Geographic Disambiguation of Plant Entities in Historical Pharmacopoeias

Zoé PURSON-PASQUALINI

Supervisors : Agnès Braud and Florence Le Ber

Université de Strasbourg
M1 Informatique

Academic Year 2024-2025

Abstract

This report explores the problem of identifying plant species in vague or ambiguous references in historical Arabic pharmacopoeias. Building on previous work using named entity recognition (**NER**) and linking (**NED/NEL**), we integrate geographic criteria to enhance disambiguation. The project relies on data from historical texts, a graph database (**Neo4j**), and geographic biodiversity resources (**GBIF**), proposing a method based on occurrence densities within geographic radii around key medieval trade cities.

1 Introduction

A **pharmacopoeia** is a collection of standards and references on composition, and preparation of medicines. The Arabic ones from the medieval period are particularly rich but also pose significant challenges to digital analysis. Ambiguities in the **naming**, **translation**, and **historical context** often hinder the automatic interpretation of ingredients.

Among these, plant-based ingredients are the most numerous and varied. A single vernacular name (e.g. *thym*) can correspond to multiple modern taxa (e.g., *Thymus vulgaris*, *Thymus citriodorus*, etc.). Our goal is to resolve such ambiguities using external knowledge.

Prior work, based on Karim El Haff's thesis, has proposed a pipeline made up of several stages combining the use of Babelfy, Babel Net, Wikidata, and GBIF, with a first filtering based on a radius around the physician's city. This project extends El Haff's approach by integrating geohistorical data (ports, trade routes), modeling plant occurrence densities within 50 km radii around trade-accessible cities, and a Neo4j-based network representation coupled with interactive map visualizations.

This report presents the data sources used, the processing methods implemented, and the results obtained in this exploratory research project.

2 Related Work

In the biomedical and biodiversity domains, **Babel Net** and **GBIF** have emerged as key tools for linking vernacular plant names to structured taxonomic entities. El Haff (2023) proposed a pipeline combining **NER**, **Babelfy**, **Babel Net** linking, **Wikidata** alignment, and GBIF filtering based on geographical proximity to the physician's city. This provides a promising direction but still lacks consideration of trade networks and regional availability across extended routes.

Our work builds on this foundation by **introducing a graph-based representation of historical geography and commerce**. This model allows for geographic filtering that extends beyond fixed locations **to include trade-accessible cities and ports**.

The **disambiguation** of plant-based entities in historical pharmacopoeias is gaining renewed interest, particularly in the context of antibiotic resistance and the search for alternative remedies. Indeed, **correctly identifying the ingredients described in these texts is crucial** for unlocking potential therapeutic insights.

A landmark example is the discovery of artemisinin, a potent antimalarial compound derived from *Artemisia annua*, extracted through the reinterpretation of ancient Chinese remedies. In fact, during the Vietnam War, a single-celled parasite that causes malaria had become resistant to chloroquine, and with all the conditions of war and the humidity of Vietnam, it spread a lot. This breakthrough led to the 2015 Nobel Prize in Medicine being awarded to Tu YOUYOU, marking a major success in modern validation of traditional pharmacological knowledge.

This example illustrates how ancient pharmacological knowledge, once properly interpreted, can lead to major medical discoveries. However, one of the main obstacles lies in the ambiguity of historical texts, especially regarding plant names that vary across time, region, and language.

To address this challenge, recent computational methods have been developed to extract, normalize, and link named entities found in ancient texts. Named Entity Recognition (NER) and Named Entity Linking (NEL) techniques, often using multilingual semantic resources such as Babel Net and Wikidata, have enabled the association of vernacular plant names with scientific taxa.

3 Data and Tools

3.1 Data

For this project, I used multiple datasets.

Sabur's Pharmacopoeia: I used the annotated and translated version of the pharmacopoeia by Sabur Ibn Sahl, which contains over 36,000 tokens. As in the thesis by Karim El Haff, I focused exclusively on named entities related to plants. These entities were processed through a disambiguation pipeline inspired by the same work.

Atlas of Medieval Islamic Ports and Maritime Routes: I extracted data from this atlas, cleaning and structuring it for integration into the graph database. This dataset provides valuable information on maritime exchanges and port locations.

