Revista Española de Documentación Científica
46(4), octubre-diciembre 2023, e370
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2023.4.1393

ESTUDIOS / *RESEARCH STUDIES*

GeoAcademy: web platform and algorithm for automatic detection and location of geographic coordinates and toponyms in scientific articles

Jesús Cascón-Katchadourian*, Carlos Rodríguez-Domínguez**, Francisco Carranza-García***, Daniel Torres-Salinas****

*University of Granada, Faculty of communication and documentation, Dept of Information and Communication (Spain)
Correo-e: cascon@ugr.es | ORCID iD: 0000-0002-3388-7862
University of Granada, ETSIIT, Dept of Software Engineering (Spain)
**Correo-e: carlosrodriguez@ugr.es | ORCID iD: 0000-0001-5626-3115
***Correo-e: carranzafr@ugr.es ORCID iD:0000-0003-0876-3494
****University of Granada, Faculty of communication and documentation, Dept of Information and Communication (Spain)
Correo-e: torressalinas@ugr.es | ORCID iD: 0000-0001-8790-3314

Recibido: 27-11-2023; 2ª versión: 13-02-2023; Aceptado 01-03-2023; Publicado: 02-10-2023

Cómo citar este artículo/Citation: Cascón-Katchadourian, J., Rodríguez-Domínguez, C., Carranza-García, F., Torres-Salinas, D. (2023). Pro GeoAcademy: web platform and algorithm for automatic detection and location of geographic coordinates and toponyms in scientific article. *Revista Española de Documentación Científica*, 46 (4), e370. https://doi.org/10.3989/redc.2023.4.1393

Abstract: The following study relates the qualities and uses of the GeoAcademy Project, a program designed with the aim of geolocating scientific articles automatically, such articles would be found in Scopus, Web of Science, or similar databases. An algorithm has been developed with the intention of capturing geographical coordinates or toponyms contained within the documents in order to perform reliable geolocation. In the methodology, we describe the stages of the project that have been necessary so as to build a sample database concerning the Sierra Nevada (Spain), as well as the development of the algorithm. The technical data regarding the employment of the algorithm on the sample documents and its levels of success are included in the results, as is an explanation of the platform containing web maps which can be utilised to show the texts which have been geolocated. In conclusion we outline the obstacles faced, potential bibliometric uses and the advantages it offers as a reference resource and source of information.

Keywords: Automatic geolocation; geographical coordinates; toponyms; algorithm; scientific articles; Scopus.

GeoAcademy: plataforma web y algoritmo para la detección automática y localización de coordenadas geográficas en artículos científicos

Resumen: El siguiente estudio describe las cualidades y usos del proyecto GeoAcademy, un programa diseñado con el objetivo de geolocalizar artículos científicos automáticamente, dichos artículos se descargarían de bases de datos científicas generales como Scopus o Web of Science. Esta geolocalización se realiza sobre el contenido del documento, ya sea mediante la captura de posibles coordenadas geográficas que tenga el documento, o topónimos que puedan aparecer en el documento a través de un algoritmo creado a tal efecto. En la metodología explicamos los pasos que se han dado en este proyecto para crear una base de datos de muestra con artículos que tratan sobre Sierra Nevada (España) y la creación y diseño del algoritmo. Los resultados muestran los datos técnicos de la aplicación del algoritmo sobre la base de datos y su tasa de éxito, así como una descripción de la plataforma creada para visualizar gráficamente los documentos geolocalizados en un mapa web. Finalmente, en la discusión, definimos las dificultades encontradas, las posibles aplicaciones bibliométricas y su utilidad como herramienta de consulta y recuperación de información.

Palabras clave: Geolocalización automática; coordenadas geográficas; topónimos; algoritmo; artículo científico; Scopus.

1. INTRODUCTION

In the scientific world and in society in general there is currently a great interest in geolocation for various purposes. Geolocation of documentation is not exclusive to science but rather has occurred previously in other areas and types of documentation such as news (Buscaldi and Magnini, 2010; D'Ignacio et al., 2014; Imani et al., 2017; Teitler et al., 2008), historical documents through digital humanities (Cascón-Katchadourian et al., 2019), social media content (Dredze et al., 2013; Zhang and Gelernter, 2014; Middleton et al., 2018) and many others. One of the causes of interest in geolocation is due to the great documentary explosion that has occurred in recent decades. With so much information available comprehension has become difficult, leading to the search for alternatives, many of them graphical, to visualize and filter the information. Science does not escape this global trend. Its recent popularization is linked to the appearance and massive use over the last decade of websites, apps and social networks that offer services depending on the user's needs and surroundings. Examples of such resources, among others, are Foursquare, TripAdvisor, GoogleMaps.

Geolocation involves determining the geographical location of information so that it can be processed (RamosVacca and Bucheli-Guerrero, 2015). Advancements in aerial photogrammetry and digitized cartography (Cortés-José, 2001), as well as new map printing methods, have contributed to this. The advent of programs such as Google Maps, Bing Maps, and OpenStreetMap, along with mapping creation tools like OpenLayer, Leaflet, CartoDB, and MapTiler (Cascón-Katchadourian and Ruiz Rodríguez, 2016), has further advanced the field. In scientific evaluation and bibliometrics, geolocation is often used to identify the geographical origin of scientific articles based on their authors (Catini et al., 2015). However, few studies have analyzed scientific publications based on their geographical coordinates or toponyms and mapped them.

