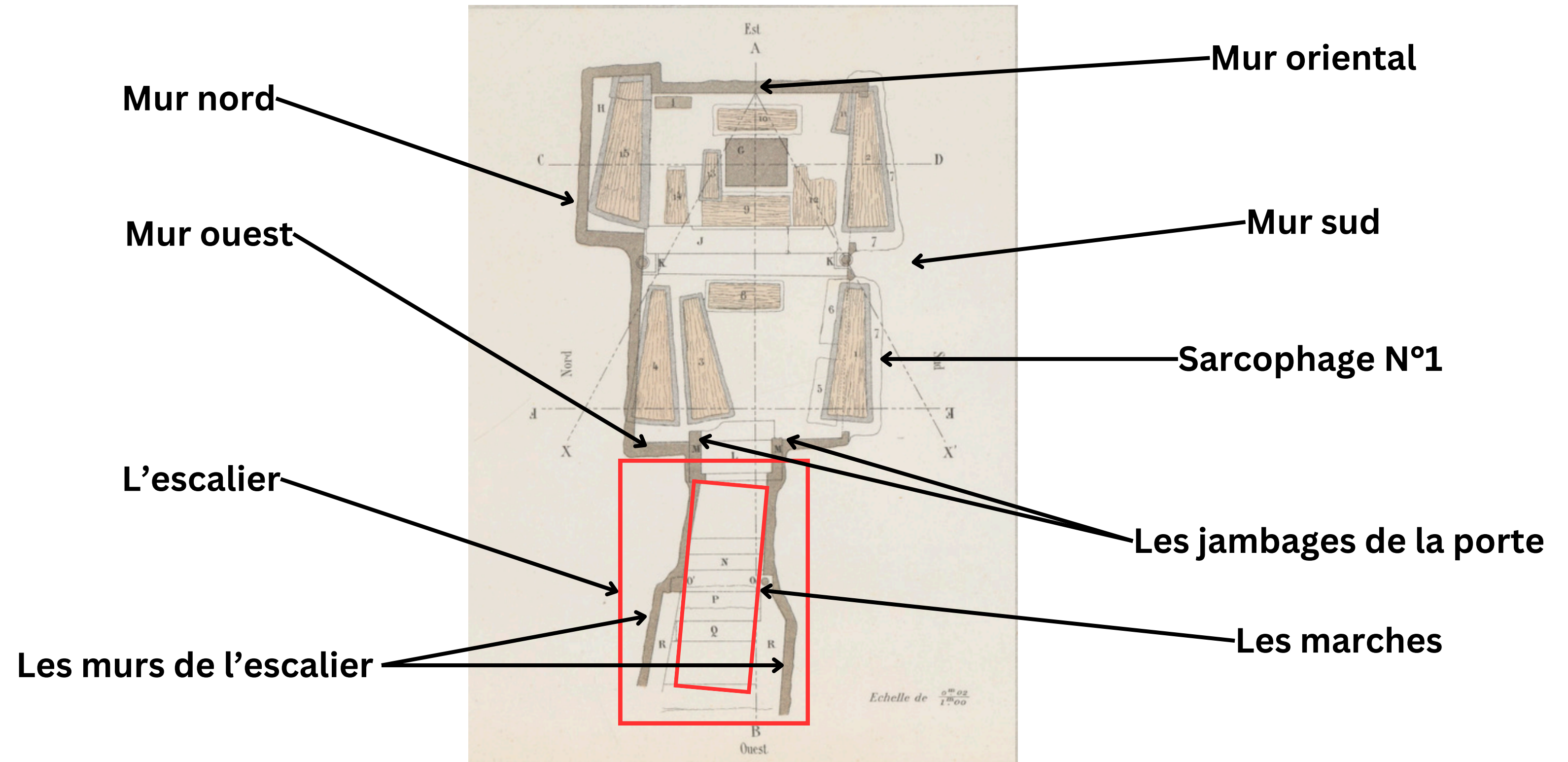
Extensions :

- CRMarchaeo: Fouille archéologique
- CRMsci: Les observations scientifiques
- CRMba: Bâtiments archéologiques

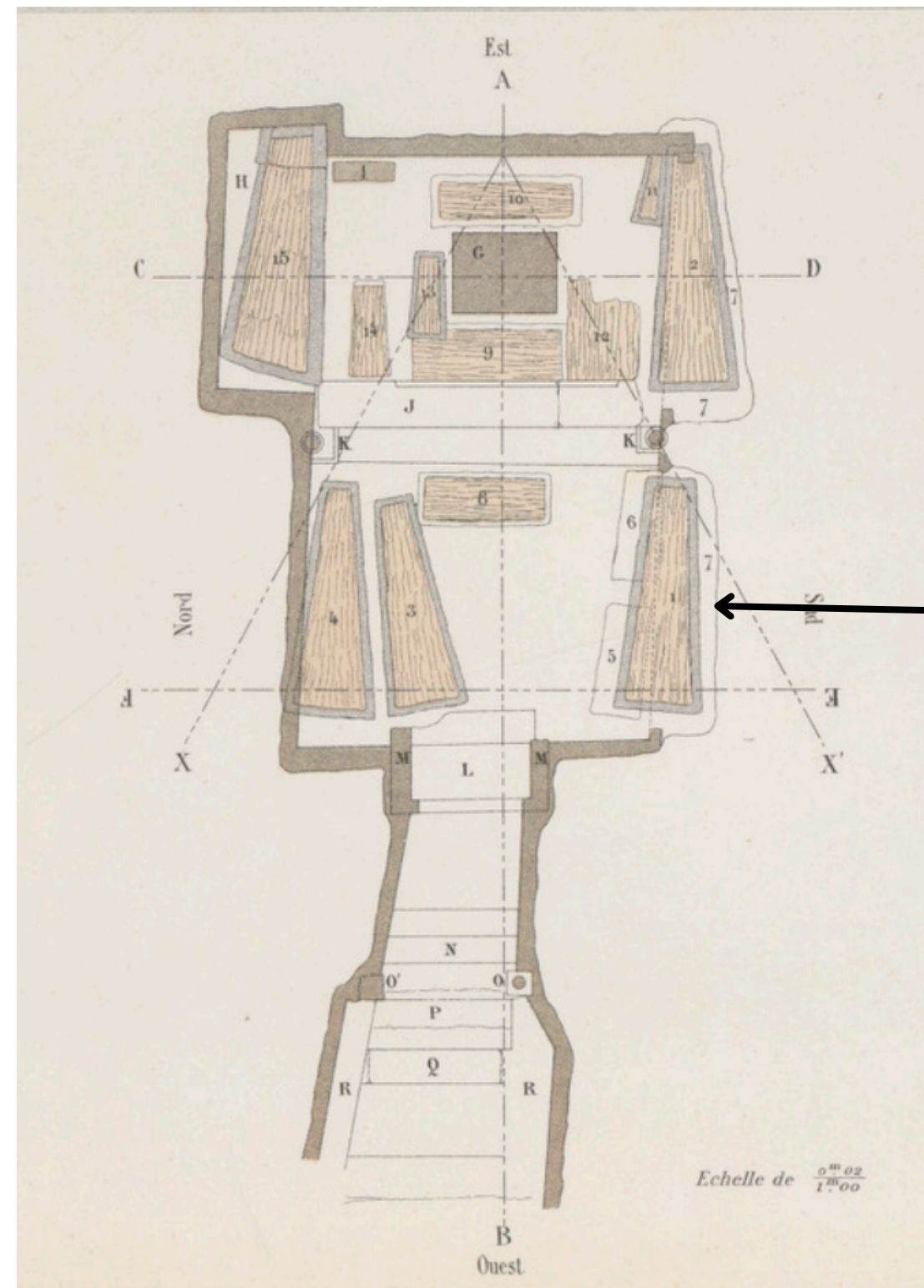


Comment l'utilise-t-on ?