Medieval Trade Route Map: To provide historical geographical context relevant to Sabur's era (8th century), I utilized a map illustrating medieval trading routes. While a contemporary map from the 8th century is unavailable, the trade networks depicted in later medieval periods often built upon and maintained earlier established routes. Therefore, this map serves as a reasonable proxy to identify key geographical connections and accessible cities from Baghdad, Sabur's primary location. Based on these routes, I selected a subset of cities accessible from Baghdad and focused my geographic modeling around them.

3.2 Tools

To combine and enrich these datasets, my workflow integrated several interconnected tools: **Babelfy** for semantic linking and disambiguation, assigning a unique Babel ID to each ingredient term in the historical texts. These Babel IDs then enabled mapping to **Wikidata** entities via **Babel Net**, allowing us to extract **GBIF (Global Biodiversity Information Facility)** taxonomic identifiers and subsequently retrieve all species within those taxa. Geographic coordinates for ports and historical cities were retrieved using the **OpenStreetMap (Nominatim API)**, with manual corrections applied as needed. Finally, all the data was integrated into a **Neo4j** graph database using **Python** and **pandas** for preprocessing.

The following sections describe some of the tools in more detail.

3.2.1 GBIF

The Global Biodiversity Information Facility (GBIF) has been our main source of georeferenced data on species occurrences. It is an international network and data infrastructure funded by governments worldwide, providing open access information on all types of life on planet Earth. However, it's crucial to acknowledge that not all countries participate, including some of primary interest to this project. The closest participating countries to our region of focus are Armenia and Georgia. Consequently, the available data for non-participating countries tends to be less comprehensive compared to participating nations like France or the United States, as occurrences are primarily contributed by external entities. For instance, regarding Iraq, the majority of the contributed data originates from Great Britain and Northern Ireland, followed by Iraq itself. Iraq has 527 datasets comprising nearly 158,000 occurrence records, contributed by 39 different countries and zones. In contrast, the United States, a participating country, has over 5,000 datasets with more than 1 billion occurrence records, contributed by 67 different countries or zones regarding species found within the United States.

3.2.2 Neo4j Graph Database

Neo4j is a **graph-oriented database management system (NoSQL)** that is particularly well suited for storing and querying interconnected data by representing them as nodes, relationships, and properties. Neo4j was chosen to maintain consistency with El Haff's thesis. Furthermore, compared to traditional relational databases (SQL), even though they can also store related data, graph databases like Neo4j offer a more natural representation for the inherent network of relationships between plants, taxa, cities, and trade routes, which can be visualized and understood as a graph. For instance, a plant belongs to a taxon via a *IS* relationship. A detailed schema illustrating some of these relationships is presented here. Graph databases like Neo4j often exhibit superior performance for queries involving deep chains of connections. To illustrate this, consider the question: *"Which plants were available in Aden and how many of them were also traded via Muscat?"* The following **Cypher query** (Neo4j's query language) aims to answer a related question and its result is provided below:

```
MATCH (q:City {name: "Quilon"})<-[rq:WERE_LIKELY_FOUND_IN]-(t:TaxonName)
WHERE rq.score > 0
WITH collect(DISTINCT t.name) AS quilon_taxons

MATCH (a:City {name: "Aden"})<-[ra:WERE_LIKELY_FOUND_IN]-(t:TaxonName)
WHERE ra.score > 0
WITH quilon_taxons, collect(DISTINCT t.name) AS aden_taxons
WITH [x IN quilon_taxons WHERE x IN aden_taxons] AS common_taxons_count
RETURN size(common_taxons_count), common_taxons_count
```

4 Methodology

4.1 Reconstruction and Exploration of the Original Graph

My first task was to reconstruct the Neo4j graph database used in Karim El Haff's thesis, which links named plant entities to scientific taxonomies using Babel Net, Wikidata, and GBIF identifiers. This reconstruction was primarily based on **Karim El Haff's script** and the **CSV files provided at the start of the project** to populate his database. The process involved executing and readjusting the script; the initial execution failed on my computer due to performance limitations and the need to set up the Neo4j database. I also changed the names of ambiguous nodes such as **ScientificName** to **TaxonName** for better understanding, and created constraints for uniqueness in nodes. Successfully rebuilding the graph allowed me to gain a comprehensive understanding of its underlying structure and the relationships it encoded, specifically the connections between vernacular plant names and their corresponding scientific taxonomies.

