Concept Note: Arches Search AI Integration

LOCALIZED CHAT USER INTERFACE FOR NATURAL LANGUAGE QUERIES OF THE EAMENA ARCHES DATABASE VIA LARGE LANGUAGE MODEL TECHNOLOGY

Prepared For

School of Archeology
University of Oxford
1 South Parks Road, Oxford OX1 3TG
www.arch.ox.ac.uk

Prepared By

Scholium Technologies LLC
720 E 20th St. Apt. A
Oakland, CA USA 94606
scholiumtech.com | +1 (510) 934-9826

Concept overview

1. Background

The Arches platform is a powerful data management and discovery platform. Arches Search directly leverages the structure of the data model (aka the Resource Model) via its various data types, relationships, and concept values available to instances of the Resource Model. While this design is mechanically sound, the Search interface can seem esoteric to users unfamiliar with structures of the Resource models or patterns in the data. To motivate users to adopt Arches, an ideal entry-point would help familiarise them with the following:

- The tools available in search (search filters)
- The facets of search queries (nodes and data types)
- A survey of the data available in the database to build interest

Scholium Technologies propose a solution for such an entry-point to the Arches Search experience. Emerging technologies in the form of Large Language Models ("LLM") make agents available for general tasks on command. These agents excel at understanding written language and parsing, transforming, and mapping media to nearly any other format. With the combination of static extraction algorithms and a live, embedded LLM agent-powered chat bot, Scholium Technologies envision a user interface that connects a human-friendly query-input to the robust searching capabilities of Arches.

2. Proposed Objectives

Scholium Technologies’ objective is to increase engagement with and ease of use of EAMENA data by the general public and in particular:

- Remove ontology- and Arches-specific contextual knowledge as prerequisites to query the EAMENA database in Arches Search
- Ensure the augmented Search experience is seamlessly localised for Arabic-speaking users
- Provide an abstracted Natural Language query interface to use Arches Advanced Search filter

3. Implementation Overview

This project would create a connection between the Arches instance EAMENA and a Large Language Model (LLM) agent via Application Programming Interface (API). Simply put: the LLM agent would parse a Natural Language query, evaluate it for completeness, and transform it into the appropriate Arches Search Filter(s) based on the data types determined during the process. See Figure A:

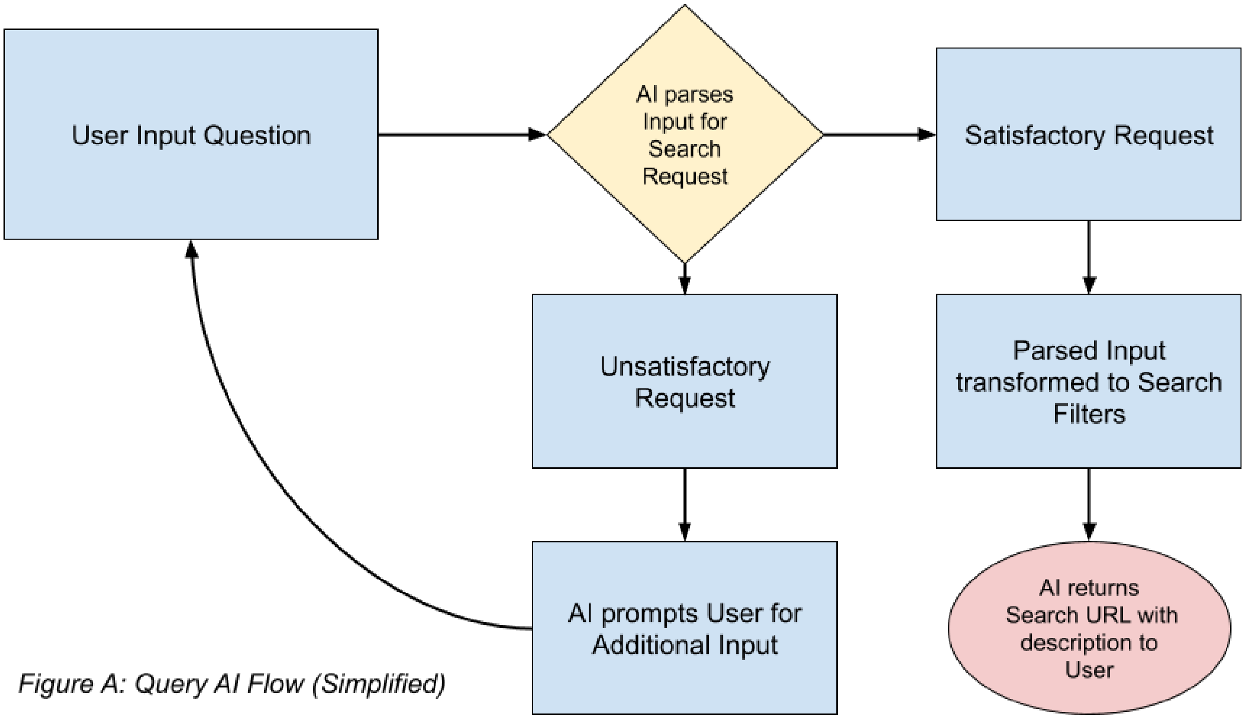

User Input Question
AI parses Input for Search Request
Satisfactory Request
Unsatisfactory Request
Parsed Input transformed to Search Filters
AI prompts User for Additional Input
AI returns Search URL with description to User

Figure A: Query AI Flow (Simplified)

Central Problem: Unstructured Data → Structured Data

The central problem this implementation solves is: *How can an unstructured Natural Language query be accurately and efficiently mapped to structured Arches Search Filters?* LLM agents excel at this particular problem because they can evaluate specific contexts to determine extraction definitions and processes on their own. However, offloading 100% of the work to LLM agents instead of static software represents an increased cost in both time and money, therefore a hybrid approach is optimal.

Arches Ontology Metadata

Arches resource instance data is already well-structured by data type, ontology entity type, and ontology relationship type. Each Arches Search Filter corresponds to specific data types. For LLM agents to process a Natural Language query in any context, an agent must parse it for *Intents*, *Entity types*, and *Prepositional/relational context*. Arches integrates into this extraction process exceptionally well because a resource model's native ontology metadata is already a kind of entity and relational classification scheme down to the node and node edge (relationship).