The aim of this article is to design an algorithm-based tool that helps scientists to classify scientific articles with geographic information. With current search systems, a manual filter is necessary in order to identify the locations they are interested in, which may be needlessly time consuming. An algorithm is needed to improve the accuracy of information retrieval systems in terms of geographic information. Likewise, being able to visualise scientific corpus would facilitate geographic meta-analysis (Gerstner et al., 2017). It would also facilitate the discovery of areas of the earth that have previously been studied to varying degrees (Fisher et al., 2011; Martin et al., 2012). In this work we analyze the geolocation of the content of the document, that is, geolocate the documentation according to the places or locations that are observed in it. We are not interested in the geolocation of the address of the authors of the articles.

Here a subsection called Related Works has been added where initiatives similar to ours and the differences they have with our project are explained. Secondly, the three fundamental objectives pursued by this research are given as well as a justification for it being carried out. Thirdly, the methodology explains step by step the search process, collection processing, as well as everything related to the development of the algorithm and its application to the database. Fourthly, in the results section, we present the main achievements of this research with the support of two tables with the success rates and frequency of place names and a figure to show the index of the website created for this purpose. Finally, in the discussions we reflect the limitations, a comparison of the results and future lines of research.

1.1 Related Work

We would like to draw attention to two current initiatives due to their significance and because they have the support of several institutions in their development. The first initiative is GEOUP4, which is a "web portal shows on an interactive map the geolocation of the academic items of the repositories of the UPC, UPCT, UPM and UPV polytechnic universities grouped in the UP4 Association" (GEOUP4, 2022). The Universitat de Girona and the Carlos III University of Madrid also collaborate with it. The user can consult the different academic works and publications (eg. thesis, articles, conferences, master's thesis, degree thesis) that are related to the territory in a geolocalized way. They can also access the author's original document found in the repositories of the libraries of the respective universities. Its objectives include increasing the visibility and impact of the academic production of the participating universities, as well as making available data related to the study and impact of the territory in relation to the academic production of the UP4 universities.

The other project is JournalMap (Karl et al., 2012; JournalMap, 2022) a cooperative project between the USDAARS Jornada Experimental Range in Las Cruces, NM and the Idaho Chapter of The Nature Conservancy. This is one of the components of another tool called Landscape Toolbox. It is a project that geolocates scientific production based on the geographical location where the study is carried

out to observe which areas have been studied and in which there are gaps. It also provides the possibility of searching for scientific literature of similar areas, in the sense that they share characteristics of soil, climate, slope, altitude and other variables that function as layers of the system. This is done with both automatic and manual techniques.

The present paper links with these two projects. Our study is original and useful since it aims to geolocate the scientific articles in an automatic way through algorithms created for this purpose and not in a manual or semiautomatic way. As for the uses of geolocation related scientific knowledge, there is great interest in using this information as input for public policy making, analysis of scientific geographic clusters, or to evaluate the distribution and impact of the research conducted. For example, in Spain there is a bidding process for investment in national parks designed by the Ministry for the Ecological Transition and Demographic Challenges (Ministerio para la transición ecológica y el reto demográfico, 2022) where €1,693,271.50 is distributed for research projects. For those carrying out this program, it would be useful to have a mapping or cartography of the national parks and the research that has been done into them, to identify little-studied areas and high-impact areas where it is more efficient to invest money. This question has been approached from numerous disciplines (Inoue et al., 2013), both for its diverse applications and for the possibility of making it scalable to interdisciplinary work (Ramos-Vacca and Bucheli-Guerrero, 2015). Currently, it is one of the fields with the greatest impact and growth within computer science (Bordogna et al., 2012) (Bornmann et al., 2011) (Ramos-Vacca and Bucheli-Guerrero, 2015).

There are many studies and reports that tell us about geolocation of scientific knowledge from the point of view of who has written the article, to which institution it belongs, as well as its city or country of origin. Very popular, for example, are the national and international rankings that have been published based on these metrics: Shanghai Ranking (ShanghaiRanking, 2022), Scimago Institution Ranking (Scimago, 2022), UMultirank (Umultirank, 2022) and CWTS Leiden Ranking (CWTS, 2022). Also well-known are the bibliometric databases and suites that allow these types of studies, such as Incites Essential Science Indicators (ESI) (Clarivate, 2022). Some of these rankings or bibliometric products have also taken advantage of geolocation to create geographic web applications that show the impact of articles published by each institution on a map, whether they be universities or research-focused institutions. This is the case for Mapping Research Excellence (Bornmann et al., 2021; Mapping Research Excellence, 2022) which is based on Scopus data taken from the SCImago Institutions Ranking (Scimago, 2021), as well as altmetrics whose source is Mendeley. The interface shows a circle for each institution. The size of the circle is related to the number of papers published by that institution. The color of the circle is related to the percentage of papers that are highly cited according to Scopus or highly marked in Mendeley. All this is based on the result of multilevel logistic regression models.