These queries aimed to understand the nature and density of the connections between plant entities (vernacular names extracted from historical texts). The purpose of this exploration was twofold: firstly, to validate the integrity and completeness of the reconstructed graph, ensuring that the links established in the original thesis were correctly reproduced. Secondly, it allowed me to identify potential areas for improvement or expansion in my own work, such as understanding which types of plants or regions were well-represented and where the existing linkages might be sparse or absent, thus informing the direction of my subsequent research and the development of my own graph enhancements.

4.2 Cleaning and Integration of Maritime Trade Data

To incorporate geographic context, I turned to the *Atlas des ports et itinéraires maritimes de l'Islam médiéval*. This dataset describes nearly 500 historical ports and their associated trading exchanges located from Spain to the Arabian Peninsula.

The original data was embedded in a web interface backed by structured JSON files **found here**. I extracted the relevant content and converted it to CSV for ease of manipulation. A critical cleaning step followed, including:

- Removing inconsistencies such as faulty locations (for example, ports located in the Arctic and others in the middle of the sea) and normalizing port names,
- Resolving modern equivalents of medieval names using secondary resources such as **Wikipedia**, **Google Maps** and **historical websites**. **OpenStreetMap** was also used for geographical verification.
- Filtering ports based on relevant time periods (e.g., pre-Islamic to 9th century).

Following the cleaning step, I combined the information related to trade exchanges and port details. This process reduced the initial set of 497 ports to 33 relevant ports. Subsequently, I translated the pertinent columns (i.e., object, spice, food, and comment) into English, as the original data was in French. I then cross-referenced this new dataset with the plant **genre** present in my initial dataset from the six cities. This step aimed to identify the most relevant harbors for our study based on the traded goods. This process concluded with a dataset of 6 harbors. The names of these harbors and their distances from Baghdad are detailed in Table 4.

4.3 Selection of Key Cities for Geographical Anchoring

Due to the ambiguities in trade-based circulation of ingredients, I opted to use not only the physician's city (Baghdad) but also additional trade-accessible cities as geographic anchors. Based on a medieval trade map 4 and historical consistency, I selected six key cities:

Baghdad, Muscat, Aden, Hamadan, Mecca, Quilon

Each of these cities was selected for certain reasons as detailed in the array 3 and was geolocated using the OpenStreetMap (Nominatim) API, with manual verification for ambiguous or modernized names. Mecca's geolocation required manual verification due to the existence of an eponymous monastery in Italy.

4.4 Linkage to GBIF and Geographic Filtering

To disambiguate plant names, candidate taxonomic identifiers were first retrieved from Wikidata (property `P846`) based on their Babel Net linkages. For each taxon, GBIF identifiers were then obtained through targeted Wikidata SPARQL queries leveraging this same property. The `P846` statement in Wikidata is the designated field that directly links a Wikidata item to its corresponding GBIF (Global Biodiversity Information Facility) identifier. For instance, the Babel Net entry for 'sorrel' (`bn:00059966n`) yielded Wikidata item `Q157378`, from which GBIF ID `2891642` was extracted via the `P846` statement.

4.5 Scoring Method for Taxonomic Disambiguation

Following the methodology proposed by El Haff, I applied a scoring system based on the density of taxonomic occurrences retrieved from GBIF near major old towns, i.e. those selected above. It is important to acknowledge that the hypothesis that physicians were more likely to use plants naturally occurring or commonly available in their immediate surrounding regions is a simplification. Historical trade networks undoubtedly facilitated the availability of non-local plants. However, this assumption provides a basis for geographically informed disambiguation and allows for the testing of this proximity-based signal within the dataset.

In contrast to El Haff's model, which used three concentric radii (50 km, 500 km, and 1500 km), I opted to simplify this scheme by selecting a single fixed radius of 50 km around each city. This decision was motivated by two main reasons. Firstly, the selected anchor cities (e.g., Muscat, Aden, Mecca) already **span a wide geographical region around Baghdad** as described in Annex A, thus integrating some notion of extended availability through their positioning on historical trade networks. Secondly, the available data on modeled exchanges (routes and ports) were not sufficient to reconstruct actual trade chains. Therefore, implementing multiple nested radii based on this limited trade data would likely introduce more statistical noise than meaningful signal in the disambiguation of ambiguous entities.

