HAL
open science

NETSCITY: a geospatial application to analyse and map world scale production and collaboration data between cities

Marion Maisonobe, Laurent Jégou, Nikita Yakimovich, Guillaume Cabanac

To cite this version:

Marion Maisonobe, Laurent Jégou, Nikita Yakimovich, Guillaume Cabanac. NETSCITY: a geospatial application to analyse and map world scale production and collaboration data between cities. International Conference on Scientometrics and Informetrics (ISSI 2019), Sep 2019, Rome, Italy. hal-02301035

HAL Id: hal-02301035

https://hal.science/hal-02301035v1

Submitted on 30 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This is a postprint of:

M. Maisonobe, L. Jégou, N. Yakimovich et G. Cabanac. *NETSCITY: a geospatial application to analyse and map world scale production and collaboration data between cities.* ISSI'19: 17th International Conference on Scientometrics and Informetrics. 2019.

NETSCITY: a geospatial application to analyse and map world scale production and collaboration data between cities

Marion Maisonobe[1], Laurent Jégou[2], Nikita Yakimovich[2], and Guillaume Cabanac[3]

[1] *marion.maisonobe@cnrs.fr*
CNRS, Géographie-cités UMR 8504 CNRS, Paris (France)

[2] *laurent.jegou@univ-tlse2.fr*, *nikita110598@mail.ru*
University of Toulouse, LISST UMR 5193 CNRS, Université Toulouse Jean Jaurès, Toulouse (France)

[3] *guillaume.cabanac@univ-tlse3.fr*
University of Toulouse, Computer Science department, IRIT UMR 5505 CNRS, Toulouse (France)

Abstract

We present NETSCITY, an online application to analyse and visualise world scale scientific production and collaboration data between cities. Contrary to existing tools that mainly focus on displaying co-occurrence networks, NETSCITY especially focuses on processing the geographical information comprised in bibliometric data. NETSCITY proposes a fully integrated solution to parse and clean the authors' addresses, comprised in a set of references, geocoding them at the city level, clustering them at the requested level of analysis (urban areas or countries) and mapping them either on a world base map or in a relational space. In the first part of the paper, we stress the originality and design of the NETSCITY application in terms of geocoding, clustering, counting, and mapping methods. In the second part, we detail its main functionality as a geoweb application. Eventually, we show the results one can obtain by applying NETSCITY on a set of references extracted from the online version of the Web of Science Core Collection.

Introduction

Practitioners in the field of bibliometrics contributed a variety of software tools to simplify the processing and mapping of bibliographic data. We can think of the online dashboards available through the Web of Science Core Collection and Scopus websites. There is also a variety of free bibliometric mapping software tools such as VOSViewer (van Eck & Waltman, 2010) and CiteSpace (Chen, 2006). According to the review proposed by Cobo et al. (2011), which focuses on these latter tools together with Bibexcel, CoPalRed, IN-SPIRE, Leydesdorff's programs, Network Workbench Tool, Sci^2 Tool, and Vantage Point, existing software tools offer four different types of bibliometric analyses: burst detection, geospatial analysis, network analysis, and temporal analysis. Among the nine tools reviewed by Cobo et al., network analysis is the most prominent one. Indeed, existing tools primarily allow computing and/or mapping of bibliometric networks of words, authors, and journals derived from co-occurrence data. Geospatial analysis, on the contrary, is the less common type of available analyses. It is included in only three of the reviewed software tools: CiteSpace, Sci^2 Tool, and VantagePoint. The first two allow network data visualisation over a world map using Google Earth Maps or Yahoo! Maps. They also include geocoding capabilities but the choice of the spatial resolution is only dual: either the data are clustered at the country level, or they are clustered at the street level – street addresses used in the publications are then converted to coordinates in latitude and longitude (Chen, 2016). In addition to geocoding and mapping capabilities, NETSCITY, the online application for geographic analysis of research output that we present in this article, features an intermediary level of geographic aggregation, which is the urban area level. Actually, NETSCITY tackles four main issues in the business of mapping bibliographic data:

the geocoding accuracy, the geographic aggregation, the counting/fractioning issue, and the mapping issue.

NETSCITY is a free geoweb application developed to promote the spatial scientometrics method designed by our team to research on the world geography of scientific production and collaboration. Over the past 10 years, our team has been working on geocoding, clustering, and mapping the contemporary geography of scientific activities using bibliometric data. To measure production share and collaboration intensity, we delineated urban area perimeters covering geocoded addresses. This aggregation step proved necessary to work with statistically comparable geographic entities at the world level, which administrative municipalities or street addresses are not. While sharing a set of 495 urban area perimeters in an open access publication (Maisonobe et al., 2018) we have developed NETSCITY, which allows any user:

- to upload a set of references to scientific materials;
- to geocode them;
- to aggregate them at the urban area and the country levels;
- to compute the number of publications per spatial unit and the number of collaborations between spatial units using various counting methods;
- to visualise these variables both on a geographic and on a network map.