But, extracting and representing geographic locations automatically remains a little studied problem (Kmoch et al., 2018; Tamames and Lorenzo, 2010; Leveling, 2015). In the few studies on the subject, several problems exist such as not using the full text (Fisher et al., 2011), only geographic coordinates are identified (Karl, 2018; Page, 2010) or too low performance when full text is added (Kmoch et al., 2018). A recently published article (Acheson and Purves, 2021) takes an interesting approach to the problem by extracting and representing geographical toponyms with the application of multiple techniques. However, this approach has limitations since it does not address the issue of geographic coordinates, which is a key issue to resolve.

2. OBJECTIVES

Taking into account, as just analyzed, that there is a growing interest in mapping and geolocating science, either from a private perspective or from an institutional perspective, this article proposes new forms of geolocation that serve as a tool for analysis, visualization and retrieval of a line of research. More specifically, there are three main objectives for this project:

- (1) to develop an algorithm that allows the coordinates and toponyms to be extracted from a collection of documents to identify exactly which places these studies deal with.
- (2) The second objective is that once the locations have been identified by the algorithm, they will be displayed on a map through an online platform.
- (3) In the third objective, the platform will integrate a layer of bibliometric and/or altmetric information that will allow one to know the volume of production according to coordinates as well as different data on the scientific and social impact.

Therefore, we will present the results of three objectives applied to a set of scientific studies concerning the Sierra Nevada mountain range (Granada, Spain). Our study is original and useful because it aims to geolocate the scientific production in an

automatic way by means of algorithms created for this purpose and not in a manual or semi-automatic way.

3. MATERIAL AND METHODS

It should be mentioned that in this paper we apply a work on a specific line of research or a specific geographic collection. Thus, as opposed to the traditional analysis based on topics or subjects, such as those generated by bibliometric software like VosViewer (Van Eck and Waltman, 2010) and Scimat (Cobo et al., 2012), we propose to visualize a line of research through a map using coordinates and toponyms, using a geographic search engine for this purpose.

To create a collection of documentary information about Sierra Nevada, a search was done in the Scopus database. The Scopus database has been chosen for the facilities it offers for exporting full-text documents, thanks to Scopus Document Download Manager. The first step was to perform a search in the Scopus database with the following search equation: (TITLE-ABS-KEY ("sierra nevada") AND TITLE-ABS-KEY (spain OR granada)) AND (LIMIT-TO (DOCTYPE, "article")). With this type of search, the objective was to find scientific articles that dealt with the Sierra Nevada mountain range located in the province of Granada, Spain. As there were other mountain ranges with the same name in other parts of the world, the term Spain or Granada was added to the search. It searched through scientific articles, without a time limit. This search found 623 articles. The second step was the processing of the information for storage. After the pertinent manual verification work (the geolocation of the documents is automatic, not the selection of the sample), there are 447 documents with a complete associated pdf (not only the abstract or title) and which are openly accessible, the other 176 (623-447) records do not have an associated pdf, are incomplete or are not an open publication, of which 424 are about the Spanish Sierra Nevada and not about the Sierra Nevada of California or Peru.

The third step was the identification of the coordinates. Although there are many types of coordinate systems, the most common that this project has found in the sample are the following types and subtypes of coordinates: the geographic coordinate system and the coordinate system UTM (Universal Transverse Mercator). The geographic coordinate system is subdivided into sexagesimal (degrees, minutes, and seconds) and decimal (degrees and decimals). The UTM coordinate has multiple variants in that the authors express them in different ways, with or without the letters of the cardinal points, specifying the use or not (in our case it is 30S). Finally, in environmental studies, they usually use a subdivision of the use 30S, which is used in army maps, and which is reflected in articles with VG, VF, WG and WF. Finally, our database contains 424 bibliographic references linked to their corresponding PDFs. This database has been used for the design and training of the algorithm. The database is accessible at https://doi.org/10.5281/zenodo.5633963

For the geolocation of the contributions, we have developed a Text Mining Algorithm based on the extraction of information through regular expressions and toponyms - keywords (see Figure 1). To carry out this automatic geolocation process, we design an algorithm based on 4 stages:

I-Preprocessing: in this stage, the contributions of the sample are processed in PDF format, converting the text of each one to a processable format (plain text), normalized following the next steps:

a. Convert all text to lowercase.
b. Remove punctuation marks, such as periods, commas, exclamation marks, etc.
c. Remove non-alphanumeric characters, such as numbers, symbols, emoticons, etc.
d. Remove stop words, such as "the", "of", "and", etc.
e. Stemming and/or lemmatization, which involves reducing words to their root or lemma, to facilitate processing by a language model.
f. Optionally, transform the words into tokens or numerical sequences, to be processed by machine learning algorithms.

For the conversion of PDF files to plain text, we have used *pdftotext* which is an open-source command-line utility and *Python* (version 3.10.0) as programming language.

II-Extraction: search and extraction of geographical references in the text using two methods, (i) regular expressions for the search of geographical coordinates in their different formats (decimal, sexagesimal and UTM) and (ii) search for the frequency of appearance of place names. The database of toponyms for this study has been downloaded from the Nomenclator Geográfico de Andalucía (NGA) of the Instituto de Estadística y Cartografía de Andalucía (Instituto de Estadística y Cartografía de Andalucía, 2022). This database has been downloaded using the interoperable Web Feature Server (WFS) services. With the help of the QGIS software and using a SQL-like WHERE, the toponyms that interest us have been downloaded and exported in xlsx spreadsheet format.