ANNOTATION



ANNOTATION

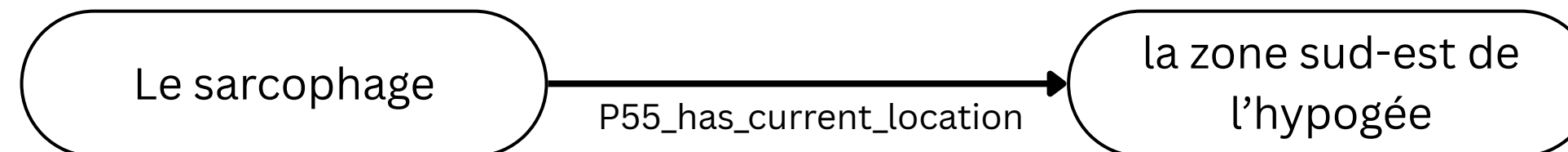


← Sarcophage N°1

LA MODÉLISATION / ANNOTATION

SARCOPHAGE N°1

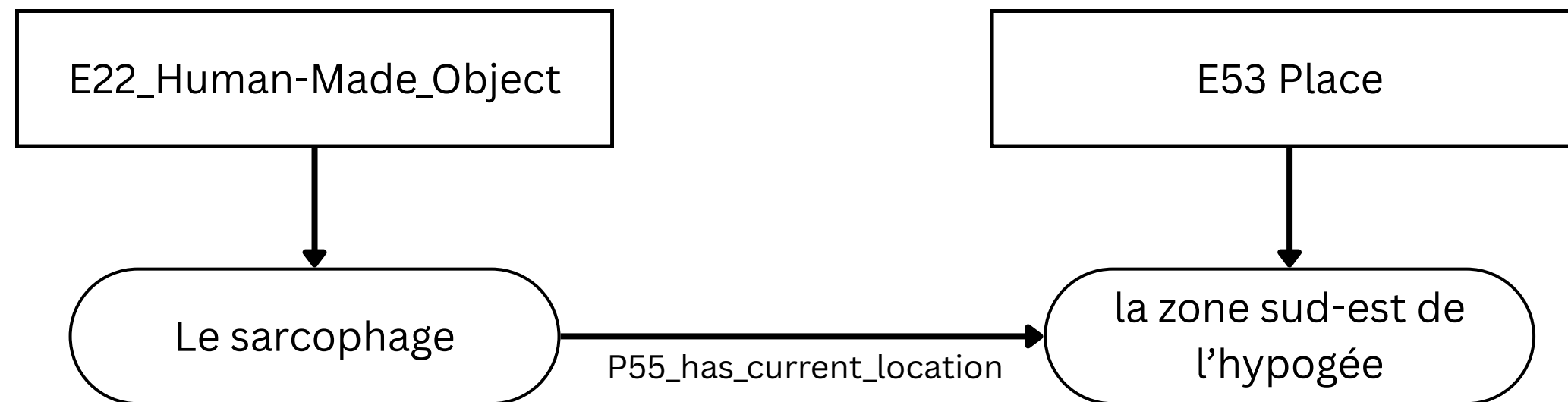
Le sarcophage est situé dans la zone sud-est de l'hypogée.



LA MODÉLISATION / ANNOTATION

SARCOPHAGE N°1

Le sarcophage est situé dans la zone sud-est de l'hypogée.

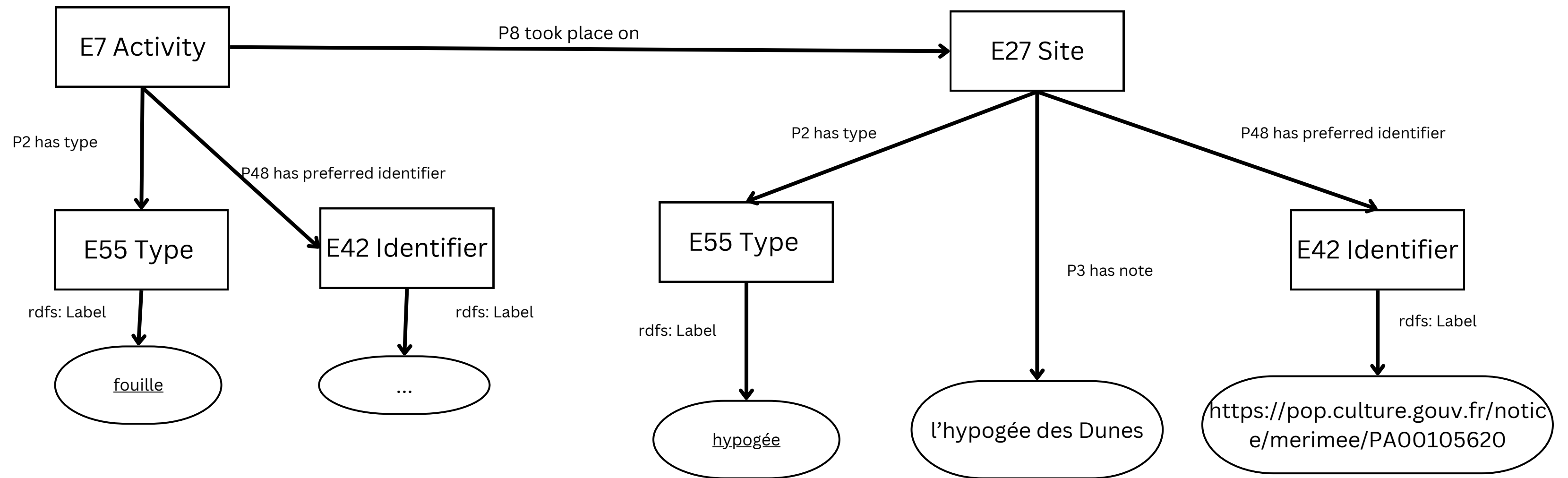


LA MODÉLISATION / ANNOTATION

Exemple 2

Les fouilles ont été réalisées dans l'hypogée des Dunes.

10 Annotations



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction Non Guidée

Context:

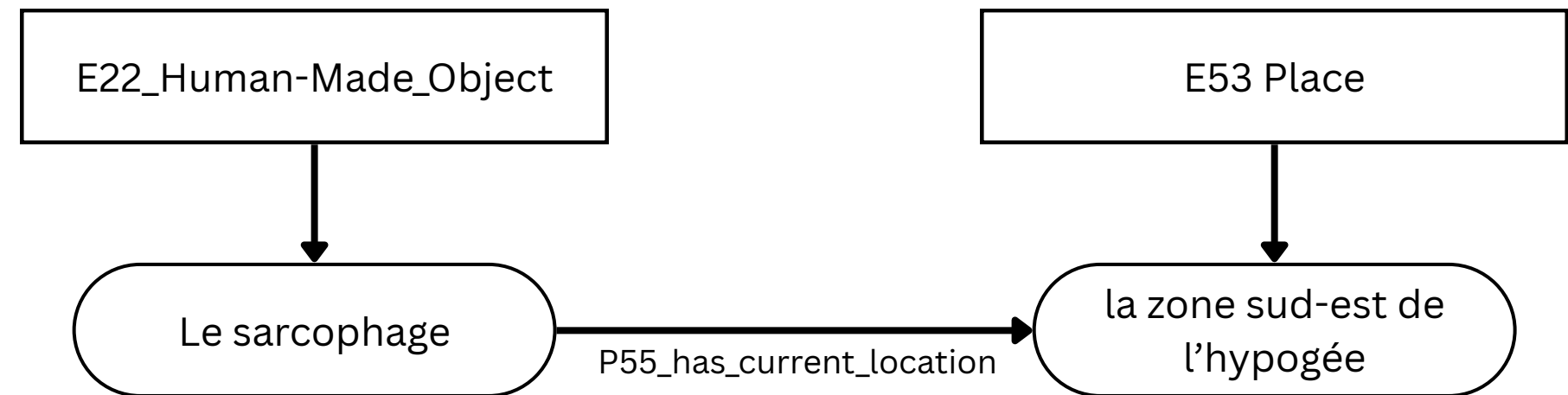
Vous êtes un expert en ontologies, en modélisation de données et en extraction de connaissances.
Votre tâche consiste à transformer des données tabulaires structurées (CSV) en triplets RDF, en vous basant sur l'ontologie CIDOC CRM.

