Geo.IA: Artificial Geo-Intelligence Platform to Solve Citizens Problems and Facilitate Strategic Decision Making in the Public Administration

Andrés Montoyo, Rafael Muñoz and Yoan Gutiérrez

Departament of Software and Computing Systems. University of Alicante, Spain. Crta. San Vicentte del Raspeig s/n, Alicante, Spain

Abstract

The objective of Geo-IA is to research, design and implement a Geo-Smart Artificial Intelligence technology platform for public and private business organizations. The GeoIA project presents a geolocation platform that integrates technological innovation to support a strategy for the creation of a Smart Territories. To do this, Text Mining, Machine Learning (including deep learning) and Natural Language Processing technologies are deployed. The functionality of the geolocation platform is to analyze, integrate, share data, visualize and represent territorial indicators, with the aim of facilitating the monitoring and fulfillment of territorial strategies. In short, GeoIA promotes interoperability between public administration bodies and also provides citizens with mechanisms to access information of interest, where the magnitude of the integrated and interrelated data permits. GeoIA also provides digital knowledge (tools, linked information, semantics, virtual assistants) for use by public administrations to enhance their decision making through greater knowledge of the environment and to improve services to citizens.

Keywords

Ontologies, semantics, semantic document profile, entity recognition, knowledge discovery, machine learning,

1. Introduction

A smart society uses advances in information and communication technologies (ICT) to promote sustainable improvements that better the lives of citizens. A "smart" environment provides people with advanced solutions to solve problems or apply innovative technologies to create efficient and adaptable services, connected cities and communities, informed, participative and satisfied citizens and, above all, intelligent solutions for the provision of services. Thus, in order to achieve these objectives and respond to the needs of citizens, 21st-century public administration must deepen the digitization and improvement of its services by promoting digital transformation as an organizational system or ecosystem in which all stakeholders (citizens, municipalities and companies) must develop at the same speed. However, the public sector adapts to change less rapidly than the private sector. It is therefore a worthwhile endeavour to provide digital knowledge (tools, linked information, semantics, virtual assistants) to public administration bodies so that they can improve services to citizens and thereby enhance public administration decision making through greater knowledge of the environment

With the above goal in mind, this project will apply a discipline based on continuous evolution that is central to Artificial Intelligence (AI), and grounded in three basic pillars: (i) data and text mining that will integrate and extract information from different heterogeneous data sources to transform it into an understandable structure for further use, (ii) a prediction system, to make the best decisions of future actions in public administration, as well as citizens' behaviors for decision making based on machine learning (ML) and, (iii) a simplified text generation system to prescribe citizens' needs based on natural language processing (NLP). The starting hypotheses to address this project will be the following:

- (H1) Ontologies, the backbone technology that supports Linked Data and the semantic web, provide a means to overcome some of these challenges and thus help to obtain accurate profiles to produce satisfactory recommendations and personalized services.
- (H2) An ontology-centric knowledge base allows the integration of data from heterogeneous sources and enables their analysis through advanced reasoning and inference processes.
- (H3) Natural Language Processing, especially focused on entity-oriented machine learning, is essential to automate the construction and population of these knowledge bases, as well as to automatically generate simplified and synthesized text.

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

montoyo@dlsi.ua.es (A. Montoyo); rafael@dlsi.ua.es (R. Muñoz); ygutierrez@dlsi.ua.es (Y. Gutiérrez)

0000-0002-3076-0890 (A. Montoyo); 0000-0001-8127-9012 (R. Muñoz); 0000-0002-4052-7427 (Y. Gutiérrez)

- (H4) Knowledge-based Artificial GeoIntelligence applied to public sector sources such as OpenData, GIS, Meteorology and Social Media data will be essential to generate predictive and prescriptive models to improve decision making processes in Public Administration.