III-Transformation: to georeferenced the results obtained on an interactive map; all the results are processed to unify them in decimal format (latitude, longitude). To convert between the different geographic representations, we have used Proj4js which is a library to transform point coordinates from one coordinate system to another, including datum transformations.

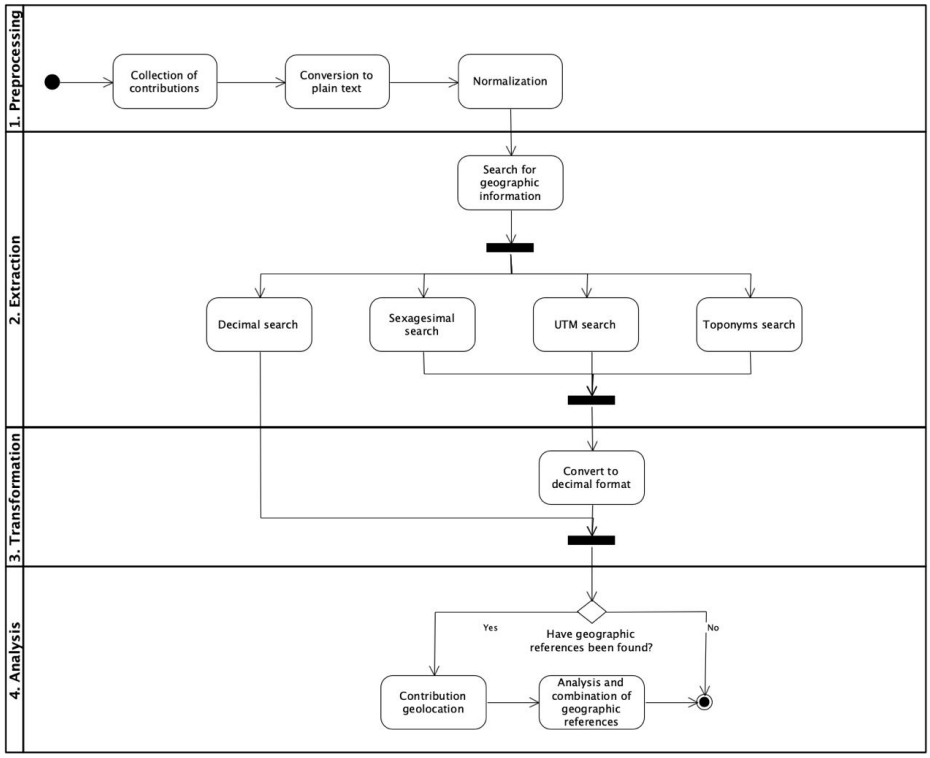
IV-Analysis: finally, once we have all the information that has been extracted in the same format, a small analysis is carried out to decide, through heuristic behavior, if it is possible to geolocate the contribution. For this, we have mainly followed three criteria: (i) if geographic coordinates and place names have been found, we combine this information to select the coordinate that is textually closest to the most frequent place name; (ii) if only geographic coordinates have been identified, it is geolocated within the most referenced area; and (iii) if only toponyms have been recognized, we geolocate it in the one with the highest frequency as long as it exceeds a certain threshold (this threshold is calculated as the simple mean of the frequencies of the toponyms in the entire sample of this study).

It is important to note that, in this first version of the algorithm, we base the geographic reference extraction process on the combination of information from two methods:

1. **Regular expressions**: we have defined a set of search patterns to find geographic coordinates (in their different formats) in the texts of the selected articles. For example, the regular expression **^([-+]?)([\d]{1,2})(((\.)(\d+)(,)))(\s*)(([-+]?)([\d]{1,3})((\.)(\d+))?)$** as search pattern will find all coordinates specified in texts in "latitude, longitude" form like "37.1809792, -3.6087813". Thanks to this set of regular expressions and the Reverse Geocoding API, we have been able to calculate the distance between the locations found to find out the relevant geographic and contextual information for this study, including finding out if the georeferencing is from a specific area or from exact geographic locations.

2. **Toponyms**: toponyms have been identified by searching for the frequency of appearance of the words included in the database and their subsequent geographic location using the coordinates provided in this database or, failing that, by georeferencing through the Google Geocoding API. In this first version we start from an already created database, we do not use techniques such as NER (Named Entity Recognition) to extract relevant toponyms from the articles but we are working to incorporate this type of techniques in a following version, combined with NLP techniques to acquire more semantic information of the texts.

Figure 1: Text Mining Algorithm.



Once this automatic process is finished and in order to check if there is a relationship between the georeferencing of contributions and their bibliometric information, we have included all the available metrics through an Almetrics report.

To do this we have linked the results of the bibliometric report with the dataset of contributions through its DOI, presenting special interest to the variables: (i) Altmetric Attention Score, (ii) Twitter mentions, (iii) Mendeley readers and (iv) citations. This information aggregation process has been carried out with automatic processing that matchs the Altmetrics report with the contributions of the initial dataset through its DOI and with the help of the Python programming language.

