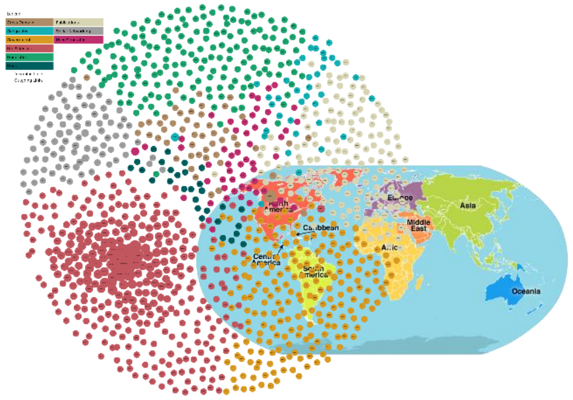
H.F.R.I.
Hellenic Foundation for
Research & Innovation

Description of the funded research project

1st Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers and Procure High-Value Research Equipment



Title of the research project: Scalable Answering of Questions Expressed in Natural Language over Large Geographic Knowledge Bases

Principal Investigator: Prof. Manolis Koubarakis

Reader-friendly title: GeoQA

Scientific Area: Mathematics and Information Sciences

Institution and Country: National and Kapodistrian University of Athens, Greece

Host Institution: National and Kapodistrian University of Athens

Collaborating Institution(s): Max Planck Institute for Informatics

Project webpage
(if applicable): http://geoqa.di.uoa.gr/

Budget: 200.000 €

Duration: 36 months

Research Project Synopsis

The main objective of the project GeoQA is twofold:

1. To show how to extend existing knowledge graphs with geographic knowledge found in important geospatial datasets available on the Web today.
2. To develop techniques and systems for answering complex non-factoid questions over such geo-knowledge graphs effectively (with high precision and recall) and efficiently (with very short response times).

In this project:

- We *create the knowledge graph YAGO2geo* by extending the well-known knowledge graph YAGO2 with data as found in multiple geospatial datasets.
- We develop *scalable techniques and software for keeping YAGO2geo up-to-date,* as some of the data sources used (e.g. OpenStreetMap) are highly dynamic.
- We generate a *gold standard corpus of geographic questions in natural language and their answers* which will contain more than 1000 questions.
- We develop the prototype question answering engine GeoQA2, based on natural language processing and knowledge graph embedding techniques, which will allow for intuitive visualization of the answers.

Project originality

- YAGO2geo will be the *richest knowledge graph* in terms of geospatial knowledge for the European Union and USA. It integrates data (thematic and spatial attributes)-as *linked data*- from: YAGO2; geographical administrative data provided by official sources of Greece, the United Kingdom, the Republic of Ireland and USA; the Global Administrative Areas dataset and OpenStreetMap.

- This is the first time that scalable knowledge graph update techniques are studied in the geospatial domain.

- The *corpus of questions and answers* that we develop, is the first one that is solely developed for the geospatial domain and it will be a point of reference for future research internationally.

- The *knowledge graph embedding techniques* that we develop encode *for the first time* geospatial information present in the entities, properties and literals of the knowledge graph to this extent (e.g. encodings of polygons).

- GeoQA2 will be the *first such engine internationally* to be able to answer *complex non-factoid geographic-based questions.* Very popular search engines, like Google, are struggling to give immediate answers to complex geographic questions like “What is the length of the river which crosses the city of Larissa in Greece?”.

- We will further optimize our engine by solving the multiple query optimization problem for the GeoSPARQL query language for which there is currently no research in this area.

Expected results & Research Project Impact

Expected results:

1. The richest knowledge graph on geospatial data.
2. An automatic knowledge graph update system for large knowledge graphs.
3. A gold standard corpus of geographic questions in natural language and their answers.
4. Knowledge Graph Embedding techniques for geospatial data.
5. An effective and efficient question answering system for geospatial question in natural language.

The *scientific impact* will be significant because:

- Currently there is no knowledge graph with a significant body of geospatial knowledge. YAGO2geo will be the first one for this purpose.
- There is only one question answering system (GeoQA) in this domain, which has been developed by this group and forms the starting point of this project.

When developed, the YAGO2geo, the corpus of geospatial questions/answers and the GeoQA2 engine will become points of reference in this area.

The *economic impact* will also be significant Greece and Europe because:

- M.Sc. and Ph.D. students will be supported during their studies.
- The availability of GeoQA will allow the team of Prof. Koubarakis to develop intelligent assistants specialized in geospatial knowledge that can be exploited commercially by a start-up of the University of Athens.
- YAGO2geo and GeoQA2 can also be exploited by any European search engine to develop domain specific functionality (e.g., for tourism)
- It will be possible to customize the GeoQA2 engine for specific applications, such as satellite image datasets. Satellite data is probably the most important digital resource available to mankind today (e.g., mitigation of climate change effects).
- Therefore, the results of the GeoQA project will have a very high financial impact in Greece and Europe in the future.

The importance of this funding

This project provides the chance to the PI and his team to develop new results and applications to this research area. Also, it gives the chance to new researchers to occupy themselves in this country and strengthen their curriculum vitae.

H.F.R.I.
Hellenic Foundation for
Research & Innovation

COMMUNICATION

185 Syggrou Ave. & 2 Sardeon St. 2
171 21, N. Smyrni, Greece
+30 210 64 12 410, 420
communication@elidek.gr
www.elidek.gr