1.1. Project objectives

The main objective of the project is the research, design and implementation of a GeoIntelligence Artificial Intelligence (GeoIA) technology platform for public and private business organizations. Specifically, the platform will carry out the following activities:

- The design and development of a flexible, scalable and robust technological architecture for data integrating from different sectors of the public administration, such as Geographic Information Systems (GIS), cadastral data, census data, infrastructure data, consumption data, and endless sector data.
- The application and optimization of machine learning algorithms to perform predictive and prescriptive analytics to predict events or make sound strategic decisions.
- The development of data visualization technologies to simplify communication with citizens by offering information in a simplified and synthesized form so that it can be easily understood.
- The design and execution of a pilot test to validate the technologies developed in a key area such as public administration.

Thus, it is essential for public and private sector organizations to develop a data and text integration model capable of defining and creating semantic networks of geolocalized digital entities, understanding digital entity as any concept that has associated information that characterizes it (a company, person, city, building, organization, etc.). These tools will allow the detection and extraction of semantic relationships between digital entities, obtaining the information from different types of sources (unstructured, structured and linked open data), as well as determining the quality, consistency and veracity of these relationships. In addition, from the previously created knowledge bases, machine learning techniques and algorithms will be adapted to work with spatio-temporal data to define GeoIA knowledge-based models.

2. State of the art

The participating team behind this proposal has been involved in several national and international research projects, in which digital entities and their contextual language have been modeled through their identification, characterization, representation and exploitation. The participating team has also worked in projects related to knowledge generation and bias in language modeling, both technologies of great relevance for the present project proposal. Specifically, they are the REDES[1] project (TIN2015-65136-C2-1-R, TIN2015-65136-C2-2-R) and LIVING-LANG[2] (RTI2018-094653-B-C22). The goal of REDES was to go a step beyond the representation of Digital Entities, enhancing the development of technologies to automatically discover Digital Entities from different heterogeneous sources to populate semantic structures and link them to shared data. This process involved defining Digital Entities for different domains and processing heterogeneous information from the web, the social web and the web of data. The digital entities were then semantically enriched and spatiotemporally controlled and, finally, the generated information was integrated into the digital entity model. REDES project was the starting point to generate the following publications in indexed journals by the research team also involved in GeoIA: [1], [2], [3], [4] and [5].

One of the main contributions of LIVING-LANG, which is relevant to the present project, is the characterization of digital entities through the use of human language in key dimensions for the social contextualization of these entities. These multidimensional features allow relating entities from different perspectives, improving the understanding of the exchanged content and creating new knowledge in the analysis of these related structures. LIVING-LANG project generates these publications indexed in journals of which the author is part of the GeoIA project team: [6], [7], [8] and [9].

Figure 1 provides a summary of how the above projects have led to a number of Artificial Intelligence technologies, in particular Language Technologies, which have evolved incrementally.

3. Proposed GeoIA System

We propose the design and development of an Artificial GeoIntelligence system that focuses on the extraction of knowledge from heterogeneous sources and is capable of integrating different types of data, i.e. sectors of public administration, such as Geographic Information Systems (GIS), cadastral data, census data, infrastructure data, consumption data, and a host of sector data. This type of system has a direct impact on eco-digital transition policies, in addition to being able to offer a direct service to the population through the use of virtual assistants, e.g. chatbots.

[1] https://gplsi.dlsi.ua.es/proyectos/redes/
[2] https://gplsi.dlsi.ua.es/proyectos/livinglang/

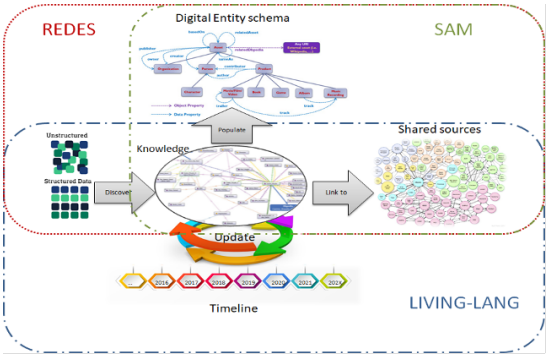

Figure 1: Knowledge discovery workflow based on the integration of technologies developed in the projects: REDES and LIVING-LANG.