For each taxon, city pair and harbor selected, I queried GBIF to retrieve all occurrences within a 50 km radius and computed the **density** as follows :

$$\rho_{t,r} = \frac{n_{t,r}}{\pi r^2}$$

where $n_{t,r}$ is the number of georeferenced occurrences of taxon t within the radius $r = 50\,\text{km}$. Each density was then normalized by the maximum observed density across all taxa for that radius:

$$S_{t,r} = \left(\frac{\rho_{t,r}}{\rho_{\max,r}} \right) \cdot C_r$$

with $C_r = 0.9$, the empirical coefficient from the original El Haff pipeline.

Finally, a small adjustment based on the **total number of occurrences** for each taxon across all cities was included to promote consistency. The adjusted score is defined as:

$$S_t^{\text{adj}} = S_{t,r} + 0.1 \cdot \left(\frac{n_t}{n_{\max}} \right)$$

where n_t is the total number of geo-referenced occurrences for the taxon across all cities, and $n_{\max}$ is the maximum value across all taxa. This allows favoring widely observed taxa while still preserving local specificity.

The final score therefore combines:

- **Local density** within the fixed 50 km radius;
- **Relative abundance (total number of occurrences)**, promoting taxa with stronger geospatial presence;
- **Geographical Relevance of Historically Significant Cities**, selected for their geographic.

4.6 Graph Enrichment and Exploration

4.6.1 Visualization

To make the results more interpretable, I generated interactive HTML maps using Folium, showing occurrences around each city with colored scoring zones and markers for plants. The map below shows the presence of the *Alhagi* taxon in different cities. The absence of a red-circle marker on a city indicates the absence of the plant or taxa within a 50 km radius.

Figure 1: Map representing the presence of the taxon *Alhagi* in the 6 different cities

We can see that the *Alhagi* taxon is only present within a 50-kilometer radius around Mecca, with a maximum score of 0.9 and a maximum density of 0.00013.

4.6.2 Graph Enrichment

With the addition of the ports and the six cities, I enriched the Neo4j graph with new nodes and relationships (visualized in Figure 2). Specifically: Each **City** and **Harbor** node was geolocated and linked to the most likely plant taxa via the **WERE_LIKELY_FOUND_IN** relationship (as detailed in Figure 2). **Harbor** nodes were also connected to individual **Ingredient** nodes using the **EXCHANGED_IN** relationship (as shown in Figure 2), allowing the graph to represent potential exchange routes of plant materials.

To ensure consistency across the graph, the unique **TaxonName** nodes created earlier are linked to a corresponding **Plant** node through the **IS** relationship. This structure supports the representation of multiple names referring to the same botanical entity.

Finally, by querying common **TaxonName** nodes shared between different cities or ports, I was able to identify overlapping plant knowledge or trade networks, offering valuable insights into regional medicinal practices and their interconnections.

Regarding the proposed inheritance for City and Harbor: While an inheritance structure could be considered, the current distinct representation allows for specific attributes unique to each entity type (e.g., geographical coordinates are crucial for both, but harbors might have specific trade volume data, while cities might have historical significance scores). However, this is a valid point for potential future refinements of the model.

Regarding the differentiation between Ingredient, TaxonName, and Plant:

- An **Ingredient** node represents a specific term for a substance (often plant-based, but also other materials like animal-derived *scorpion* or *fly*) as it appears in the historical texts. These terms are often vernacular names and can be ambiguous.
- A **TaxonName** node represents a normalized, scientific-like label for a group of plants. This is the result of the disambiguation process.
- A **Plant** node represents a broader botanical entity, potentially encompassing multiple **TaxonName** entries that are considered to refer to the same plant.

This distinction is crucial for navigating the different levels of plant identification within the graph, from historical mentions to standardized botanical entities.

The **IS** relationship connects a specific **TaxonName** to a broader **Plant** entity, indicating that the taxonomic name refers to that plant. The **WERE_LIKELY_FOUND_IN** relationship, on the other hand, connects a **TaxonName** to a **City** or **Harbor** based on the scoring method, indicating the probable presence of that taxonomic identification in the vicinity of that location. This probabilistic link accounts for the inherent ambiguity in historical texts.

Note that Figure 2 provides a simplified view of the graph structure. It highlights the key node types and relationships but does not include all the subtypes and secondary attributes implemented in the final database.