4. RESULTS

Table I shows a summary of the results of the algorithm of automatic detection of geolocations being applied to the collection of 424 scientific articles on Sierra Nevada. In total, the algorithm has been able to geolocate 157 articles with coordinates, 37% of the total number of articles analysed in the study. In this research, all the articles that contain geographic coordinates (205) have been manually identified to test the platform, in their different variables, included in text or image. In table 1 we can see the performance of our algorithm in detail on the total number of articles with geographic coordinates. One of the tools used to increase the number of geolocated articles and elements has been the use of place names that have allowed us to geolocate 512 place names of 220 different articles. There are 5.84 place names on average per scientific article. Finally, the number of works that have been geolocated using the algorithm is 377, that is to say 88.9% of the original document collection has been located, either by coordinates or toponyms. It should be mentioned that one of the fundamental aspects to improve the success rate of the algorithm has been the use of place names.

Cases for which an article may not be georeferenced in our proposed pipeline:

- Not having coordinates identifiable by regular expressions because the coordinate expression does not match the pattern specified in the regular expression set or they are inside images.
- That you do not find toponyms or that they do not have sufficient frequency and importance to constitute a georeferencing.

With the completion of objective 1, the team shifted their focus to developing a software application that would allow users to view scientific articles on a digital map based on the geographical information they contain. The resulting portal, which is currently in beta format and can be accessed at https://geoacademy.everyware.es/ (Geoacademy, 2022), features the collection of documents about Sierra Nevada that was analyzed in the study. By clicking on each marker on the map (as shown in Figure 2), users can access information about the paper and a link to view it. The filter function allows users to search for articles based on type (such as articles, conference papers, or theses), or by keyword. The portal also includes a traditional search engine with filters and a search radius, as well as tabs that describe the project and its

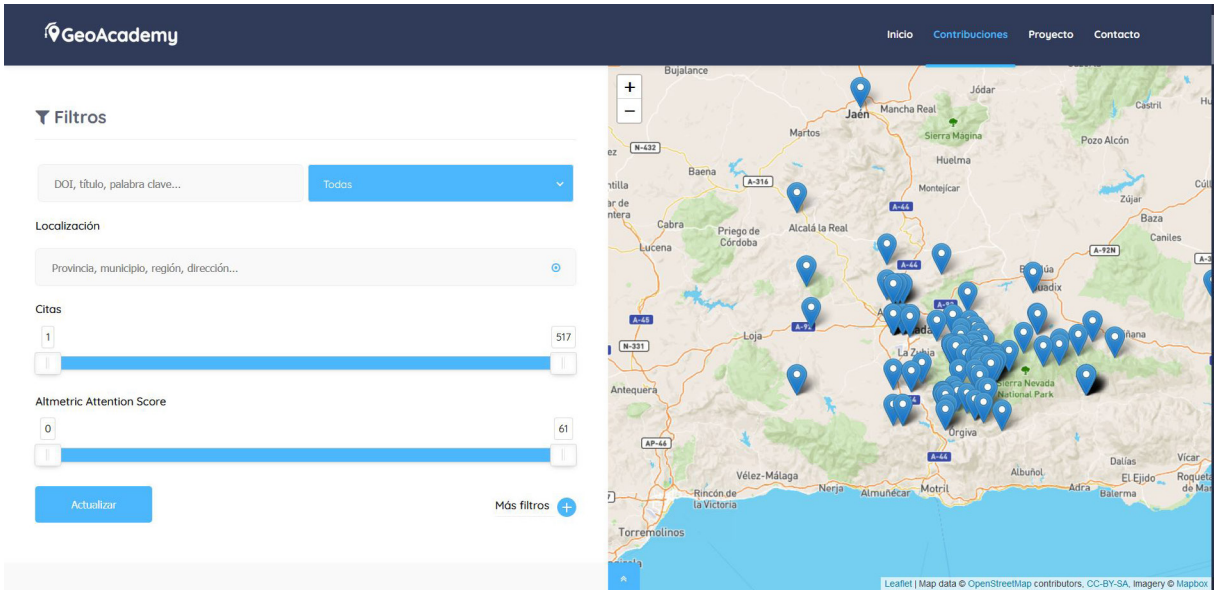
Table I: Indicators and general statistics in the algorithm training process.

A) General indicators related to geolocation by coordinates	
A1. Number of articles analysed in the study	424
A2. Number of articles containing geographical coordinates (157 + 43 (images) + 5 (UTM))	205
A.3 Number of articles containing geographical coordinates identified by the algorithm	157
A.4. Percentage of articles geolocated through geographical coordinates (A3/A2)	76.59%
A.5. Percentage of articles geolocated through geographical coordinates (without images (43))	96.91%
A.6. Percentage of articles geolocated through geographic coordinates over the total number of articles (A3/A1)	37.02%
B) Indicators related to geolocation by place names	
B.1. Number of place names identified by the algorithm	512
B.2 Number of articles geolocated by place names	220
B3. Percentage of articles geolocated by place names (B2/A1)	55.9%
C) Findings	
C.1 Number of geolocalisations achieved (coordinates + place names) (A.3+B.2)	377
C.2. Percentage of articles geolocated from the total number of articles (C1/A1)	88.9%

Table II: A metrics summary of the main toponyms from Andalusian Institute of Statistics and Cartography included in the Sierra Nevada collection.