Input data (CSV):

{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus



Résultats obtenus

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction Non Guidée

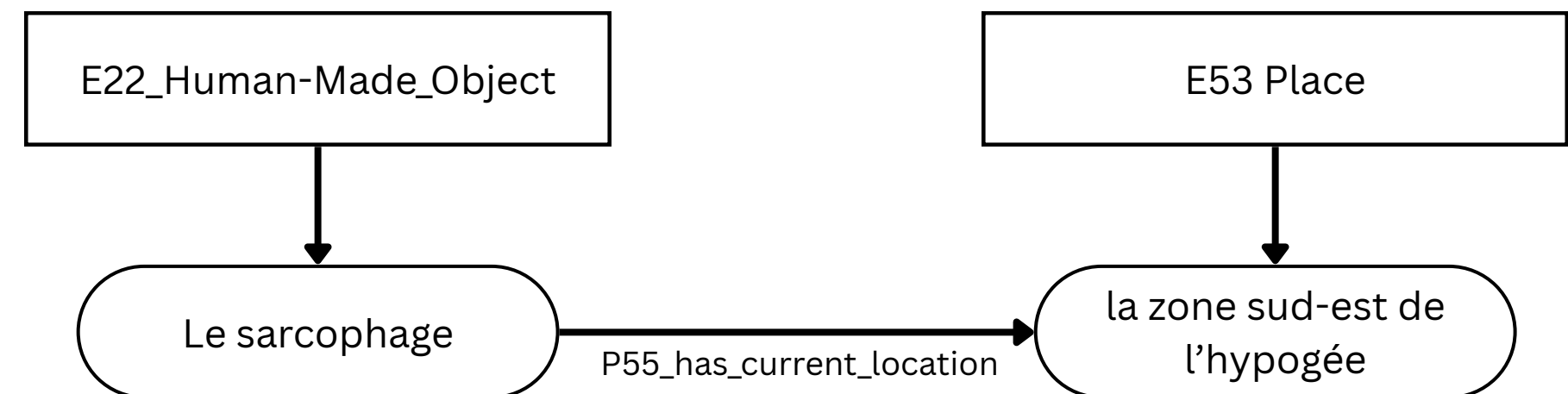
Context:

Vous êtes un expert en ontologies, en modélisation de données et en extraction de connaissances.
Votre tâche consiste à transformer des données tabulaires structurées (CSV) en triplets RDF, en vous basant sur l'ontologie CIDOC CRM.

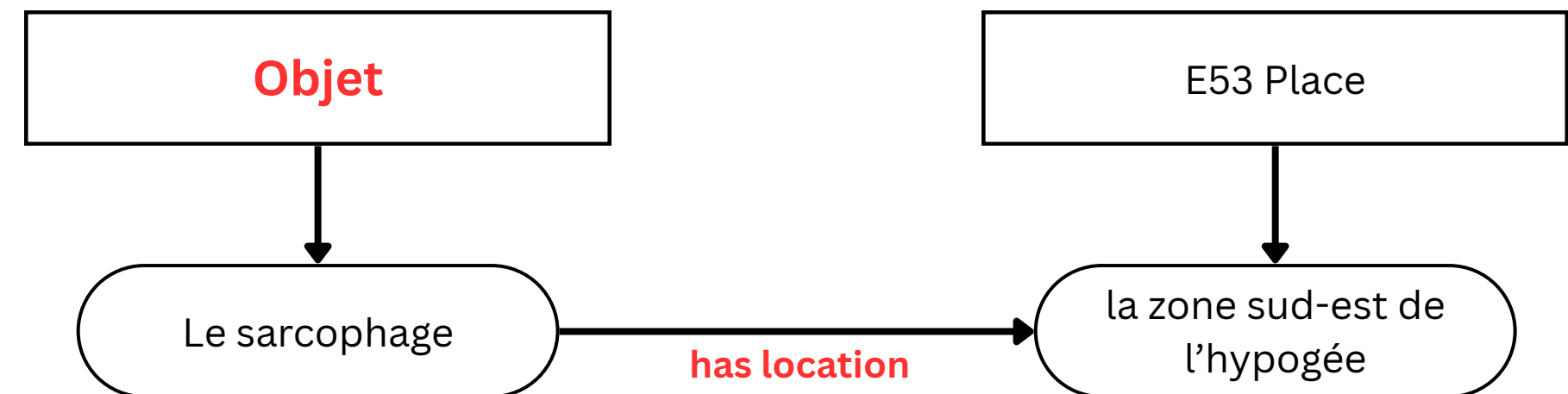
Input data (CSV):
{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus



Résultats obtenus



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction Non Guidée

Context:

Vous êtes un expert en ontologies, en modélisation de données et en extraction de connaissances.
Votre tâche consiste à transformer des données tabulaires structurées (CSV) en triplets RDF, en vous basant sur l'ontologie CIDOC CRM.

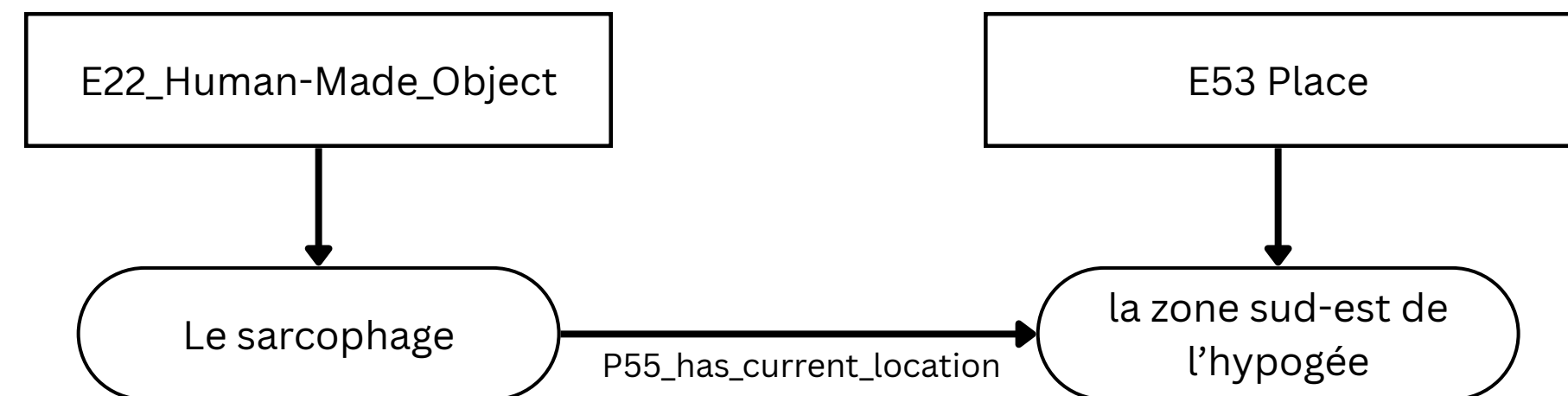
Input data (CSV):
{input_data}

Output: RDF JSON-LD (JSON-LD format)