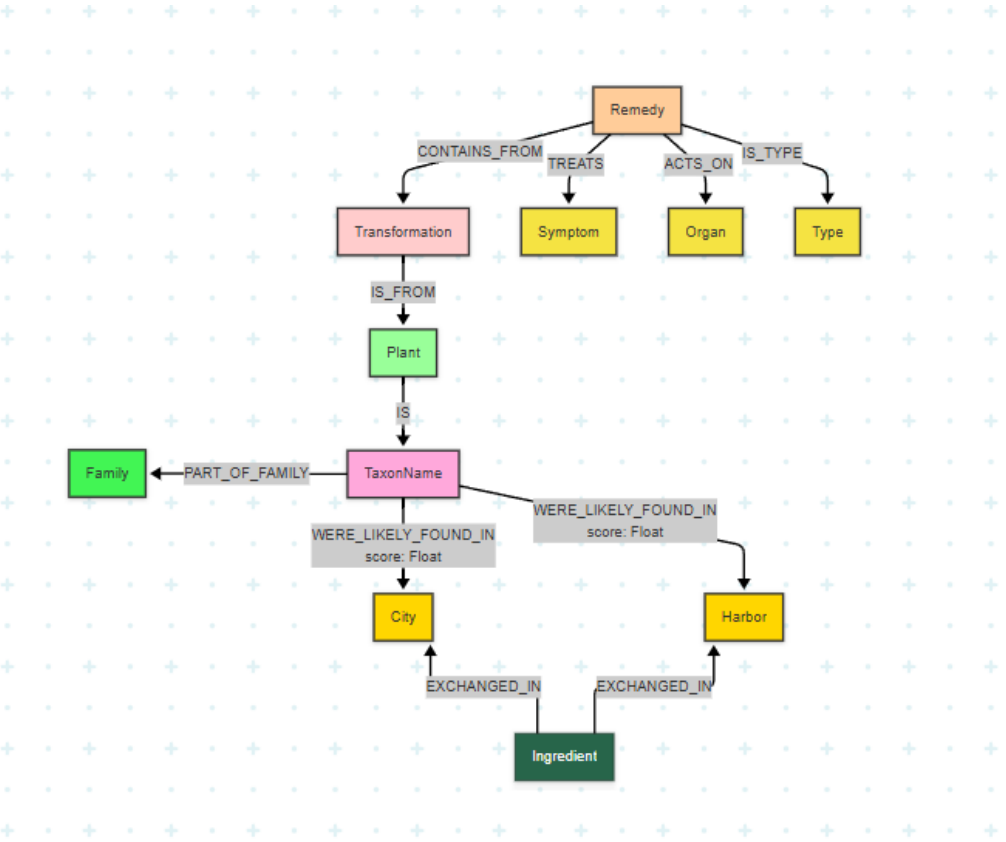

Figure 2: Simplified graph model schema of the Neo4j database, showing main node types, relationships, and key relationship properties such as `score`. This diagram does not include all nodes and properties implemented in the final model.

The quantities of each node and relationship type are summarized in Table 1.

5 Results and Analysis

The structured database built during this project includes several hundred entities, representing both botanical, geographical, and therapeutic information. The majority of nodes are related to plant entities, including generic plants and their subparts (such as leaves, bark, seeds, or fruit pulp), but also ingredients that do not belong in the plant category. Each plant or ingredient may be linked to multiple taxonomic names, which ensures compatibility with scientific nomenclature as well as the traditional or historical vocabulary found in ancient texts.

The integration of six historic towns and a selection of ancient ports, geolocated within the graph, enables the establishment of links to taxon names through the **WERE_LIKELY_FOUND_IN** relationship. This relationship is weighted by a composite metric called `final_score`, which combines local and global occurrence information for each taxon.

The score is computed by summing two components: a locally derived score based on the occurrence density of the taxon (normalized by the maximum observed density and adjusted by a coefficient), and a global ratio representing the overall frequency of the taxon across all studied cities. The formula is:

$$\texttt{final\_score} = \texttt{global\_score} + 0.1 \cdot \left(\frac{n_t}{n_{\max}} \right)$$

where: `global_score` is the maximum local density score $S_{t,r}$ observed for the given taxon across all cities/ports within its 50 km radius. This component captures the highest local geospatial presence for that specific taxon. n_t is the total number of georeferenced occurrences of the taxon across all cities, $n_{\max}$ is the highest total occurrence value among all taxa.