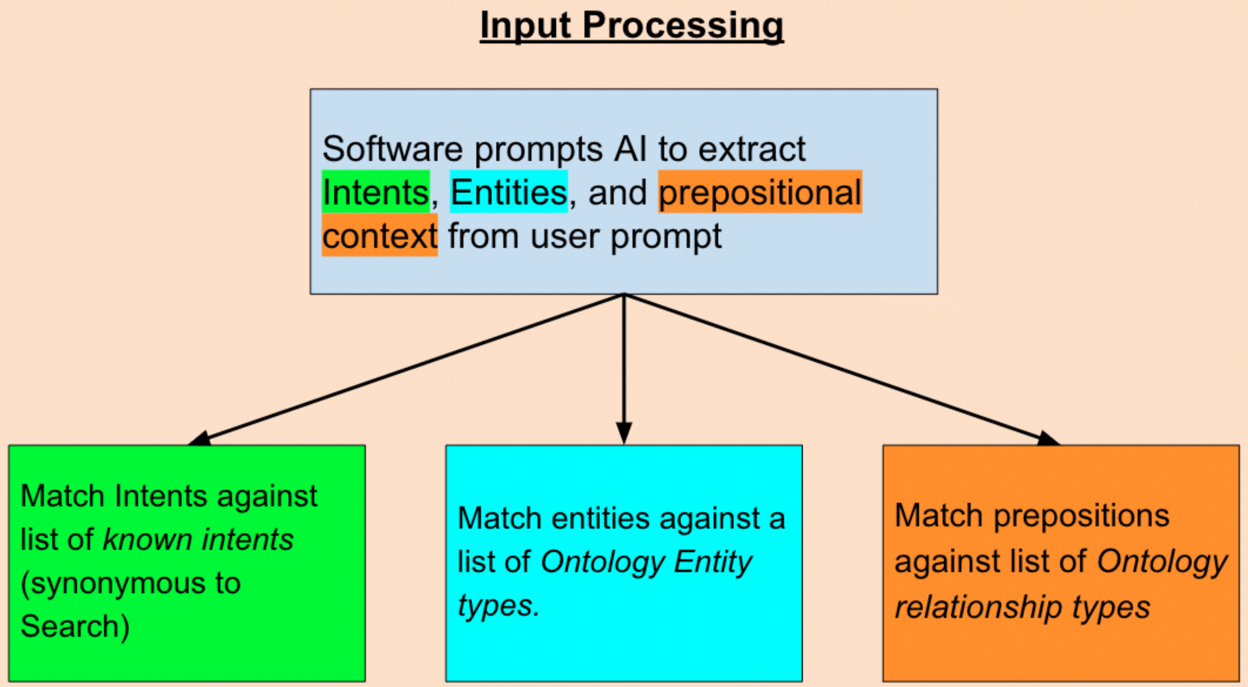

Input Processing
Software prompts AI to extract Intents, Entities, and prepositional context from user prompt
Match Intents against list of *known intents* (synonymous to Search)
Match entities against a list of *Ontology Entity types.*
Match prepositions against list of *Ontology relationship types*

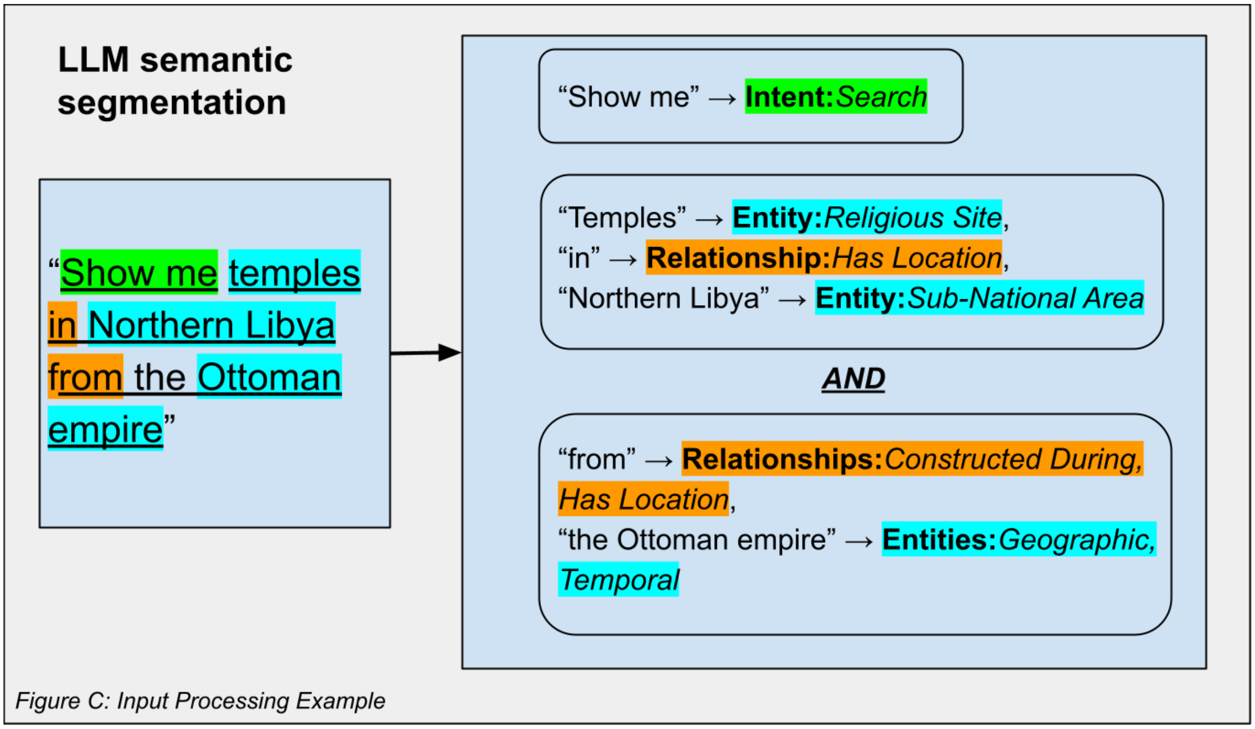
Figure B: Input Processing

The ontologies' metadata effectively map a parsed natural language query onto known nodes and their data types. The logic is roughly:

- Entity Types → Node Ontology Property
- Prepositional context → Node edge Ontology Property (aka Parent-Child relationship)

In cases where a search parameter does not correspond to data in the Arches database, the LLM agent can supply data for it on the fly. See the images below demonstrating an example sequence:

Example 1A: User queries “Show me temples in Northern Libya.” The AI agent extracts Search-oriented Intent, typed entities, and relationship context.