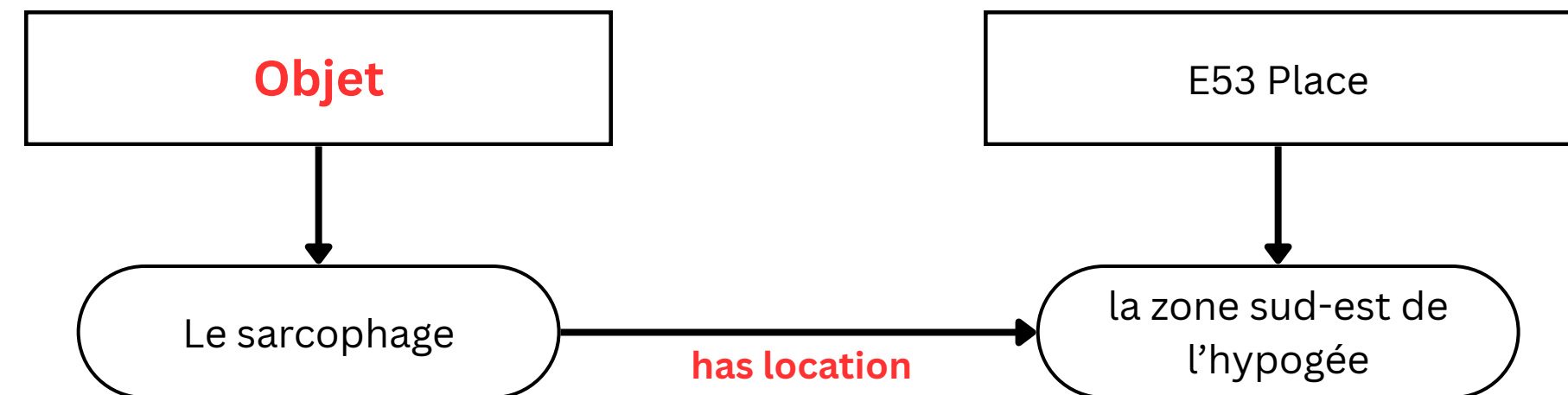
Limites:

- Hallucinations
- Structure RDF incohérente

Résultats attendus



Résultats obtenus



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction guidée avec l'ontologie complète du CIDOC CRM

Context:

You are an expert in ontologies, data modeling, and knowledge extraction.

Your task is to transform structured tabular data (CSV) into RDF triples based on CIDOC-CRM ontology.

Strict Instructions:

- Strictly follow the ontology structure and semantics provided ...

Steps to be followed:

- Analyze the given ontology...

Ontology (format JSON):

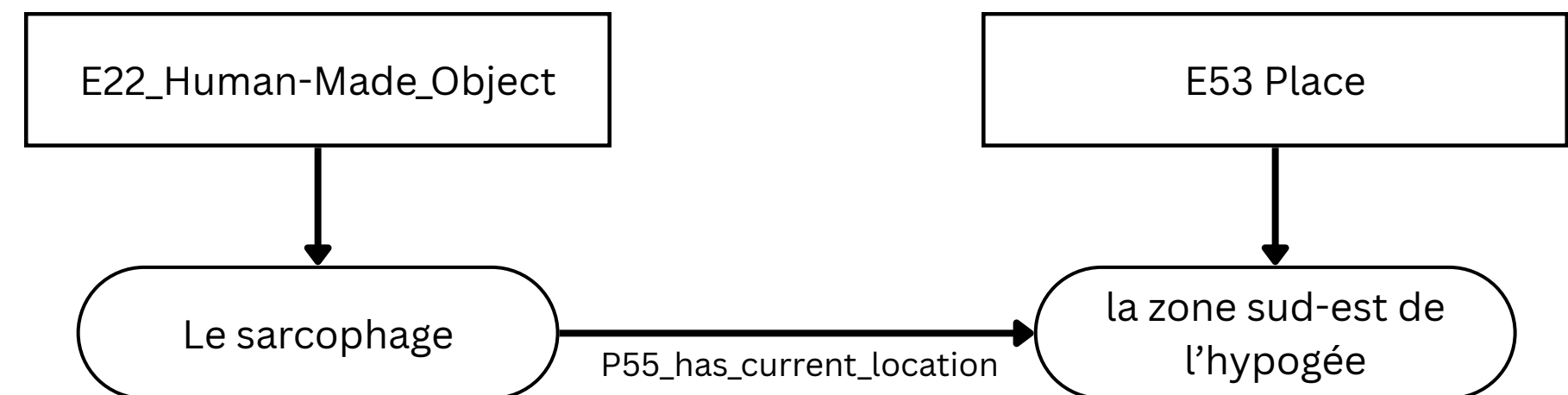
{full_ontology_description}

Input data (CSV):

{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus



Résultats obtenus

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction guidée avec l'ontologie complète du CIDOC CRM

Context:

You are an expert in ontologies, data modeling, and knowledge extraction.
Your task is to transform structured tabular data (CSV) into RDF triples based on CIDOC-CRM ontology.

Strict Instructions:

- Strictly follow the ontology structure and semantics provided ...

Steps to be followed:

- Analyze the given ontology...

Ontology (format JSON):

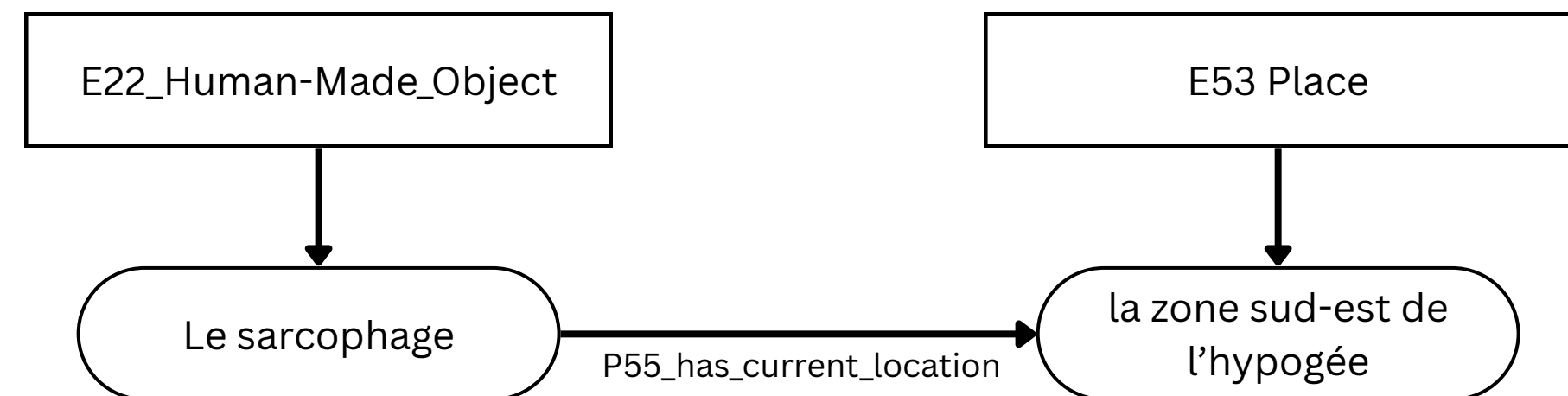
{full_ontology_description}

Input data (CSV):

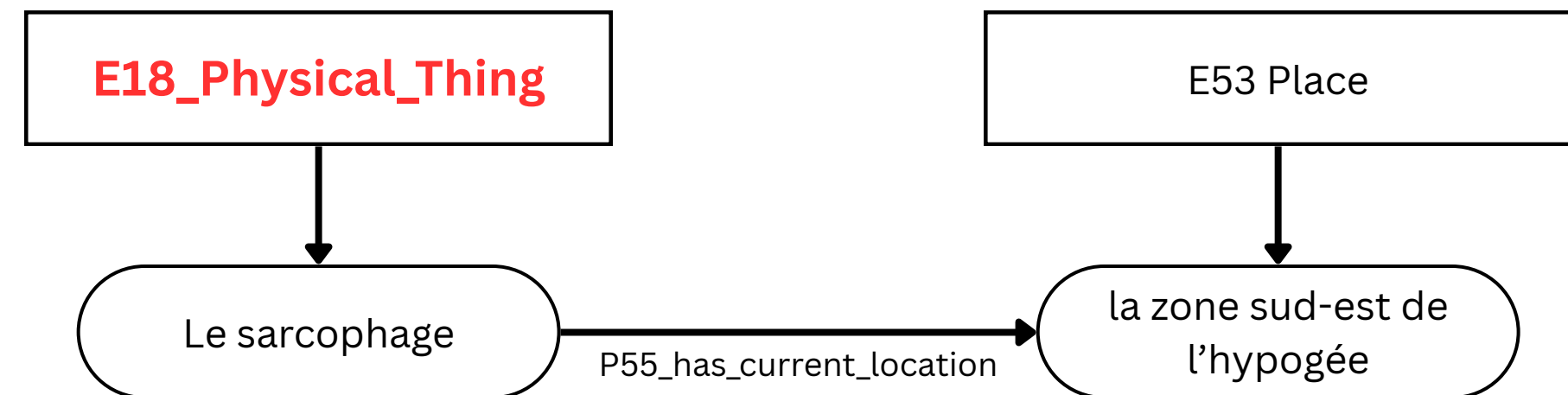
{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus



Résultats obtenus



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction guidée avec l'ontologie complète du CIDOC CRM

Context:
You are an expert in ontologies, data modeling, and knowledge extraction.
Your task is to transform structured tabular data (CSV) into RDF triples based on CIDOC-CRM ontology.