This `final_score`, ranging from 0 to 1, reflects the likelihood of a meaningful association, incorporating both the density of taxon occurrences within a 50 km radius of the locality and the taxon's overall abundance. Most scores cluster around a mean of 0.3, indicating a predominance of moderate associations—likely reflecting the inherent uncertainties in historical data and the complex nature of species distributions. Nonetheless, exceptionally high scores, such as 0.9 for *Alhagi* in Mecca Figure 1, point to a strong probability of local presence and, by extension, the potential for historical export to locations such as Baghdad.

The graph model also captures trade flows. Ingredients - whether botanical or not - are linked to ports through specific exchange relationships. This aspect of the model makes it possible to study not only the medicinal knowledge contained in remedies, but also the potential circulation of materials between regions that have been named in the pharmacopoeia.

This structure not only allows flexible queries over complex connections, but also paves the way for visual exploration, scoring comparisons, and even geographic clustering of remedies and ingredients.

Entity / Relationship Type	**Count**
Ingredient nodes	987
Remedy nodes	292
TaxonName nodes	395
City nodes	6
Harbor nodes	6
`WERE_LIKELY_FOUND_IN` relationships	1 215
`EXCHANGED_IN` relationships	7
Total nodes	2 565
Total relationships	6 576

Table 1: Summary of key node and relationship counts in the Neo4j graph

One particularly useful result of this structure is the ability to compare the botanical knowledge between cities. For instance, by comparing the sets of taxa most strongly associated with each city, we found that Quilon and Aden share 11 taxonomic entities with a non-zero association score. This was established using the **Cypher** query presented in the previous section. These common taxa include widely known ingredients such as *cyperus*, *aloe*, or *jasmine*, but also less expected terms such as *scorpion* or *fly*. This overlap suggests a significant convergence in botanical knowledge or trade between these two historical cities. Alternatively, it could also indicate that similar environmental conditions in these geographically separated regions might have led to the presence of the same plant species. A similar comparative approach was applied to all of

the other cities with Baghdad and is detailed in Annex C, where a list of shared taxa and the nature of the modeled link (land or sea) are provided. It is important to note that the presence of shared taxa does not definitively prove trade. As discussed earlier, similar environmental conditions, cultivation practices in both regions, or the widespread knowledge and use of certain plants could also explain these overlaps. The 'Link Between' column offers a potential, albeit indirect, indicator of possible trade routes, but further historical evidence would be needed to confirm such exchanges for specific taxa.

A similar comparative approach was also applied to the botanical knowledge between Baghdad and the selected ports. For instance, comparing Baghdad and the harbor of SHARMA yielded 43 shared taxa. These comparisons with the harbors are detailed in Annex D.

From a global perspective, the graph offers a networked view of the material and conceptual circulation of pharmacopoeias. Some cities act as hubs with a dense number of links to various taxa, while others appear more peripheral. Similarly, certain taxa are highly connected across multiple locations, indicating widespread use or significance, whereas others are very localized, appearing only once.

Furthermore, and importantly, the graph enables us to identify taxa present in various towns and ports but not found within a 50 km radius of Baghdad. This provides insights into the potential extent of trade at that time (though this remains a hypothetical exploration).

Harbor/City	Number of taxons not present near Baghdad	Distance to Baghdad
Muscat	9	1740 km
Aden	8	2274 km
Hamadan	37	434 km
Mecca	19	1391 km
Quilon	25	4268 km
SHARMÂ	35	3139 km
MÎNÂB	0	1397 km
AL-BALID	7	2053 km
RAS AL-HADD	1	1922 km
ATHTHAR	17	1802 km
AL-MAKHÂ'	0	2216 km

Table 2: Number of taxons present near cities or harbors but not near Baghdad with the distance from the cities/harbors to Baghdad

In summary, this table suggests that taxon composition varies considerably between the different cities/ports studied and the Baghdad region. Geographical distance does not appear to be the main factor determining these differences. This observation might be influenced by the fact that the GBIF data, while the best available, represents contemporary species occurrences and might not perfectly reflect historical distributions.

6 Conclusion and Perspectives

This project demonstrated the feasibility and value of integrating geohistorical and biodiversity data into a semantic graph to refine the disambiguation of plant ingredients in medieval Arabic pharmacopoeias. The reconstruction of the existing graph, enriched by the addition of data on ports and trade routes, and the application of a scoring method based on the density of geographical occurrence, facilitated the establishment of significant links and enabled the comparison of botanical knowledge between different towns and ports of the period.