LLM semantic segmentation

“Show me temples in Northern Libya from the Ottoman empire”

“Show me” → **Intent:***Search*

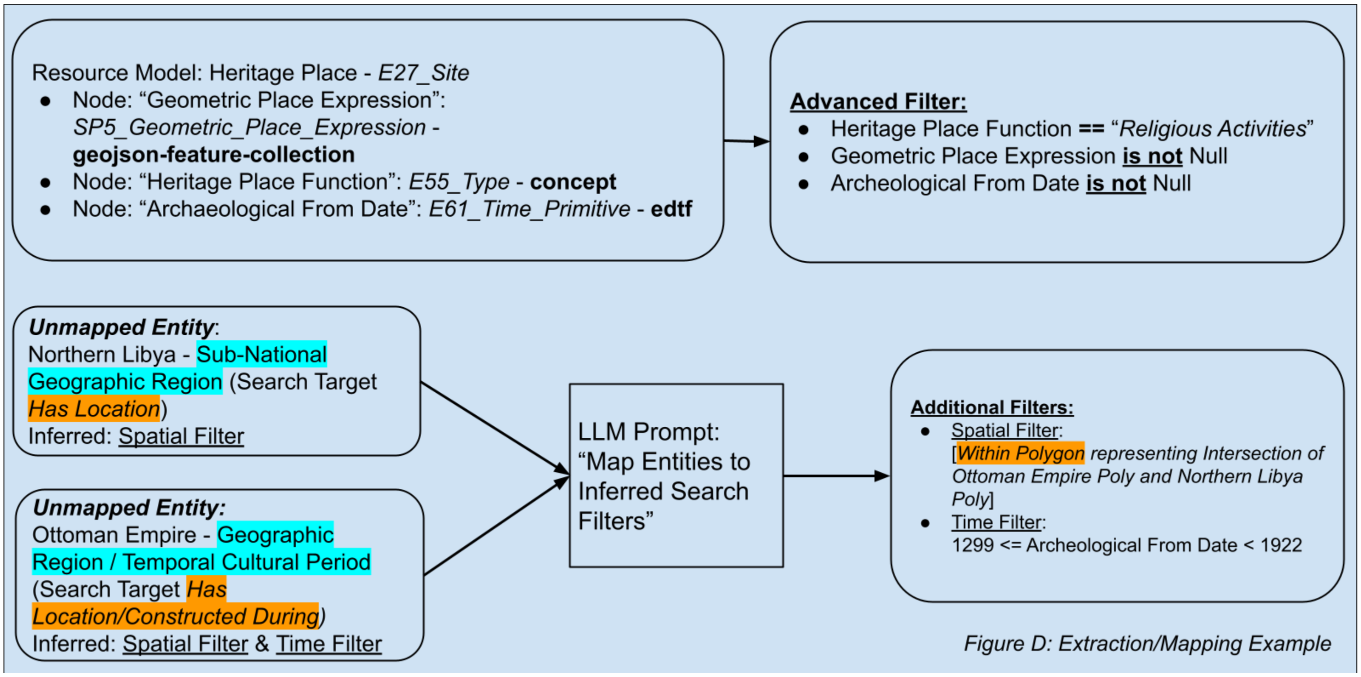
“Temples” → **Entity:***Religious Site*,
“in” → **Relationship:***Has Location*,
“Northern Libya” → **Entity:***Sub-National Area*

AND

“from” → **Relationships:***Constructed During, Has Location*,
“the Ottoman empire” → **Entities:***Geographic, Temporal*

Figure C: Input Processing Example

Example 1B: Software handles extracted entities with metadata, creates Advanced Filter for mapped node values and requests additional search parameters from LLM agent.



Resource Model: Heritage Place - *E27_Site*
- Node: “Geometric Place Expression”: *SP5_Geometric_Place_Expression* - **geojson-feature-collection**
- Node: “Heritage Place Function”: *E55_Type* - **concept**
- Node: “Archaeological From Date”: *E61_Time_Primitive* - **edtf**

Advanced Filter:
- Heritage Place Function **==** “*Religious Activities*”
- Geometric Place Expression **is not** Null
- Archeological From Date **is not** Null

Unmapped Entity:
Northern Libya - Sub-National Geographic Region (Search Target *Has Location*)
Inferred: Spatial Filter

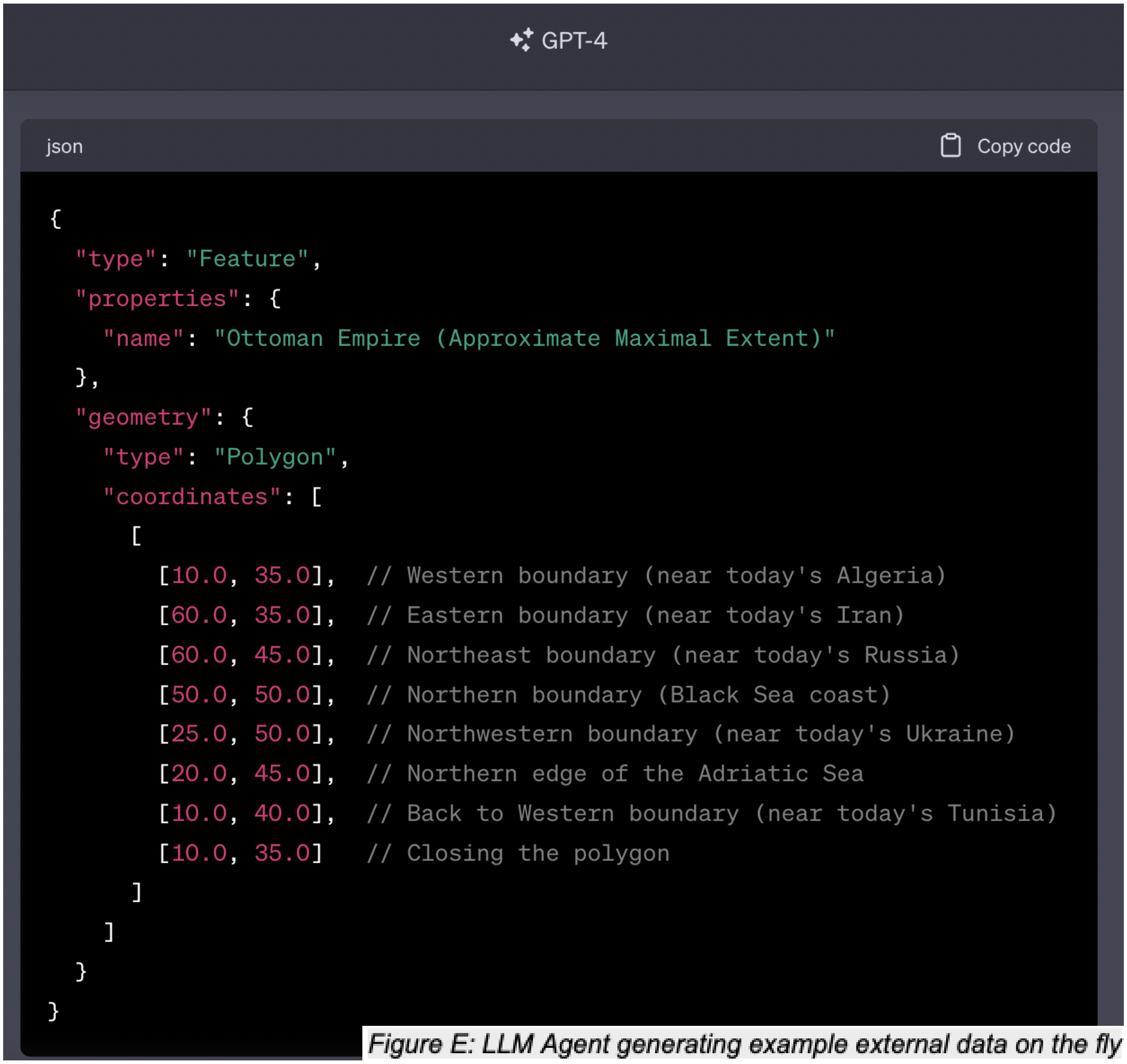
Unmapped Entity:
Ottoman Empire - Geographic Region / Temporal Cultural Period (Search Target *Has Location/Constructed During*)
Inferred: Spatial Filter & Time Filter