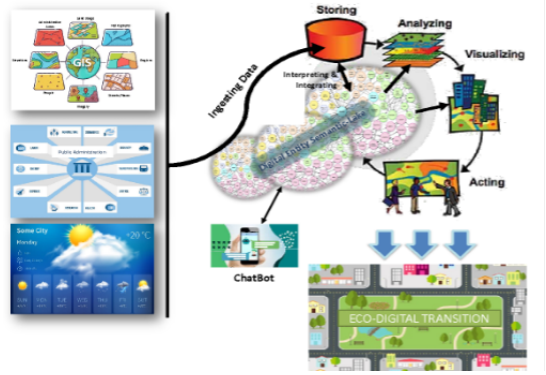

Figure 2: GeoIA solution in the context of digital transition.

Figure 2 is a macro view of how GeoIA's knowledge-based solutions are presented to support the digital transition towards optimizing operations in organizations by incorporating digital technologies.

The GeoIA project, as an open and accessible data storage and processing center, will allow a controlled collaborative model between the different agents for the achievement of the project objectives, based on obtaining knowledge through the integration of the data collected and subsequently analyzed. Addressing each of the different phases requires an expert consortium both in the techniques involved and in the experience for its application to real areas of society. For this reason, the consortium is formed by: Instituto Tecnológico de Informática (ITI), the multinational company GFT, the company 1MillionBot, the company Gente Comunicación and the University of Alicante. The phases into which the system is divided are presented in Figure 3.

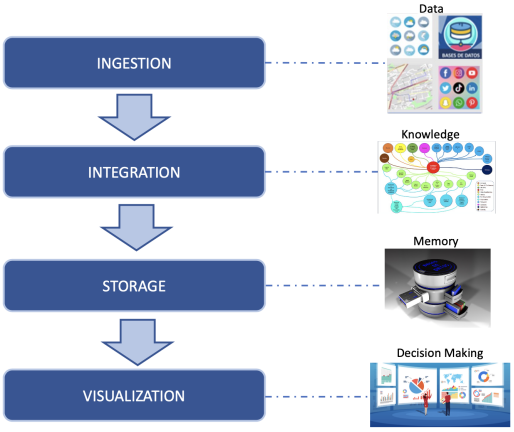

Figure 3: System Phases.

3.1. Ingestion

The data to be processed can be of different nature and format. It is therefore necessary to implement data cleansing, normalization, and transformation techniques in order to be able to enter data into the platform in a standardized way. Cleaning consists of correcting possible errors in the data, sometimes manual or often caused by trying different encoding conversions (i.e. UTF-8, EPSG[3], etc.). In cases where the data is not recognizable because it has undergone a significant alteration, it is eliminated. Standardization consists of unifying characteristics and information with the same meaning, under uniform criteria. For example: Alicante is equal to ALC. Transformation consists of extracting the data from the source document, recognizing its original format, and placing it in a common format among all the data incorporated into the platform, for example, GeoPandas[4]. Examples of types of data sources that may be relevant for this project are weather data API (e.g. OpenWeatherMap[5], GIS (e.g. GeoNet[6]), census, city cadastre, air quality data (e.g. Breezometer[7], social media user comments (e.g. Twitter [8] and Google[9], and other data sources such as the National Institute of Statistics[10] (INE).

[3] https://epsg.io/
[4] https://geopandas.org
[5] https://openweathermap.org/
[6] https://www.geonet.es/
[7] https://www.breezometer.com/
[8] https://twitter.com/home
[9] https://www.google.com/
[10] https://www.ine.es/

3.2. Integration

The data incorporated into the platform lacks semantic information. Therefore, from a previously designed model, which is able to capture the semantics of the domain, these data are instantiated, as pieces of semantic information, and linked together. The linking process consists of identifying common characteristics for each instance, and these characteristics are converted into semantic relationships.