Name	Type	Province	Frequencies in the collection	Number of citations	Twitter mentions	Mendeley mentions
Sierra Nevada	Entidad Singular INE	Granada	8771	10740	639	10424
Veleta	Elevación	Granada	1008	1819	91	1418
La Caldera	Lugar / Paraje	Granada	534	663	50	597
Río Seco	Curso Fluvial	Granada	494	1081	158	1557
Mulhacén	Elevación	Granada	441	2402	70	1481
Nevada	Municipio	Granada	293	2555	179	1781
La Sierra	Elevación	Granada	230	1529	31	700
Bérchules	Municipio	Granada	212	77	35	162
Guadix	Lugar / Paraje	Almería	211	1135	12	230
Castillo	Vértice Geodésico	Almería	184	1342	129	1369
Mecina	Población	Granada	184	664	20	537
Dílar	Municipio	Granada	183	791	33	618
Lanjarón	Municipio	Granada	169	2191	30	1799
San Juan	Lugar / Paraje	Granada	164	890	46	474
Sierra de Baza	Elevación	Almería	153	1322	11	428
Borreguiles	Servicio/Equipamiento	Granada	122	767	100	887
Válor	Municipio	Granada	119	1053	26	868
Colorado	Lugar / Paraje	Granada	110	1274	55	979
Calar	Vértice Geodésico	Granada	103	463	22	129
Dúrcal	Municipio	Granada	97	809	84	480
Monachil	Capital de Municipio	Granada	91	436	25	249
La Virgen	Lugar / Paraje	Granada	88	233	114	704
Trevélez	Municipio	Granada	83	570	39	459
Puertas	Lugar / Paraje	Almería	77	259	69	354
Calvache	Lugar / Paraje	Almería	72	597	0	232

Figure 2: GeoAcademy application showing the geolocated locations of the Sierra Nevada article collection, and bibliometric filters.



members and provide a contact form. In addition to the metadata from the databases (in this case, Scopus), each processed scientific work includes metadata generated by the algorithm, such as the decimal format coordinates and a complete list of identified place names, which are added to the bibliographic description.

The third objective of this work is to provide the platform with a layer of information capable of representing information of a bibliometric and altmetric nature, that is, of scientific and social impact. In this sense, the GeoAcademy platform has been provided with the following functionalities 1) Filter locations based on indicators (Number of citations, Altmetric Attention Score) 2) Use of geoposition markers differentiated according to values of the indicators. 3) The platform will present a bibliometric summary of the different coordinates and locations, offering bibliometric information at a level not currently seen.

An example of this last point can be seen in the following table. Table 2 shows the main toponyms ordered by frequency of occurrence in our database. These toponyms have been taken from the public databases of the Instituto Andaluz de Estadística y Cartografía. With the visualization of the table we can see the most studied toponyms of Sierra Nevada.

5. DISCUSSION

We believe that in this study a profound approach has been taken for the creation of an algorithm that is useful for researchers and society in general. With the chosen sample, its operation has been tested intensively, however we are aware there is still room for improvement.

This study contains a number of limitations that we proceed to list and which have mainly to do with processing. The first problem relates to the processing of PDF documents. We found several that did not have an associated pdf, for others the pdf requires payment, or only the title or the title and abstract are available. This limits the number of results you can cover. We hope that in the future the current trend of improvements in open science will continue, overcoming this limitation. Secondly, the coordinates have some formal standardization, whereas in practice each field of knowledge has different guidelines for writing the coordinates. There is greater variability with the UTMs that the algorithm has overcome without problems. In future updates of the algorithm we hope to resolve all the multiple variables found. Another problem is that several studies show the coordinates in image form, that makes them more difficult to extract, although in the future through OCR technology they could be captured. In future updates of the algorithm we will incorporate two Python libraries that are demonstrating promising results to extract text from images; OpenCV (Open Source Computer Vision Library) (https://opencv.org/) and Tesseract (https://github.com/tesseract-ocr/tesseract). At the linguistic level, words with multiple means in the toponymy can be problematic, which can give both false positives and false negatives. Also, the approach does not allow more generic descriptions like 'mountain ranges in France' or 'forests in Spain' to be included while these such studies may still be relevant. The most effective way to solve these last two problems is through semantic searches, disambiguation, and the creation of semantic networks.

Regarding the comparison with other proposals, the difference with our project is that in GEOUP4 each document is geolocated manually, this is done by the library services. Although they are studying ways to do it automatically with natural language processing, so far it is a manual process. JournalMap, on the other hand, uses a dual approach, geolocating with automatic and manual techniques to determine geographic locations. If the coordinates can be detected, it generates the location automatically; if they cannot be detected or there are only place names, they are geotagged manually by the authors who upload the paper to the platform. With respect to our project there are two differences, firstly our algorithm also detects place names and secondly our project does not need the authors to upload their papers, rather they are downloaded by us from the database. The success rate of the Geoacademy algorithm on all documents is 88.9%, a high percentage for this type of study, especially if we consider that it is made on a general and interdisciplinary collection of documents (within our database there are studies of earth sciences, anthropology, physics and geography, amongst others), not only on a specific and well-defined scientific field. This is one of the main criticisms of Gritta et al. (2018) about the tendency to develop tools for very specific tasks and corpus. Putting to one side the distances and taking into account the different techniques developed and forms of evaluation of each study and software, the research of Gritta et al. (2018) that compares 5 software programs offers accuracy of between 81 and 21 percent when extracting place names. Something similar happens with Acheson's research (Acheson and Purves, 2021), with a toponym extraction rate of between 81 and 86 percent depending on the corpus.