LLM Prompt: “Map Entities to Inferred Search Filters”

Additional Filters:
- Spatial Filter:
  [*Within Polygon representing Intersection of Ottoman Empire Poly and Northern Libya Poly*]
- Time Filter:
  1299 <= Archeological From Date < 1922

Figure D: Extraction/Mapping Example

Example 1C: LLM-supplies search parameter data for an unmapped entity (“Ottoman Empire - Geographic Region”) to be used in Spatial Search Filter.[1]



```json
{
  "type": "Feature",
  "properties": {
    "name": "Ottoman Empire (Approximate Maximal Extent)"
  },
  "geometry": {
    "type": "Polygon",
    "coordinates": [
      [
        [10.0, 35.0],  // Western boundary (near today's Algeria)
        [60.0, 35.0],  // Eastern boundary (near today's Iran)
        [60.0, 45.0],  // Northeast boundary (near today's Russia)
        [50.0, 50.0],  // Northern boundary (Black Sea coast)
        [25.0, 50.0],  // Northwestern boundary (near today's Ukraine)
        [20.0, 45.0],  // Northern edge of the Adriatic Sea
        [10.0, 40.0],  // Back to Western boundary (near today's Tunisia)
        [10.0, 35.0]   // Closing the polygon
      ]
    ]
  }
}
```

Figure E: LLM Agent generating example external data on the fly

[1] Figure E was taken as a screenshot from the OpenAI Chat-gpt user interface. On the backend, the LLM agent would send an identical response in json format to the software.

In a best-case scenario, the LLM agent would do only the initial parsing and mapping onto known Nodes (single-shot). In a different scenario, the LLM agent would follow up to map any unmapped entities (two-shot), especially in the case of data external to the EAMENA database like national boundary polygons. The backup strategies for the second scenario would include leveraging the ElasticSearch index for terms matching any unmapped entities as well as data type-specific guidance for the LLM agent in finalizing the mapping process.

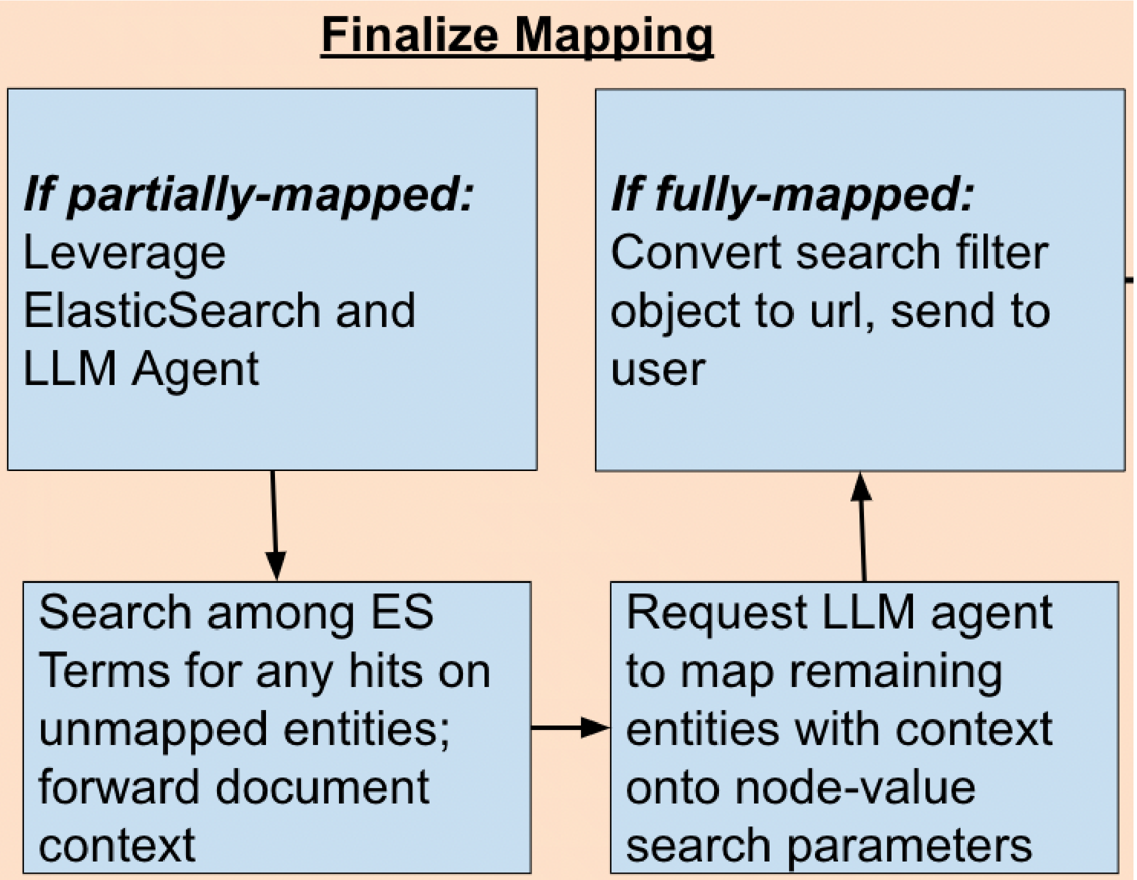

Finalize Mapping
If partially-mapped: Leverage ElasticSearch and LLM Agent
If fully-mapped: Convert search filter object to url, send to user
Search among ES Terms for any hits on unmapped entities; forward document context
Request LLM agent to map remaining entities with context onto node-value search parameters

Figure F: Follow-up Mapping Strategy

4. Budget overview

Labour Cost Estimate

Below is an overview of tasks and estimated level of effort in hours:

- Design, implementation of frontend User Interface: 8
- Design, implementation of Chat User Experience and state management: 16
- Prompt engineering and testing: 32
- Extraction and Mapping software processes, testing: 80
- Localisation (Arabic) and testing: 32

Scholium Technologies 2023 labour rates are a uniform rate of $140 per hour, £114.38 per hour[2]

Scholium Technologies estimate a calculated labour cost of 168 hours or £19,215.84.