Strict Instructions:

- Strictly follow the ontology structure and semantics provided ...

Steps to be followed:

- Analyze the given ontology...

Ontology (format JSON):

{full_ontology_description}

Input data (CSV):

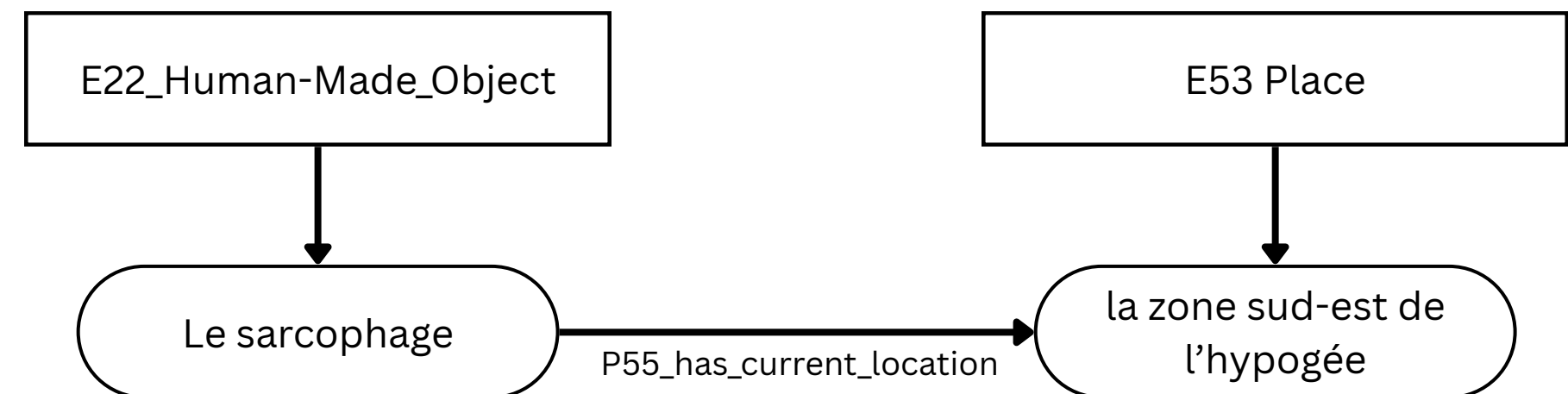
{input_data}

Output: RDF JSON-LD (JSON-LD format)

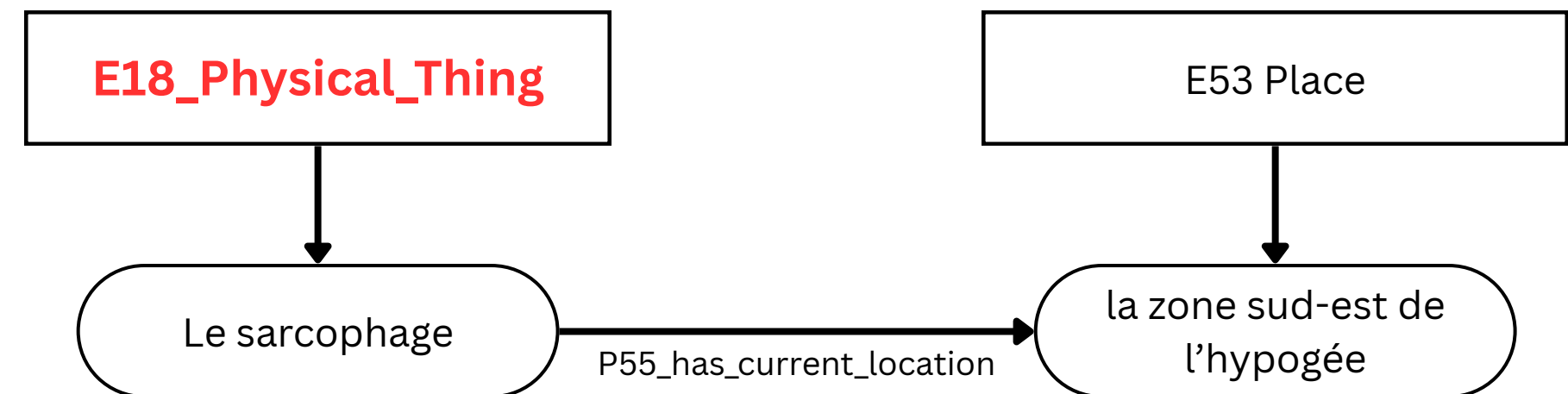
Limites:

- Prompt trop long
- Confusion due à la hiérarchie

Résultats attendus



Résultats obtenus



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction guidée avec un sous-ensemble de CIDOC CRM

Context:

You are an expert in ontologies, data modeling, and knowledge extraction.

Your task is to transform structured tabular data (CSV) into RDF triples based on CIDOC-CRM ontology.

Strict Instructions:

- Strictly follow the ontology structure and semantics provided ...

Steps to be followed:

- Analyze the given ontology...

Ontology (format JSON):

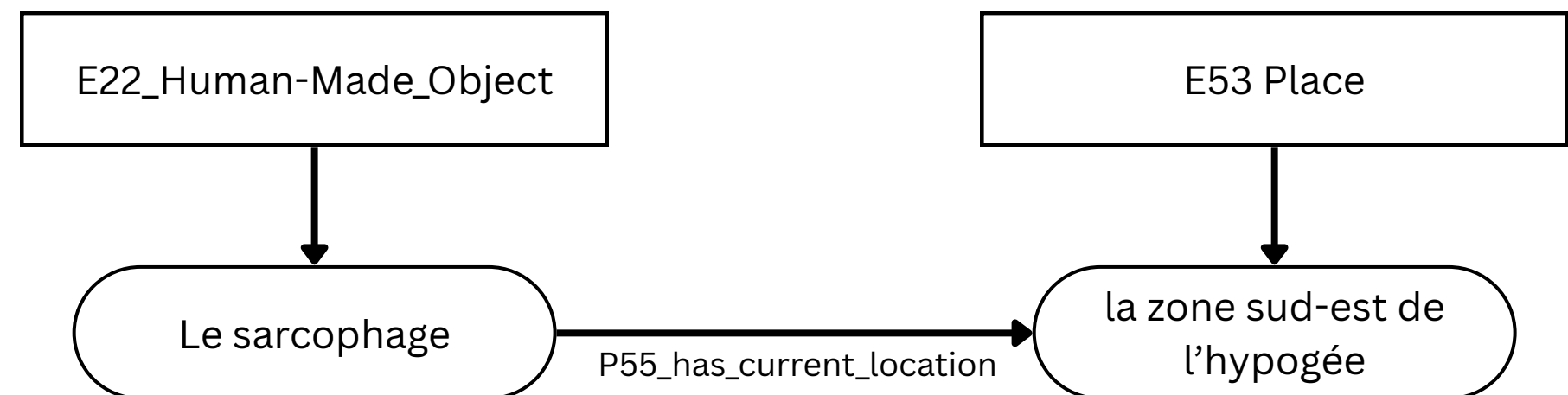
{subset_ontology_description}

Input data (CSV):