We will continue to train the algorithm with much larger documentary collections related to other ge-

ographical features in order to reduce its limitations. A potential improvement would be to create a directory of place names for each search or study, making the process semi-automatic, however there are geocoding services such as Google Geocoding API, GeoNames or OpenStreetMap (OSM) Nominatim that would help us with this task, in addition to NER (Named-Entity Recognition) that would facilitate extracting information from the previous geocoding services. Likewise, it will be applied in other contexts, such as archaeological where most of the articles include precise geographic coordinates. It could also be applied to digital humanities, since interesting projects are being carried out where historical works are geolocated through the place names that appear in them.

Future developments could include its inclusion with an industry tool such as Bibliometrix. It would also be very interesting if users of large scientific databases such as Scopus or Web of Science, having performed a search with a list of results, could see said list geographically on a map, applying our algorithm for a better user experience. That is to say, to integrate our project in their platforms. Finally, numerous studies do not work with geographical points, we are currently working on how to show these studies on our platform.

6. ACKNOWLEDGMENTS

This manuscript was supported by the University of Granada Medialab Project (No. MLAB2019-02).

AGRADECIMIENTOS

Este manuscrito ha sido financiado por el Proyecto Medialab de la Universidad de Granada (Nº MLAB2019-02).

7. AUTHOR CONTRIBUTION STATEMENT

Jesús Cascón-Katchadourian and Daniel Torres-Salinas have designed the research, the QUERY, processed the pdfs as well as written the main manuscript. Carlos Rodríguez-Domínguez and Francisco Carranza-García have designed and tested the algorithm as well as created the web platform where the results are displayed and the searches are carried out.

8. DATA AVIABILITY

This paper is a substantially extended version (at least 50% new material) of the ISSI2021 conference paper: "GeoAcademy: algorithm and platform for the automatic detection and location of geographic coordinates in scientific articles".

9. REFERENCES

Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. PloS one, 16(1), e0244918. DOI: https://doi.org/10.1371/journal.pone.0244918.

Bordogna, G., Ghisalberti, G., & Psaila, G. (2012). Geographic Information Retrieval: Modeling Uncertainty of User's Context. *Fuzzy Sets and Systems*, 196, 105-124. DOI: https://doi.org/10.1016/j.fss.2011.04.005.

Bornmann, L., Leydesdorff, L., Walch-Solimena, C., & Ettl, C. (2011). Mapping excellence in the geography of science: An approach based on Scopus data. *Journal of Informetrics*, 5(4), 537-546. DOI: https://doi.org/10.1016/j.joi.2011.05.005.

Bornmann, L., Mutz, R., Haunschild, R., Moya-Anegon, F., Clemente, M., & Stefaner, M. (2021). Mapping the impact of papers on various status groups in excellencemapping.net: a new release of the excellence mapping tool based on citation and reader scores. *Scientometric,* 126, 9305–9331. DOI: https://doi.org/10.1007/s11192-021-04141-4.

Buscaldi, D., & Magnini, B. (2010). Grounding toponyms in an italian local news corpus. *Proceedings of the 6th workshop on geographic information retrieval*, 1-5. DOI: https://doi.org/10.1145/1722080.1722099

Cascón-Katchadourian, J., & Ruiz Rodríguez, A. Á. (2016). Descripción y valoración del software MapTiler: Del mapa escaneado a la capa interactiva publicada en la web. *Profesional de la Información*, *25(*6), 970978. DOI: https://doi.org/10.3145/epi.2016.nov.13.

Cascón-Katchadourian, J., López-Herrera, A. G., Ruiz-Rodríguez, A. Á., & Herrera-Viedma, E. (2019). Proyecto Histocarto: aplicación de SIGs (georreferenciación y geolocalización) para mejorar la recuperación de la documentación histórica gráfica. *Profesional de la Información*, 28(4). DOI: https://doi.org/10.3145/epi.2019.jul.16.

Catini, R., Karamshuk, D., Penner, O., & Riccaboni, M. (2015). Identifying geographic clusters: A network analytic approach. *Research policy*, 44(9), 1749-1762. DOI: https://doi.org/10.1016/j.respol.2015.01.011.

Clarivate. (2022). *Incites.* Available at: https://incites.clarivate.com/#/landing.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8), 1609-1630. DOI: https://doi.org/10.1002/asi.22688.

Cortés-José, J. (2001). El documento cartográfico. In J. Jiménez-Pelayo, J. Monteagudo-López-Menchero, & F. J. Bonachera-Cano, *La documentación cartográfica : Tratamiento, gestión y uso*, 37-113. Huelva: Universidad de Huelva.

CWTS. (2022). *CWTS Leiden Ranking 2022.* Available at: https://www.leidenranking.com/.