Operating Cost Estimate

Below is generic example for cost projections:

- OpenAI[3] gpt-3.5 model per API call + response: $0.0035 or £0.0029
- Anticipated search API calls per search query: 1.75[4]
- Anticipated non-search API calls per search query: 0.5[5]
- Cost per 1000 search queries (including non-search API calls): £6.525

Accessibility and Controlling Costs

Because the search queries represent additional operating overhead, one implementation could look like rate-limiting chat-based search queries per IP address per date range for anonymous users. Otherwise, a user could apply to create an account with the EAMENA Arches app and have a higher rate-limit. An additional cost-reduction strategy would be to analyse recurring API calls in the case of requesting external data to complete the mapping process. Understanding common trends (e.g. specific national boundaries or cultural metadata) would indicate cost savings for localising such data in the EAMENA database.

[2] Exchange rate calculated October 9 2023, subject to change up to finalised contract

[3] Providers of LLM agents e.g. Google PaLM, Anthropic Claude, and others continue to expand offerings and affordability. OpenAI was used as an example for approximate pricing estimations.

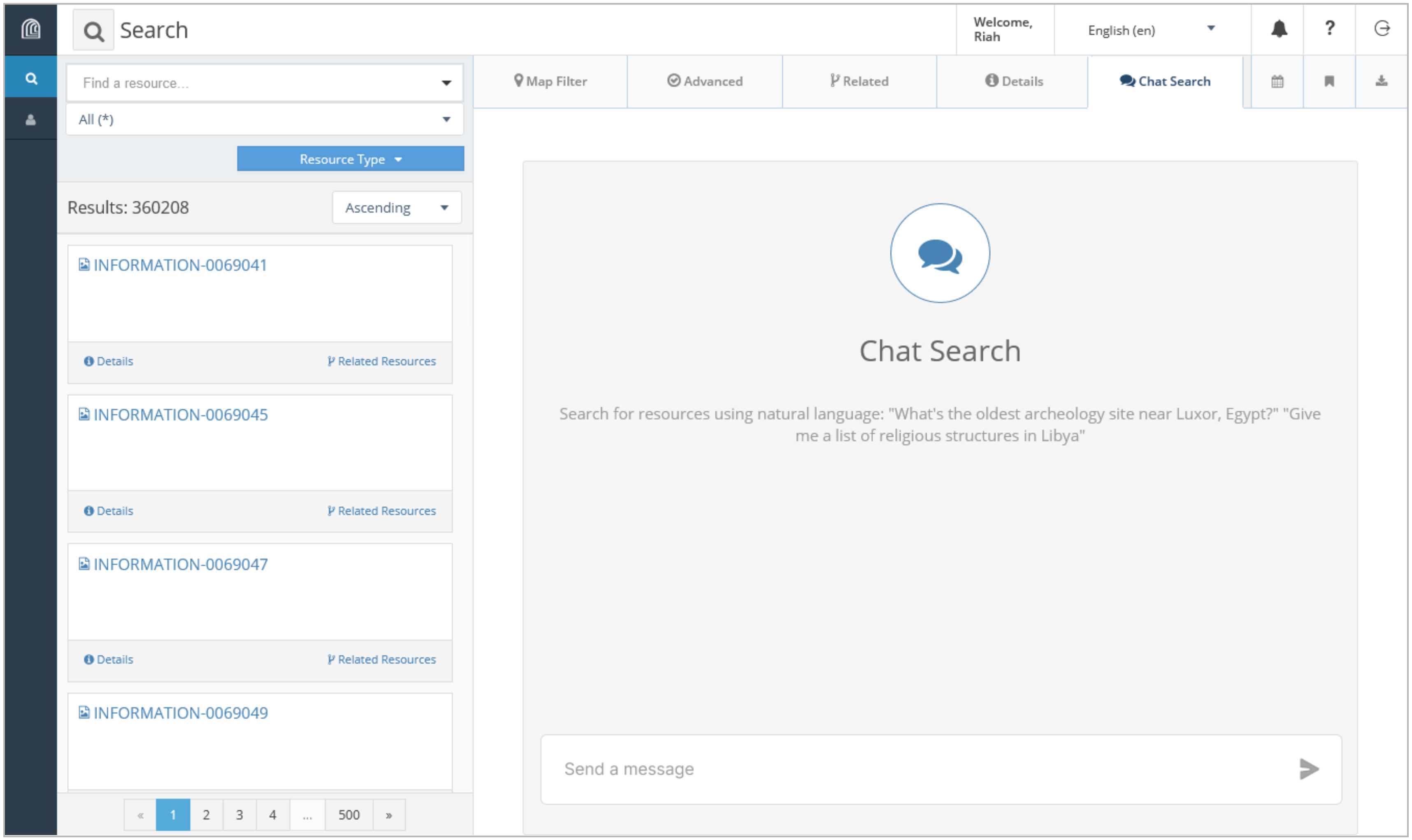
[4] If 1 out of 4 Natural Language queries can be fully mapped to search filters based on data local to the EAMENA Arches database (1 API call) and 3 out of 4 need additional search parameter data (2 API calls) the average for 4 search queries is (1(1) + 2(3)) / 4 = 7/4 = 1.75

[5] If 1 out of 4 Natural Language queries have 1 non-search query related to clarifying a set of results and 1 out of 4 fail to parse a “Search” Intent then 2/4 = 0.5