{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus

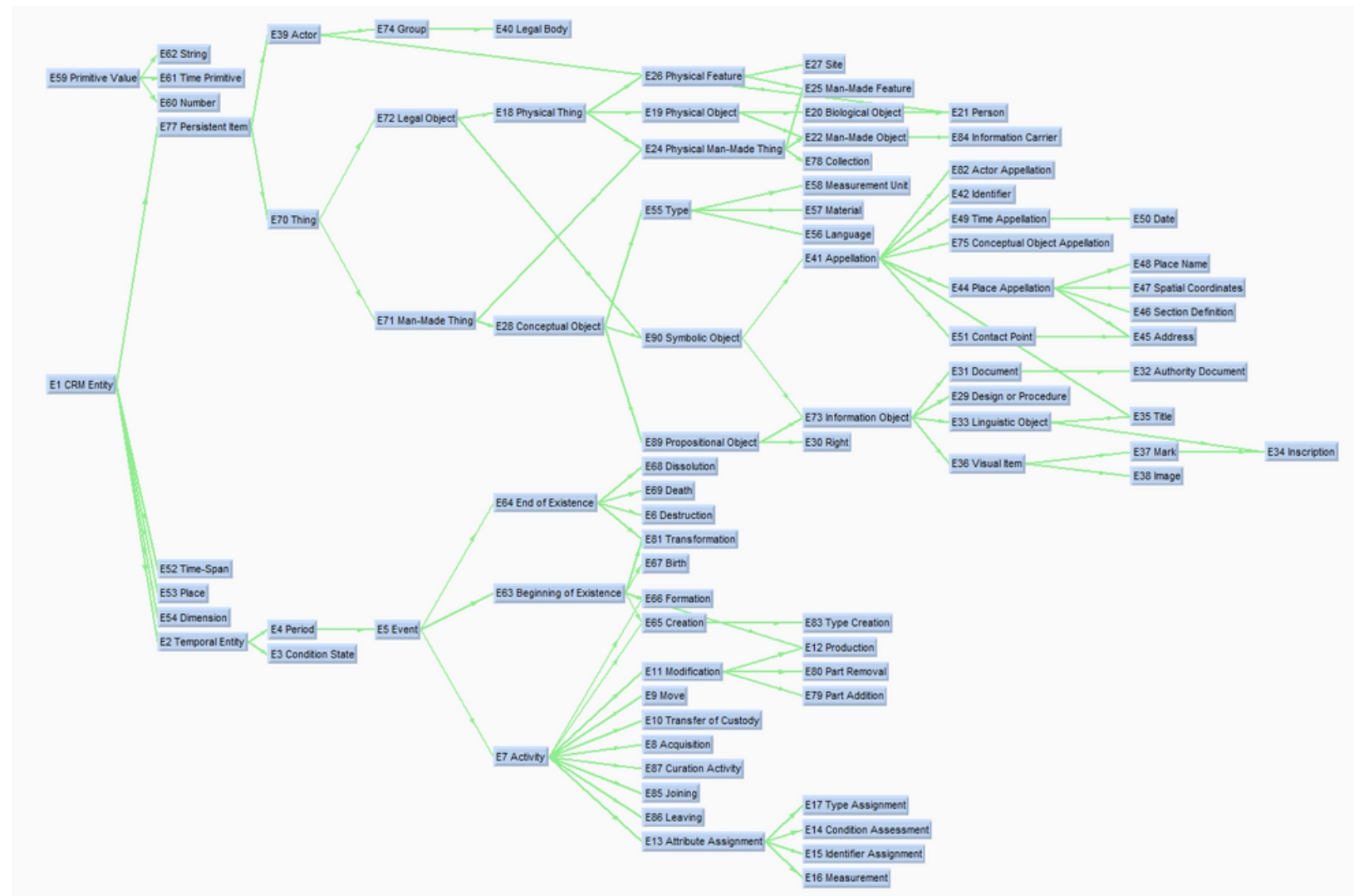


Résultats obtenus

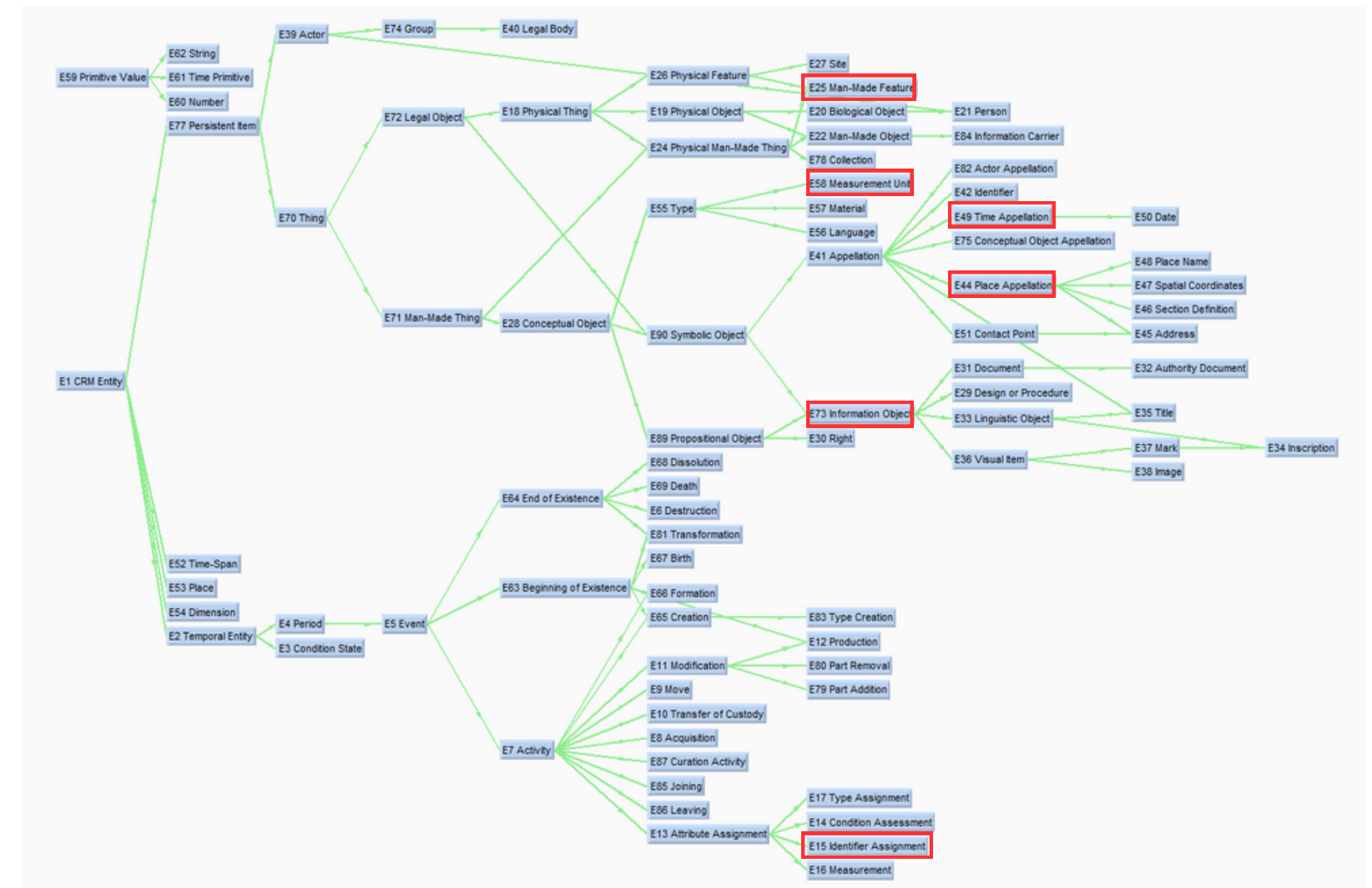
ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

CIDOC CRM



Sous-ensemble de CIDOC CRM



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction guidée avec un sous-ensemble de CIDOC CRM

Context:

You are an expert in ontologies, data modeling, and knowledge extraction.

Your task is to transform structured tabular data (CSV) into RDF triples based on CIDOC-CRM ontology.