At each step, the user is free to download the results of the current process. Some users can be interested in the results of the geocoding process only, others in the results of the aggregation process, others in the results of the counting process, and others in the results of the mapping process. NETSCITY thus fulfils a various number of spatial analytics purposes. We designed it is as a key application to ensure a better reproducibility in the spatial bibliometric field (Frenken et al., 2009; Cobo et al., 2018). We also plan to release the source code of the application.

In this paper, we first discuss the inputs of NETSCITY for the spatial bibliometric field. Then, we describe the inner workings of NETSCITY: from upstream geocoding down to mapping processes. Next, we show the results of NETSCITY when applied to a specific set of references. We conclude with future works involving additional levels of geographic aggregation, counting methods, and mapping options.

Mapping the geography of science

Geocoding

Mapping scientific activities is a five-century-old craft, with seminal maps of science showcased in Börner's (2010) *Atlas of Science*. In 2010, Loet Leydesdorff and Olle Persson (2010) published a landmark article discussing the opportunities of geocoding and mapping Web of Science and Scopus data at the city and institution levels. One year before, Frenken et al. (2009), were identifying a spatial turn in scientometrics that has been confirmed later (Frenken & Hoekman, 2014). The improving capabilities and availability of online geocoding tools, the trend toward territorialised policies of science, and the renewed enthusiasm for global benchmarking of cities and higher education institutions prompted this spatial turn. At the same time, our research team, mainly based in Toulouse (southwestern France), identified a need for strengthening the ties between bibliometrics and geomatics. Processing the spatial information included in bibliometric datasets by relying only on the results of online geocoding tools proved insufficient, such as Google Maps API (Jégou, 2014). Many errors result from trying to geocode the entire address strings included in bibliometric datasets. Let us consider the following example: an author of a 2009 publication reported the following address: "Chang Gung Mem Hosp, Tao Yuan, Taiwan". It corresponds to the Chang Gung Hospital, in Taoyuan District, Taoyuan county of Taiwan Island, 15 km west of the capital, Taipei. Google Maps API hesitates between the hospital itself and the Chang Gung Memorial metro station. Other web geocoding

services confuse this address with several locations in the district. Indeed, the district name is similar to the county name and to the island name.

If the required level of aggregation is the urban level, we proved that it is more relevant to begin by spotting the portion of the address string corresponding to the city, the province / state and the country. The latter is the easiest to find since it always ends the string. As shown in the aforementioned example, tagging the city and the province is a more complicated task because their order of appearance might differ from one address to another and the province is not always specified (non-federal countries such as France do not use them in postal addresses). Even by pre-processing the data to ensure better geocoding results, some errors or missing values still occur: some of them result from homonyms, which require additional spatial information, other result from misspelled toponyms. Here we should highlight that geocoding as we performed it is a semi-automatic process: a geographer expert reviews the results, focussing on the records affiliated to the most prominent scientific cities or the most prone to potential errors (we devoted additional work on addresses having undergone an alphabet transliteration and/or with postal system specificities like Indonesia or Taiwan). After years of geocoding work on the entire contents of the Web of Science Core Collection and on Scopus datasets, our research team improved the geocoding results one obtains when using these databases (Jégou, 2014). Capitalizing on this previous work, the geocoding tool implemented in NETSCITY detects and corrects much of the common misspellings found in Web of Science and Scopus data (by simplifying the given addresses, comparing them with a list of known misspellings variants and, in the absence of a match, by using internal and external online geocoding services) — these stem from authors, publishers, typists working for such database vendors. For the remaining errors, NETSCITY users can amend the records as they see fit (see next section).

Geographic levels of analysis

To study the world geography of science, our team strived to determine an adequate urban level of data aggregation. For studying the geography of science before 2010, most scholars were relying on the country level and only a few studies tackled it at the urban level. The latter has attracted a growing interest since the 2000s following city and regions' empowerment and the multiplication of city rankings (Bornmann & Moya-Anegón, 2018). Two types of studies can be distinguished: those focusing on a limited number of urban regions (e.g., Matthiessen et al., 2010; Hoekman et al., 2010; Nomaler et al., 2014); and those encompassing the publishing localities of the entire world (e.g., Waltman et al., 2011; Pan et al., 2012; Csomós, 2018). In the first case, the authors tend to reuse existing sets of administrative perimeters: United States MSA, European NUTS and, in the second case, the authors do not tend to aggregate the geocoding localities. To get a global panorama of the world scientific production, neither of these two approaches is satisfying. Existing sets of urban area perimeters fulfil specific purposes (whether they are defined for a national or a continental scope, or they are limited to the most populated areas only). Combining existing sets in order to encompass the entire world production might be a solution but it leads to comparative biases since the definitions used to produce these various sets are diverse. Not aggregating the data at all also leads to comparative biases since municipalities' boundaries depend on very heterogeneous criteria differing from one country to another (Maisonobe et al., 2018). For example, not aggregating the data will lead to count apart the numerous publications authored from Bethesda (Maryland, USA) of those of Washington City (DC, USA). Bethesda is a suburb of Washington, which include important medical institutions such as the NIH.