5. Proposed design

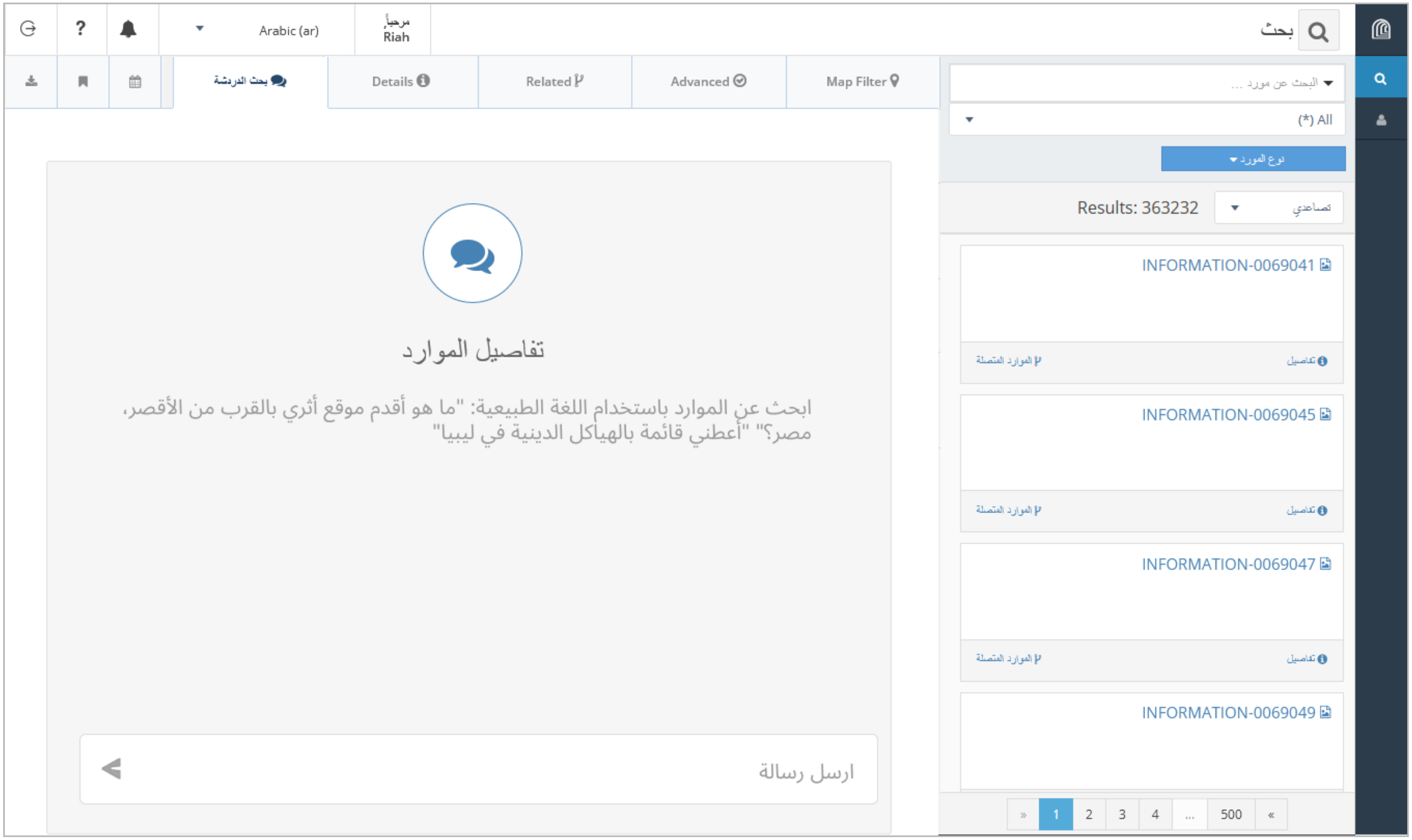
The exact details of the workflow are subject to technical feasibility, and may change slightly during implementation.

Clicking the Chat Search tab opens a chat UI with brief instructions and sample questions.



Search for resources using natural language: "What's the oldest archeology site near Luxor, Egypt?" "Give me a list of religious structures in Libya"

Arabic Example:



تفاصيل الموارد

ابحث عن الموارد باستخدام اللغة الطبيعية: "ما هو أقدم موقع أثري بالقرب من الأقصر،
مصر؟" "أعطني قائمة بالهياكل الدينية في ليبيا"

ارسل رسالة

Results: 363232

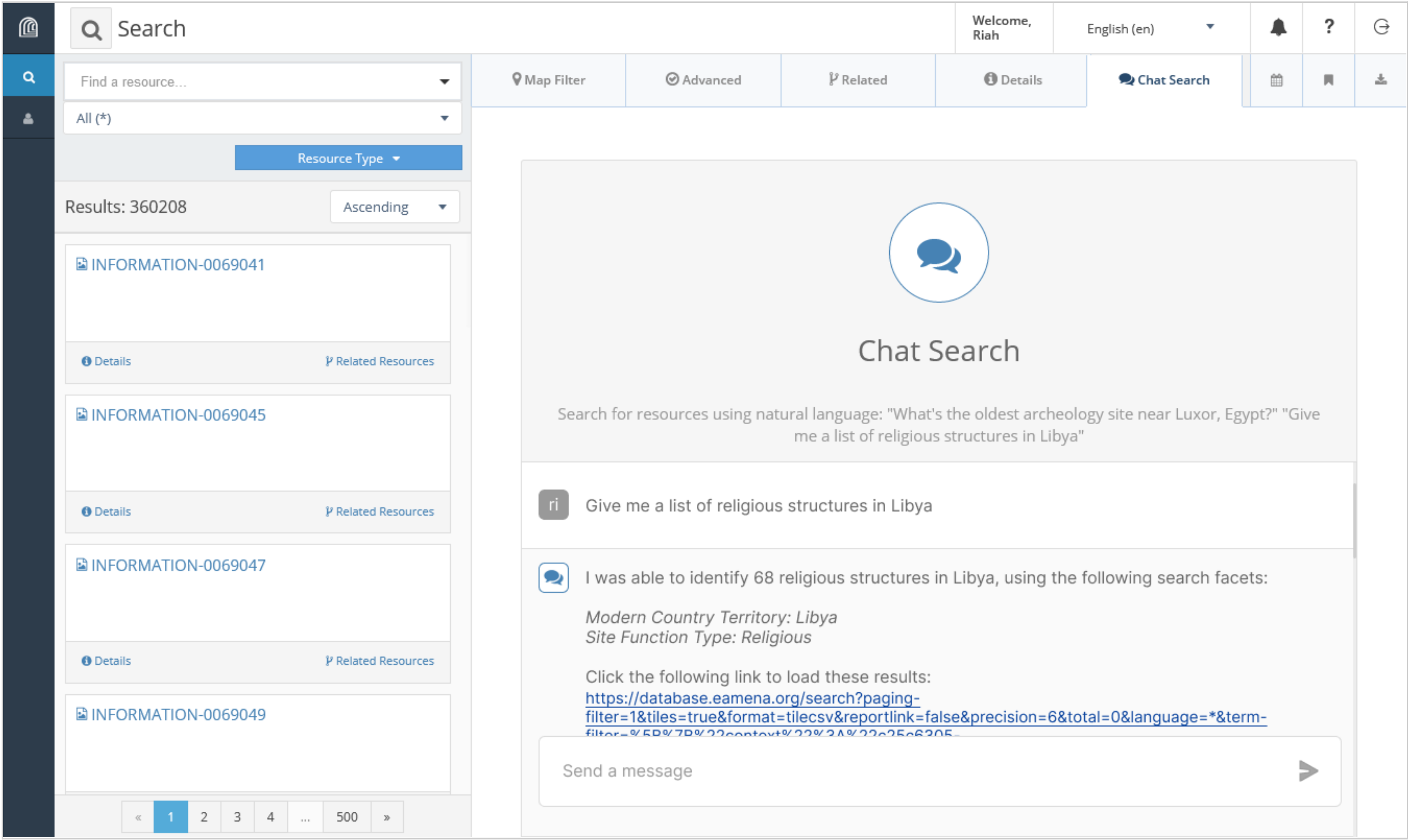
INFORMATION-0069041

INFORMATION-0069045

INFORMATION-0069047

INFORMATION-0069049

Submitting the query returns a response with a brief summary of search filters used and search URL. The user can respond for further clarification or an entirely new request.