Strict Instructions:

- Strictly follow the ontology structure and semantics provided ...

Steps to be followed:

- Analyze the given ontology...

Ontology (format JSON):

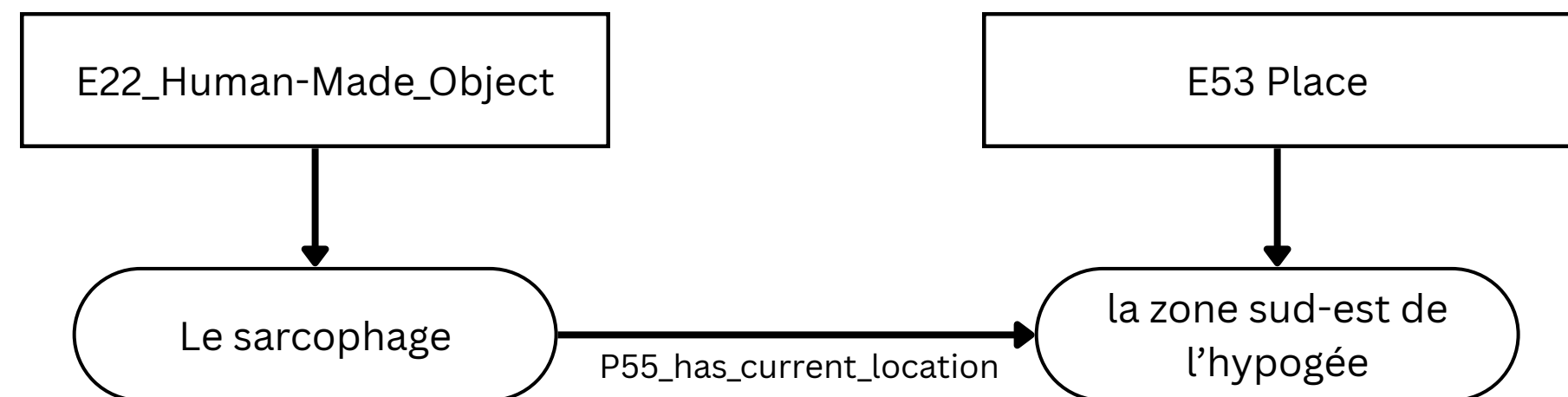
{subset_ontology_description}

Input data (CSV):

{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus



Résultats obtenus

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Expérimentation

Extraction guidée avec un sous-ensemble de CIDOC CRM

Context:

You are an expert in ontologies, data modeling, and knowledge extraction.

Your task is to transform structured tabular data (CSV) into RDF triples based on CIDOC-CRM ontology.

Strict Instructions:

- Strictly follow the ontology structure and semantics provided ...

Steps to be followed:

- Analyze the given ontology...

Ontology (format JSON):

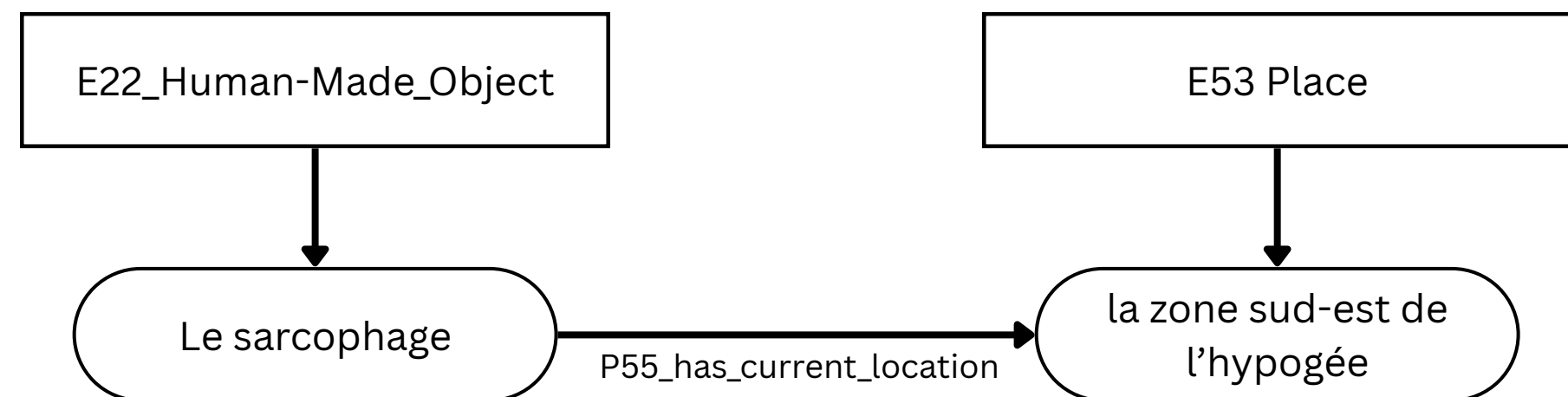
{subset_ontology_description}

Input data (CSV):

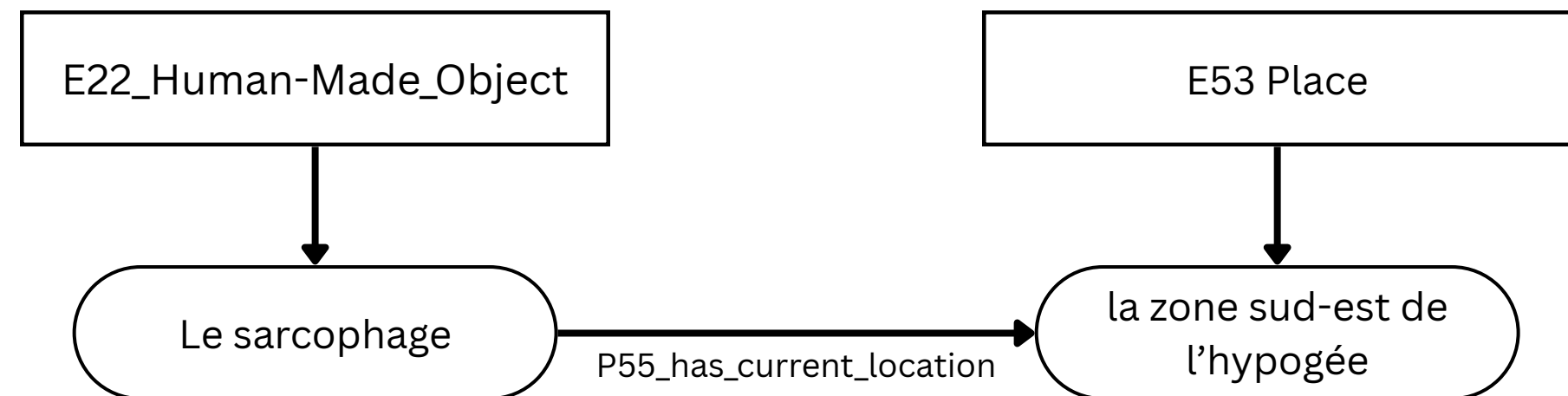
{input_data}

Output: RDF JSON-LD (JSON-LD format)

Résultats attendus



Résultats obtenus



ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Résultat (GPT4o)

Extraction Non Guidée

Aucun résultat significatif

Extraction guidée avec l'ontologie
complète du CIDOC CRM

Précision	Recall	CQs
88.1	76.8	79.3

Extraction guidée avec un sous-
ensemble de CIDOC CRM

Précision	Recall	CQs
90.2	79.1	88.5

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Résultat (GPT4o)

Extraction Non Guidée

Aucun résultat significatif

Extraction guidée avec l'ontologie complète du CIDOC CRM

Précision	Recall	CQs
88.1	76.8	79.3

Extraction guidée avec un sous-ensemble de CIDOC CRM

Précision	Recall	CQs
90.2	79.1	88.5

Exemples:

- Quelles unités stratigraphiques (US) sont localisées dans le “Mur est” ?
- Combien de faits documentés mentionnent une relation directe avec l’US05 ?
- Quels matériaux sont les plus fréquemment utilisés pour les objets de l’inventaire lapidaire ?