Although the integration of GBIF data brought a crucial geographical dimension to the disambiguation, the limitations of the data coverage for the regions of the medieval Middle East presented a challenge. The inherent ambiguity of historical plant names and the intricate nature of reconstructing the period's trade networks required methodological simplifications. These aspects could be refined in future work.

Future prospects for this work include the exploration of more complex geographical scoring models, the integration of additional historical textual data to validate plant occurrences, and the extension of this approach to other corpora of medieval pharmacopoeia. Developments in interactive visualizations, such as those showing the potential flow of ingredients across trade routes and their correlation with medicinal knowledge in different regions, could also enrich our understanding of the links between botanical knowledge, medical practices, and the commercial dynamics of the period.

7 Appendix

Annex A : Distance of cities from Baghdad

City	Geospatial challenge	Distance to Baghdad
Baghdad	Sabur's city	0 km
Muscat	A city located in the Arabic Peninsula, particularly in Oman, the city is known since the 1st century as a leading port for trade between the west and the east	1740 km
Aden	A city located in Yemen. In the tab Histoire of the Atlas des ports et itinéraires maritimes de l'Islam médiéval, it's mentionned that a tax on the spices coming from India was applicated	2274 km
Hamadan	Located accross the mountain from Baghdad	434 km
Mecca	Located in Saudi Arabia, according to the map, the town was part of the route of an Arab road	1391 km
Quilon	Located in southern india, the city was accessible by sea from Baghdad	4268 km

Table 3: List of distances and reasons why the cities had been selected

Annex B : Distance of harbors from Baghdad

Harbor	Exchanges known	Distance to Baghdad
SHARMÂ	lapis lazuli	3139 km
MÎNÂB	indigo and turmeric	1397 km
AL-BALID	fish	2053 km
RAS AL-HADD	lapis lazuli	1922 km
ATHTHAR	unspecified$_{s}pices$	1802 km
AL-MAKHÂ'	sesamum	2216 km

Table 4: List of distances between selected harbors and Baghdad

Annex C – Shared Taxa Between Baghdad and Cities

City	Link Between	Shared Taxa
Aden	land	barley, palm-leaves, castor plant, scorpion, meat sparrow, date-palm, wheat, fly, trunk skink, wet coarsely barley, cyperus, Magnoliopsida, caper (total: 13)
Quilon	sea	palm-leaves, fresh pomegranate, pomegranate, fine sorrel, castor plant, pomegranate rind, one ha sorrel, scorpion, sour pomegranate, meat sparrow, sour pomegranate, wild pomegranate, persian pomegranate, apple- pomegranate, sesame, sweet sour sourish pomegranate, sweet pomegranate, frog, fly, trunk skink, cyperus, wild sorrel, liver dog bit, Magnoliopsida, fang rabid dog, sorrel, soft-rinded pomegranate, rind sour pomegranate, fresh sesame (total: 29)
Muscat	sea	palm-leaves, wild cucumber, fig, fresh pomegranate, pomegranate, castor plant, pomegranate rind, scorpion, sour pomegranate, meat sparrow, date-palm, sour pomegranate, wild pomegranate, persian pomegranate, apple-pomegranate, sweet sour sourish pomegranate, cucumber, sweet pomegranate, frog, fly, trunk skink, cyperus, liver dog bit, king fig, Magnoliopsida, fang rabid dog, soft-rinded pomegranate, cucumber, caper, rind sour pomegranate (total: 30)
Mecca	land	palm-leaves, fig, garden peppercress, fine sorrel, castor plant, one ha sorrel, scorpion, meat sparrow, date-palm, sesame, frog, fly, trunk skink, cyperus, wild sorrel, king fig, Magnoliopsida, sorrel, caper, babylonian garden peppercress, fresh sesame (total: 21)
Hamadan	land	barley, wild cucumber, garden peppercress, castor plant, meat sparrow, wheat, cucumber, fly, clover, wet coarsely barley, beet, liver dog bit, Magnoliopsida, fang rabid dog, cucumber, babylonian garden peppercress (total: 16)

Table 5: List of shared taxonomic names between selected cities and Baghdad, based on WERE_LIKELY_FOUND_IN relationships with non-zero scores.