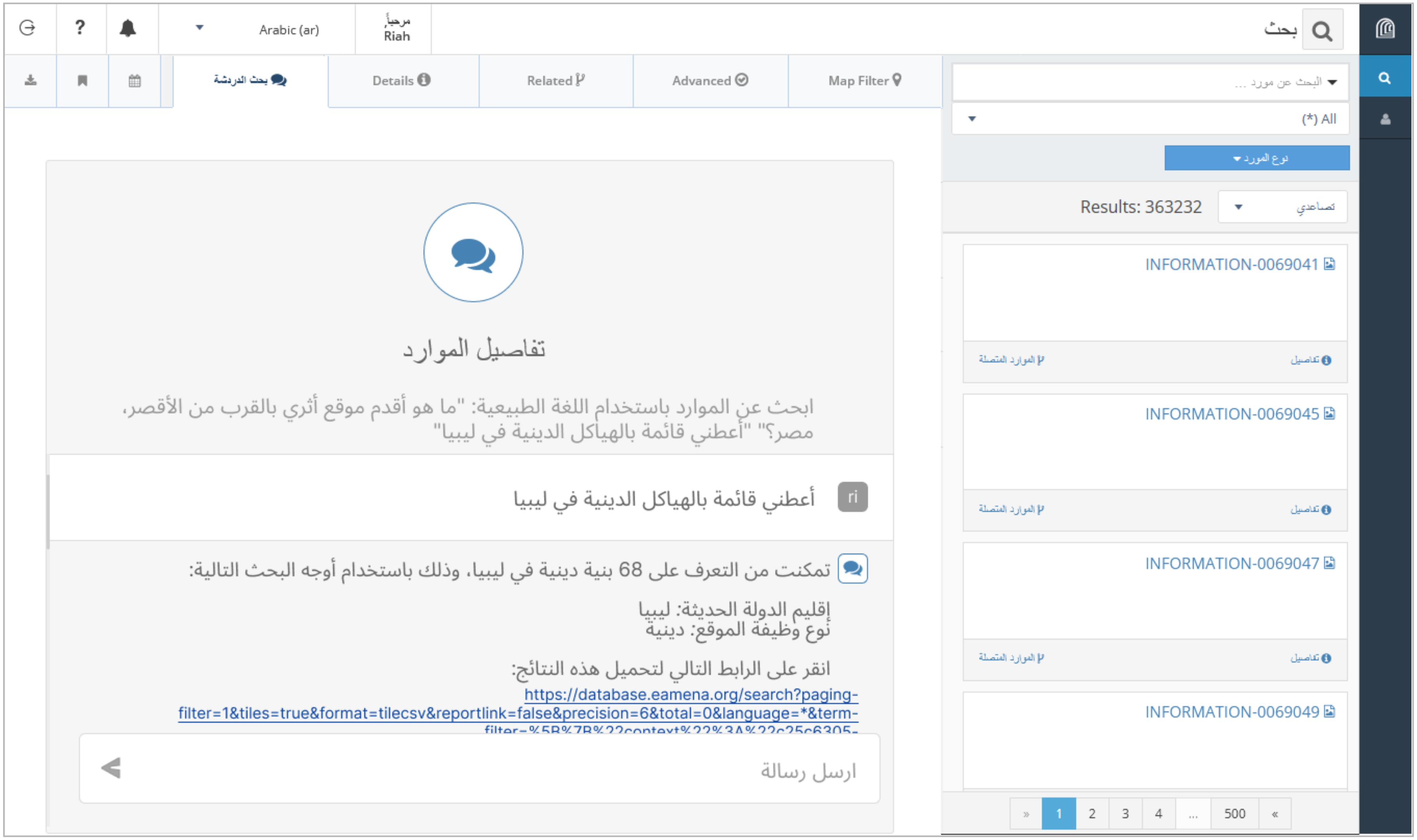
Give me a list of religious structures in Libya

I was able to identify 68 religious structures in Libya, using the following search facets:

Modern Country Territory: Libya
Site Function Type: Religious

Click the following link to load these results:
https://database.eamena.org/search?paging-filter=1&tiles=true&format=tilecsv&reportlink=false&precision=6&total=0&language=*&term-filter=%5B%7B%22context%22%3A%22c25c6305-

Arabic Example:



تفاصيل الموارد

ابحث عن الموارد باستخدام اللغة الطبيعية: "ما هو أقدم موقع أثري بالقرب من الأقصر،
مصر؟" "أعطني قائمة بالهياكل الدينية في ليبيا"