D'Ignazio, C., Bhargava, R., Zuckerman, E., & Beck, L. (2014). Cliff-clavin: Determining geographic focus for news articles. *DSpace@MIT*. Available at: https://hdl.handle.net/1721.1/123451.

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). Carmen: A twitter geolocation system with applica-

tions to public health. *Workshops at the twenty-seventh AAAI conference on artificial intelligence*.

Instituto de Estadística y Cartografía de Andalucía. (2022). *Toponimía-Nomenclátor.* Available at: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/prodCartografia/toponimia/index.htm#:~:text=El%20Nomencl%C3%A1tor%20Geogr%C3%A1fico%20de%20Andaluc%C3%ADa,%2C%20extractivas%2C%20servicios%20y%20equipamientos.

Fisher, R., Radford, B. T., Knowlton, N., Brainard, R. E., Michaelis, F. B., & Caley, M. J. (2011). Global mismatch between research effort and conservation needs of tropical coral reefs. *Conservation Letters*, 4(1), 64-72. DOI: https://doi.org/10.1111/j.1755-263X.2010.00146.x.

Geoacademy. (2022). *Geoacademy.* Available at: https://geoacademy.everyware.es/.

GEOUP4 (2022). *GEOUP4. Las universidades en el territorio. Geolocalización de la producción científica.* Available at: http://geo.up4.es/.

Gerstner, K., Moreno-Mateos, D., Gurevitch, J., Beckmann, M., Kambach, S., Jones, H. P., & Seppelt, R. (2017). Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods in Ecology and Evolution*, 8(6), 777-784. DOI: https://doi.org/10.1111/2041-210X.12758.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2), 603-623. DOI: https://doi.org/10.1007/s10579-017-9385-8.

Imani, M. B., Chandra, S., Ma, S., Khan, L., & Thuraisingham, B. (2017). Focus location extraction from political news reports with bias correction. *2017 IEEE International Conference on Big Data (Big Data)* 1956-1964. DOI: https://doi.org/10.1109/BigData.2017.8258141.

Inoue, H., Nakajima, K., & Saito, Y. U. (2019). Localization of collaborations in knowledge creation. *The Annals of Regional Science*, 62(1), 119-140. DOI: https://doi.org/10.1007/s00168-018-0889-y.

JournalMap (2022). *JournalMap.* Available at: https://www.journalmap.org/.

Karl, J. W. (2019). Mining location information from life-and earth-sciences studies to facilitate knowledge discovery. *Journal of Librarianship and Information S*cience, 51(4), 1007-1021. DOI: https://doi.org/10.1177/0961000618759413.

Karl, J. W., Unnasch, R. S., Herrick, J. E., & Gillan, J. (2012). JournalMap: geo-semantic searching for relevant knowledge. In *Ecological Society of America Proceedings*.

Kmoch, A., Uuemaa, E., Klug, H., & Cameron, S. G. (2018). Enhancing location-related hydrogeological knowledge. *ISPRS International Journal of Geo-Information*, 7(4), 132. DOI: https://doi.org/10.3390/ijgi7040132.

Leveling, J. (2015). Tagging of temporal expressions and geological features in scientific articles. *Proceedings of the 9th Workshop on Geographic Information Retrieval*, 1-10. DOI: https://doi.org/10.1145/2837689.2837701.

Mapping Research Excellence. (2022). *Excellence Maps v2*. Available at: https://www.excellencemapping.net/

Martin, L. J., Blossey, B., & Ellis, E. (2012). Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations. Frontiers in Ecology and the Environment, 10(4), 195-201. DOI: https://doi.org/10.1890/110154.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. ACM Transactions on Information. *Systems (TOIS)*, 36(4), 1-27. DOI: https://doi.org/10.1145/3202662.

Ministerio para la transición ecológica y el reto demográfico. (2022). *Programa de investigación. Convocatoria 2022.* Available at: *https://www.miteco.gob.es/es/red-parques-nacionales/programa-investigacion/convocatoria2022.aspx.*

Page, R. D. (2010). Enhanced display of scientific articles using extended metadata. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2-3), 190-195. DOI: https://doi.org/10.1016/j.websem.2010.03.004

Ramos-Vacca, I. D., & Bucheli-Guerrero, V. A.. (2015). Automatic geolocation of the scientific knowledge: Geolocarti. *10th Computing Colombian Conference (10CCC),* 416-424.

Scimago. (2022). *Ranking Methodology*. Available at: https://www.scimagoir.com/methodology.php.

ShanghaiRanking. (2022). *ShanghaiRanking.* Available at: https://www.shanghairanking.com/rankings/arwu/2021.

Tamames, J., & de Lorenzo, V. (2010). EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC bioinformatics*, 11(1), 1-10. DOI: https://doi.org/10.1186/1471-2105-11-294.

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., & Sperling, J. (2008). NewsStand: A new view on news. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, 1-10. DOI: https://doi.org/10.1145/1463434.1463458.

Umultirank. (2022). *Umultirank. Universities compared. Your way. Available at:* https://www.umultirank.org/.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics*, 84(2), 523-538. DOI: https://doi.org/10.1007/s11192-009-0146-3.

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, 9, 37-70. DOI: http://dx.doi.org/10.5311/JOSIS.2014.9.170.