3.3. Storage

The storage phase allows us to ensure that all data that has been instantiated and linked can persist as long-term memory, and also coexist with other data that has been previously incorporated. To guarantee this operation we opted for a non-SQL database, in which each instance becomes a document and each relation an index. The technology stack chosen for this platform was ElasticSearch[11] for textual information. It is a Lucene[12]-based search server that provides a full-text, distributed, multitenancy-capable search engine with a RESTful web interface and JSON documents. For integration queries GeoPandas, and Leaflet[13] y Folium[14] for map creation and visualization. The entire stack is built on the Python programming language.

3.4. Visualization

In this phase, a series of queries are defined based on the user needs indicated in the user interface. From these queries, data and metadata are obtained to generate visualizations that interweave data of different nature, source and domain, incorporated in the platform. With this ability to interlink data, it is possible to carry out cross-cutting studies and recommendations to facilitate decision-making and to propose new strategies, whether political, economic, geographic, or other, and mixtures of these areas.

One of the most important aspects to take into account in this task is semantic exploration and recommendation. Given the existence of a semantic database, it is necessary to develop mechanisms for the exploration of the semantic network, supported by SPARQL[15] or Cypher[16] queries, of document profiles and other digital entities. These mechanisms will allow to retrieve not only documents through metadata filters, but also to make aggregate queries to discover statistical trends, and to make recommendations of profiles (e.g., documents, authors, institutions, topics, named entities, etc.) through the semantic links that interconnect the network.

[11] https://www.elastic.co/
[12] https://lucene.apache.org/
[13] https://leafletjs.com/
[14] http://python-visualization.github.io/folium/
[15] https://skos.um.es/TR/rdf-sparql-query/
[16] https://neo4j.com/developer/cypher/

4. Impact GeoIA Project

This project is fully committed to the principle of efficient and rational public management, balancing control with productivity, by incorporating ICTs to improve administrative processes through advanced digitization tools and the provision of digital services to citizens and businesses. In this context, GeoIA impacts the following lines of action that are considered central to making smart cities a reality: digital public services (e-Government), paperless administration, inter-administrative cooperation and interoperability, rational management of ICT resources and promotion of technological innovation in public management. The specific actions included in the project are the promotion of the use and monitoring of the quality of the public administration's electronic services, the promotion of the integration and exchange of data and documents between administrations, the implementation of an electronic administration platform, etc. In this line, the implementation of Geo.IA (GeoArtificial Intelligence) will promote the generation of an ecosystem that revolves around Data, facilitating the development, prototyping and deployment of data analysis techniques in any domain of the economy and society. This will enable the incorporation of value from the exploitation of the data available in the project by leveraging data to improve management and decision making. The effective integration of innovation and research efforts in the project through the formed consortium will be boosted adding significant value. Geo.IA, serving as the cornerstone and vehicle, will act as the guiding thread for seamlessly incorporating these advancements into the productive and social fabric of the Valencian Community.

Geo.IA bases its value proposition on channeling the efforts of the Valencian Community in accelerating the digitization of our public sectors. The emphasis is on knowing the requirements of Valencian Community citizens in their interactions with public administration bodies through an AI platform to extract knowledge. This involves knowledge of citizens in line with their geospatial position and knowledge for the public administration to drive strategic decision making. Summarizing, this proposal aims to boost to the digitization and technological advancement of citizens and the Valencian Community's public administration, enabling the analysis of the data generated in the GeoIA system to provide evidenced based decision making.