Annex D – Shared Taxa Between Baghdad and Harbors

Harbor	Shared Taxa
SHARMÂ	barley, palm-leaves, wild cucumber, fig, francolin, garden peppercress, fresh pomegranate, pomegranate, fine sorrel, castor plant, pomegranate rind, one ha sorrel, scorpion, sour pomegranate, meat sparrow, sour pomegranate, wild pomegranate, persian pomegranate, apple-pomegranate, sesame, sweet sour sourish pomegranate, wheat, cucumber, sweet pomegranate, frog, fly, trunk skink, clover, wet coarsely barley, beet, cyperus, wild sorrel, liver dog bit, king fig, Magnoliopsida, fang rabid dog, willow, sorrel, soft-rinded pomegranate, cucumber, babylonian garden peppercress, rind sour pomegranate, fresh sesame (total: 43)
MÎNÂB	palm-leaves, francolin, castor plant, scorpion, meat sparrow, date-palm, frog, fly, trunk skink, beet, Magnoliopsida (total: 11)
AL-BALID	palm-leaves, wild cucumber, fig, fresh pomegranate, pomegranate, castor plant, pomegranate rind, scorpion, sour pomegranate, meat sparrow, date-palm, sour pomegranate, wild pomegranate, persian pomegranate, apple-pomegranate, sweet sour sourish pomegranate, wheat, cucumber, sweet pomegranate, frog, fly, trunk skink, cyperus, liver dog bit, king fig, Magnoliopsida, fang rabid dog, soft-rinded pomegranate, cucumber, caper, rind sour pomegranate (total: 31)
RAS AL-HADD	castor plant, scorpion, meat sparrow, frog, trunk skink, cyperus, Magnoliopsida (total: 7)
ATHTHAR	palm-leaves, fig, garden peppercress, fine sorrel, castor plant, one ha sorrel, scorpion, meat sparrow, date-palm, sesame, frog, fly, trunk skink, cyperus, wild sorrel, liver dog bit, king fig, Magnoliopsida, fang rabid dog, sorrel, caper, babylonian garden peppercress, fresh sesame (total: 23)
AL-MAKHÂ'	castor plant, fly, cyperus, Magnoliopsida (total: 4)

Table 6: List of shared taxonomic names between selected harbors and Baghdad, based on WERE_LIKELY_FOUND_IN relationships with non-zero scores.

References

1. K. El Haff, "Liage d'entités nommées issues de pharmacopées anciennes – une première approche," Master's thesis, Université de Strasbourg, 2023.

2. Vanessa Fokou, Karim El Haff, Agnès Braud, Xavier Dolques, Florence Le Ber, et al.. Exploring Old Arabic Remedies with Formal and Relational Concept Analysis. Concepts 2024, Cadiz, Spain, septembre 2024, Sep 2024, Cadiz, Spain. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-04622852`

3. APIM, "Atlas des ports et itinéraires maritimes de l'Islam médiéval," [Online]. Available: `https://apim.huma-num.fr/`

4. M. J. Månsson, "Map of Medieval Trade Routes," [Online]. Available: `https://easyzoom.com/imageaccess/ec482e04c2b240d4969c14156bb6836f`

5. The Nobel Prize | Women who changed science | Tu Youyou. (s. d.). `https://www.nobelprize.org/womenwhochangedscience/stories/tu-youyou`

6. J. Akoka, I. Comyn-Wattiau, and C. Du Mouza, "Conception de Bases de Données Prosopographiques en Histoire – Un État de l'Art," *Revue ouverte d'ingénierie des systèmes d'information*, vol. 1, no. 3, pp. 1–19, 2020. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-03937727`

7. J. Akoka, I. Comyn-Wattiau, S. Lamassé, and C. Du Mouza, "Conceptual Modeling of Prosopographic Databases Integrating Quality Dimensions," *Journal of Data Mining and Digital Humanities*, 2021. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-01966374v5`

8. J. Akoka, I. Comyn-Wattiau, S. Lamassé, and C. Du Mouza, "Contribution of Conceptual Modeling to Enhancing Historians' Intuition: Application to Prosopography," in *ER 2020: 39th International Conference on Conceptual Modeling*, Vienna, Austria, Nov. 2020, pp. 164–173. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-03023837`

9. V. Fokou, K. El Haff, A. Braud, X. Dolques, F. Le Ber, et al., "Exploring Old Arabic Remedies with Formal and Relational Concept Analysis," in *Concepts 2024*, Cadiz, Spain, Sep. 2024. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-04622852`