أعطني قائمة بالهياكل الدينية في ليبيا

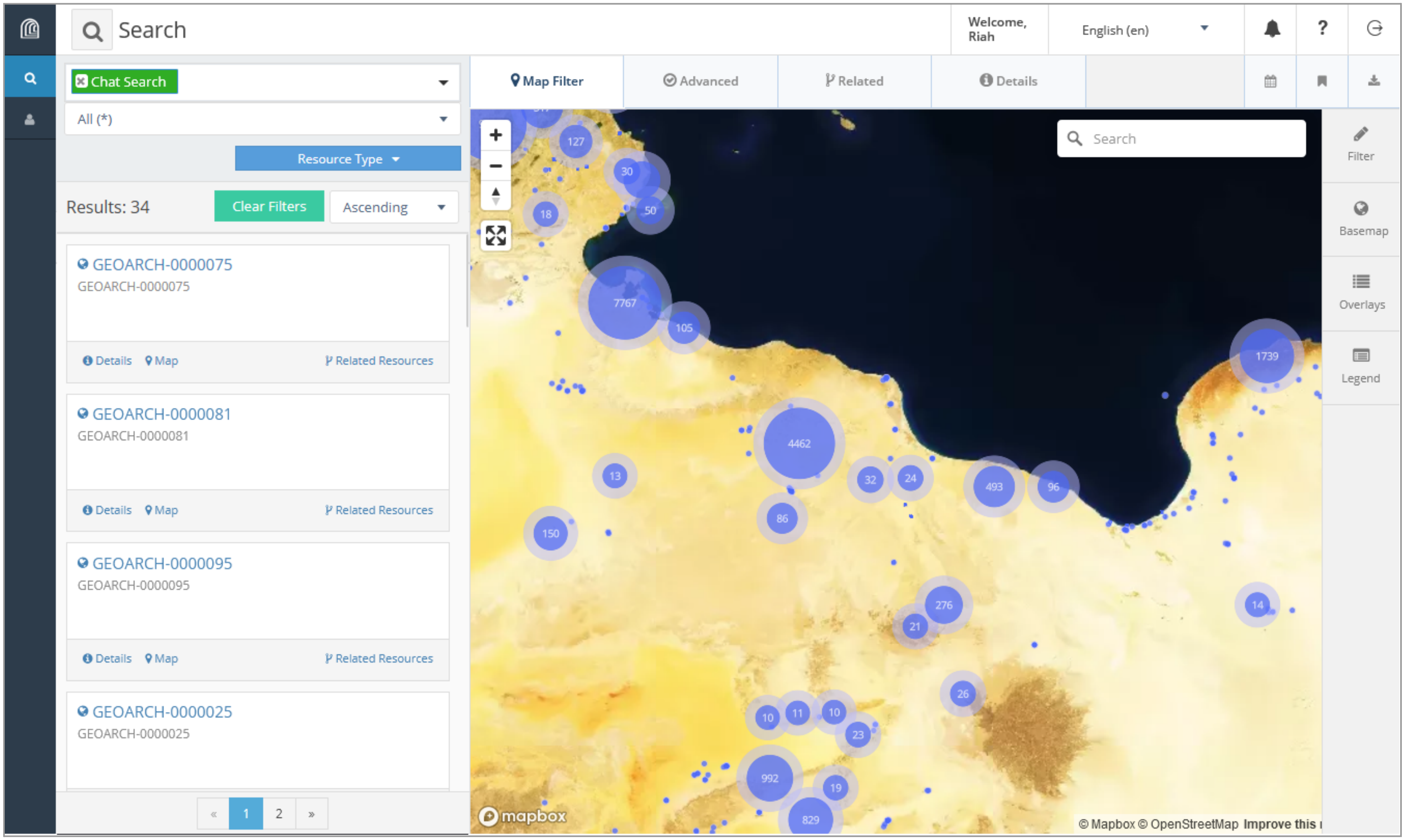
تمكنت من التعرف على 68 بنية دينية في ليبيا، وذلك باستخدام أوجه البحث التالية:

إقليم الدولة الحديثة: ليبيا
نوع وظيفة الموقع: دينية

انقر على الرابط التالي لتحميل هذه النتائج:

https://database.eamena.org/search?paging-filter=1&tiles=true&format=tilecsv&reportlink=false&precision=6&total=0&language=*&term-filter=%5B%7B%22context%22%3A%22c25c6305-

Clicking any link provided will load the map page and apply any of the query parameters



Search

Chat Search

All (*)

Results: 34

Ascending

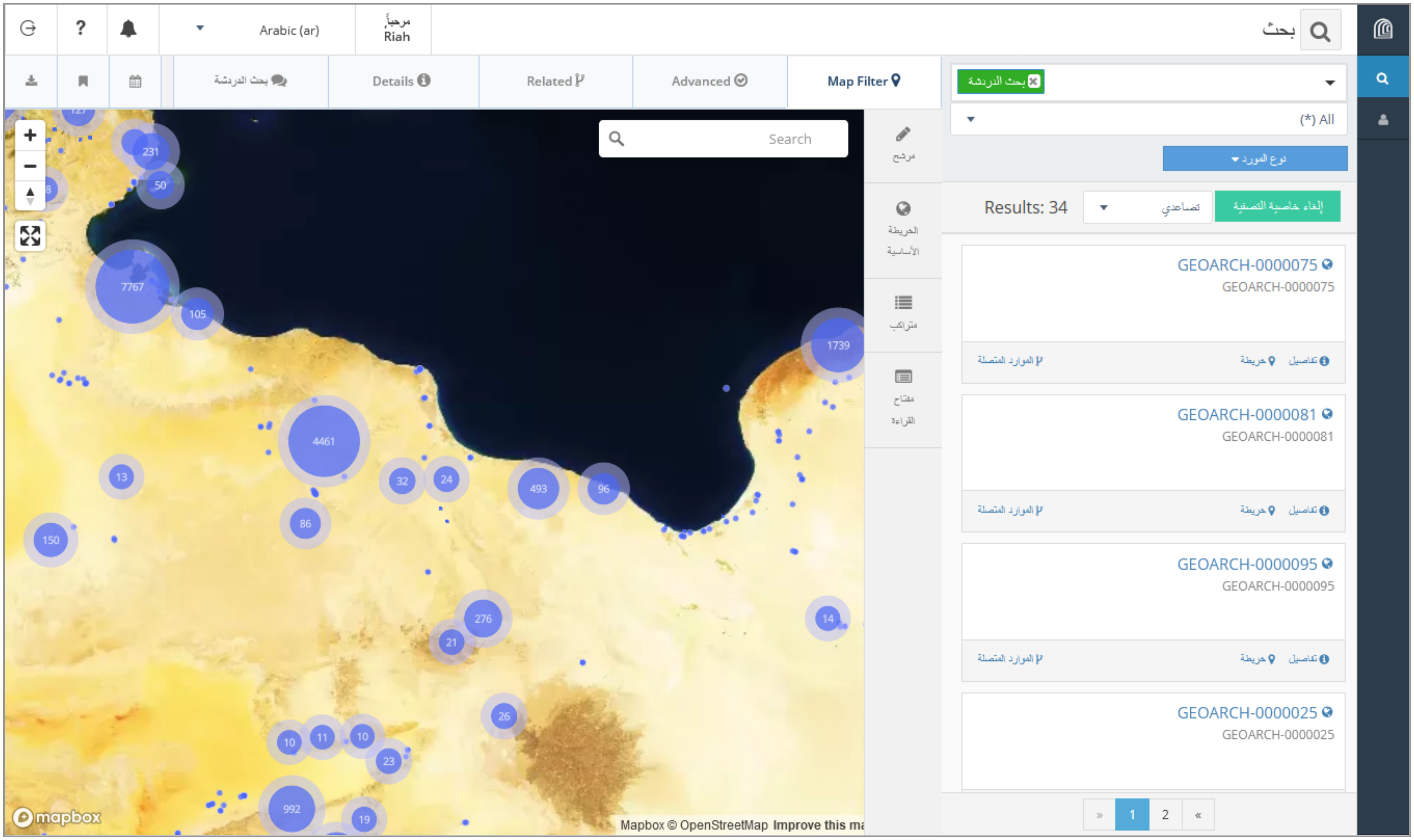
GEOARCH-0000075
GEOARCH-0000075

GEOARCH-0000081
GEOARCH-0000081

GEOARCH-0000095
GEOARCH-0000095

GEOARCH-0000025
GEOARCH-0000025

Arabic example:



Results: 34

GEOARCH-0000075
GEOARCH-0000075

GEOARCH-0000081
GEOARCH-0000081

GEOARCH-0000095
GEOARCH-0000095

GEOARCH-0000025
GEOARCH-0000025

END OF CONCEPT NOTE