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Résultat (GPT4o)

Extraction Non Guidée

Aucun résultat significatif

Extraction guidée avec l'ontologie
complète du CIDOC CRM

Précision	Recall	CQs
88.1	76.8	79.3

Extraction guidée avec un sous-
ensemble de CIDOC CRM

Précision	Recall	CQs
90.2	79.1	88.5

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Résultat (Mistral)

Extraction Non Guidée

Aucun résultat significatif

Extraction guidée avec l'ontologie
complète du CIDOC CRM

Précision	Recall	CQs
63.1	52	47

Extraction guidée avec un sous-
ensemble de CIDOC CRM

Précision	Recall	CQs
68.3	58.5	53.3

ANNOTATION SEMI-AUTOMATIQUE

Large Language Models - Impact environnemental

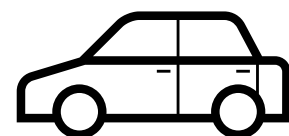


27.63 gCO₂éq

276x



5 min



102 m

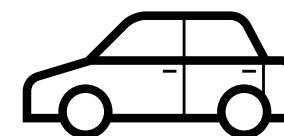


0.19 gCO₂éq

1,9x



2 sec



0.7 m

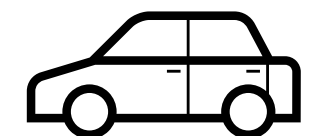


0.10 gCO₂éq

1x



1 sec



37 cm

ANNOTATION SEMI-AUTOMATIQUE

Article accepté



36^e Conférence internationale sur les bases de données
et les applications à base de systèmes experts

25 - 27 August, 2025

Towards Automating RDF Extraction for Archaeological Knowledge Graphs with LLMs

Ali Hariri^[0009–0004–9380–5138], Stéphane Jean^[0000–0002–9434–9320], and Mickaël
Baron^[0000–0002–3356–0835]

ISAE-ENSMA, Université of Poitiers, LIAS, France
{ali.hariri, stephane.jean, mickael.baron}@ensma.fr

Abstract. Archaeological sites pose significant challenges in terms of accessibility and data integration. The Semantic Web, particularly through the CIDOC CRM ontology, enables structured data sharing. However, manually constructing knowledge graphs remains a complex and labor-intensive task. This study explores semi-automatic methods leveraging large language models (LLMs) to generate knowledge graphs for archaeological data. We propose different prompting strategies based on varying levels of information about CIDOC CRM and experiment with diverse prompt patterns. Experiments conducted on real datasets demonstrate that providing LLMs with a carefully selected subset of the CIDOC CRM ontology, combined with few-shot prompt patterns, enhances RDF extraction and improves performance in answering competency questions.

Keywords: Knowledge Extraction · LLMs · Ontology · CIDOC CRM

1 Introduction

The *Hypogeum of the Dunes* in Poitiers, France, is a significant Merovingian monument that exemplifies the complexity of cultural heritage research. Its study demands interdisciplinary collaboration across archaeology, geology, epigraphy, and other related fields. However, research on this site faces challenges, including developing a shared vocabulary for data integration, aligning discipline-specific data formats, and addressing restricted physical access due to the monument's fragility. These challenges are representative of broader issues in managing and preserving cultural heritage sites. Addressing them is a key objective of the French ANR project Digitalis (ANR-22-CE38-0011-01), which aims to develop a collaborative digital infrastructure to support interdisciplinary research through advanced data integration, semantic interoperability, and remote accessibility.

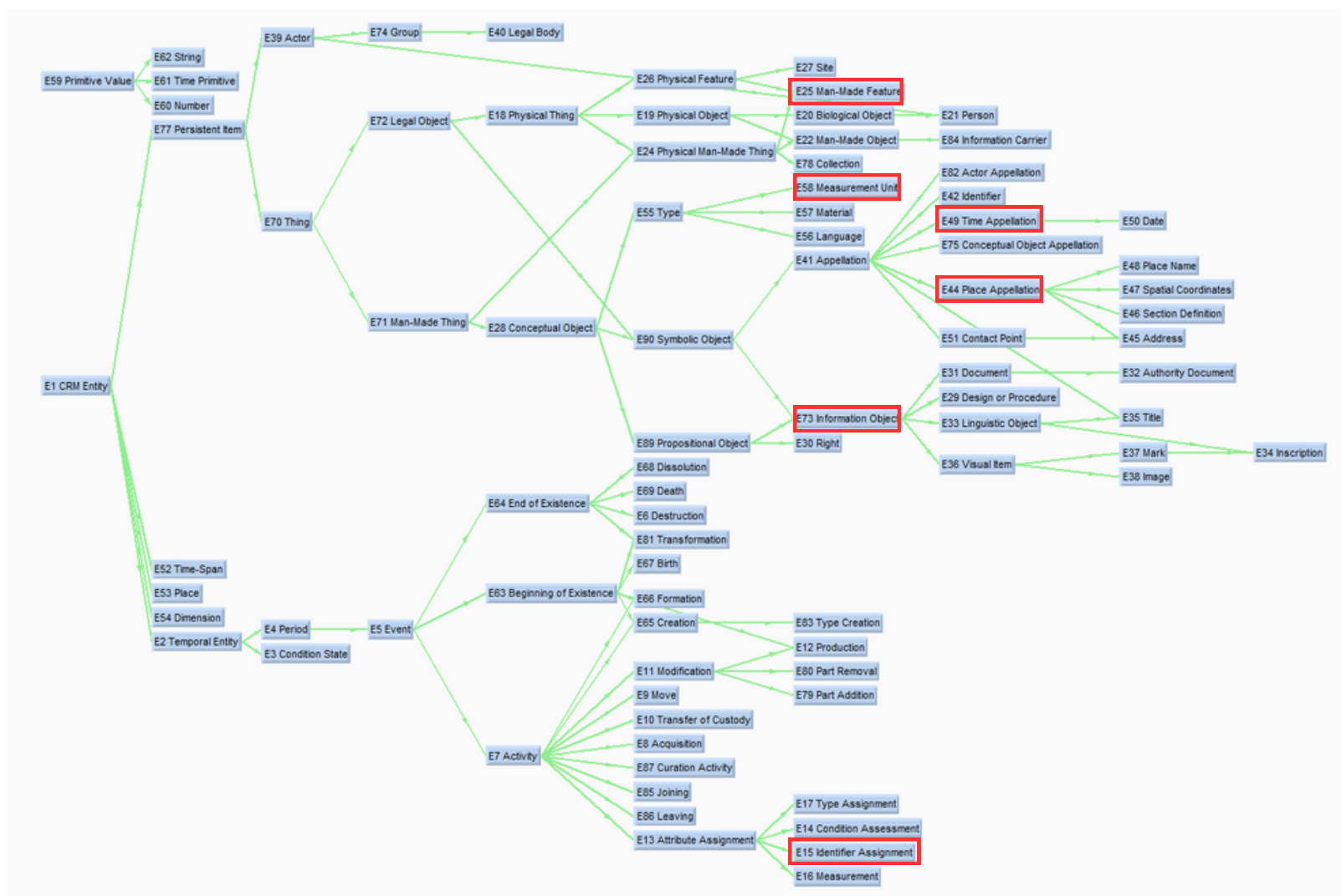
Semantic Web technologies, particularly the CIDOC Conceptual Reference Model (CIDOC CRM) ontology [3], offer a promising framework for representing and integrating heterogeneous data in a consistent and meaningful way. CIDOC CRM is designed to capture the complex relationships between cultural heritage objects, actors, and events, enabling researchers to map their data into a shared conceptual space and support remote collaboration. Moreover, CIDOC CRM supports the development of domain-specific extensions, enabling researchers to

ANNOTATION SEMI-AUTOMATIQUE

Perspectives

Sélection automatique des classes et propriétés

Fine-tuning



- Spécialisation du modèle sur La tâche d'extraction de connaissances
- Amélioration des performances
- Réduction de l'empreinte carbone

Merci