5. Conclusions

GeoIA will provide solutions to integrate data from Geographic Information Systems (GIS), public sector organizations, census, land registry, cadastre, national statistics institute, weather forecast providers, social media, etc. These data will be integrated and transformed into Digital Entities (DE) to compose a Semantic Lake of Digital Entities in which real entities coexist and interrelate with external information, DbPedia for example. This will contribute to generate an ecosystem that can be exploited by public or private organizations and third party Apps oriented to citizen services. In addition, this allows analysis, intelligent visualization of data for decision making at different scales. The models will improve traditional geographic analytics, going from knowing where and when things happen to knowing why they happen in those places. If we add to this the information coming from public administration agencies (land registry, census, environmental information, etc.), a whole ecosystem is created, enriched by the use of knowledge that can be exploited by public administration agencies or private companies and by third party Apps oriented to citizen services.

Acknowledgments

This project is funded by the Valencian Agency for Innovation(AVI) and the European Regional Development Fund(ERDF) through the project "GeoIA: Artificial GeoIntelligence platform to solve citizens problems and facilitate strategic decision making in public administrations" (INNEST/2023/11), partially funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the project NL4DISMIS: TLHs for an Equal and Accessible Inclusive Society (CIPROM/2021/021). Moreover, it was backed by the work of two COST Actions: CA19134 - "Distributed Knowledge Graphs" and CA19142 - "Leading Platform for European Citizens, Industries, Academia, and Policymakers in Media Accessibility".

References

[1] Y. Gutierrez, D. Tomas, I. Moreno, Developing an ontology schema for enriching and linking digital media assets, Future Generation Computer Systems 101 (2019) 381–397.

[2] M. Lloret-Climent, A. Montoyo, Y. Gutiérrez, R. M. Guillena, K. Alonso-Stenberg, A systemic and cybernetic perspective on causality, big data and social networks in tourism, Kybernetes 48 (2019) 287–297. URL: https://doi.org/10.1108/K-02-2018-0084. doi:10.1108/K-02-2018-0084.

[3] Y. Gutiérrez, S. Vázquez, A. Montoyo, Spreading semantic information by word sense disambiguation, Knowledge-Based Systems 132 (2017) 47–61.

[4] Y. Gutiérrez, S. Vázquez, A. Montoyo, A semantic framework for textual data enrichment, Expert Syst. Appl. 57 (2016) 248–269. URL: https://doi.org/10.1016/j.eswa.2016.03.048. doi:10.1016/J.ESWA.2016.03.048.

[5] B. Navarro-Colorado, E. Saquete, Cross-document event ordering through temporal, lexical and distributional knowledge, Knowl. Based Syst. 110 (2016) 244–254. URL: https://doi.org/10.1016/j.knosys.2016.07.032. doi:10.1016/J.KNOSYS.2016.07.032.

[6] S. Estevez-Velarde, Y. Gutiérrez, Y. Almeida-Cruz, A. Montoyo, General-purpose hierarchical optimisation of machine learning pipelines with grammatical evolution, Inf. Sci. 543 (2021) 58–71. URL: https://doi.org/10.1016/j.ins.2020.07.035. doi:10.1016/J.INS.2020.07.035.

[7] A. Piad-Morffis, Y. Gutiérrez, Y. Almeida-Cruz, R. Muñoz, A computational ecosystem to support ehealth knowledge discovery technologies in spanish, J. Biomed. Informatics 109 (2020) 103517. URL: https://doi.org/10.1016/j.jbi.2020.103517. doi:10.1016/J.JBI.2020.103517.

[8] S. Estevez-Velarde, Y. Gutiérrez, A. Montoyo, Y. Almeida-Cruz, Automl strategy based on grammatical evolution: A case study about knowledge discovery from text, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4356–4365. URL: https://doi.org/10.18653/v1/p19-1428. doi:10.18653/V1/P19-1428.

[9] A. Piad-Morffis, Y. Gutiérrez, R. Muñoz, A corpus to support ehealth knowledge discovery technologies, J. Biomed. Informatics 94 (2019). URL: https://doi.org/10.1016/j.jbi.2019.103172. doi:10.1016/J.JBI.2019.103172.