To overcome these issues our team designed a unique set of urban area perimeters encompassing all publishing localities identified after geocoding the entire 1999-2014 contents

of the Web of Science Core Collection. Our urban delineations consider the distribution of the world population density and the Euclidean distance between publishing localities. So far, NETSCITY allows its users to aggregate their data at the level of this set of urban area perimeters as well as at the country level.

Counting method

The next issue addressed by NETSCITY is the counting conundrum. Since about one third of the entire world publications are authored from more than one urban area, one needs to decide how to measure urban areas' contribution to science production. There exist many ways of assigning a number of publications to statistical units. Gauffriau et al. (2008) gives a rich overview of existing methods and highlight the need for more transparency on the methods used since the choice of a counting method significantly influences research findings in bibliometrics. When focusing on the geography of science, the elementary unit of analysis can be the author or the institution, but given the biases of existing sources, it is more often the address. Indeed, in pre-2008 bibliometric records, the list of addresses is not always linked to the list of authors. As a result, it is possible to derive a number of authors per publication as well as a number of addresses but not a number of authors per address per publication. In addition, authors' addresses do not necessarily correspond to authors' institutions since an address may refer to several institutions. Conversely, different addresses may refer to the same institution with varying research teams or university departments. Unless pre-processing the data accordingly, the postal address is the most elementary counting unit of a bibliometric dataset for geographic purpose. As a result, there are two main ways of measuring the scientific production of an urban area: adding up the number of addresses per urban area involved in the publication or considering the number of different urban areas involved in the publication. Similarly, to measure the scientific production of a country, we can consider three different values: the number of addresses per country involved in the publication, the number of urban areas per country involved in the publication, or the number of involved countries. In some disciplines such as in chemistry, only the first address or the corresponding address might be relevant. Other addresses would not be counted in this latter case. In addition to the choice of a counting unit, the counting issue requires to arbitrate between full and fractional countings (Van Hooydonk, 1997). In the first case, we consider the total number of addresses/urban areas/countries while in the second case, we fraction the credit of the publication (i.e., one unit) so that the sum of each fractioned credit total one. Since fractioning avoids counting a single country contribution multiple times, it is the most preferred method in bibliometrics. As shown by Leydesdorff & Park (2017), the choice of a counting method applies both to production data (computing a number of publications per spatial entity) and to collaboration data (computing a collaboration weight between co-authoring entities). With NETSCITY, it is possible to control for the computation of both indicators.

Mapping issue

Mapping issues addressed in bibliometric research are traditionally limited to the visualisation of co-occurrence networks in a relational space. Waltman & van Eck (2010) distinguish two main types of bibliometric maps, namely graph-based maps and distance-based maps. The former refer to mapping the presence or absence of a relation between entities while the latter refer to mapping the distance between entities according to the intensity of their relations. There exists a large range of similarity metrics to compute a relational distance between bibliometric entities: cosine, association strength, Jaccard index, Pearson index and, of course, the raw co-occurrence number (either full or fractional as shown in the previous sub-section) (Boyack et al., 2005; van Eck & Waltman, 2009). When used to produce a distance-based map, similarity metrics can be selected according to different mapping algorithms. Kamada Kawai and

Fruchterman Reingold are the most commonly used techniques in the field of bibliometric as well as in the broader field of network analysis. To adapt to power-law characteristics of most bibliometric distributions, Waltman et al. (2010) propose a unified way of mapping and clustering bibliometric networks. VoSViewer as well as other network analysis software such as Pajek include this method. Compared to this latter issue, mapping issues related to the production of geographic maps are mostly overlooked. The world maps used in spatial bibliometrics are often screenshots taken from Google Earth or Yahoo! Maps' base maps. Instead of benefitting from the capabilities of thematic cartography to map production and collaboration data, there seems to be a preference for online and interactive visualisation. However, we believe the two types of graphic representation – static and interactive – should be valued and different visualisation methods must be chosen accordingly, within the context of geomatics. In this respect, the literature in GIS and cartography about flow maps and the development of geoweb applications might be of relevance (Dodge et al., 2011). While NETSCITY displays interactive geographic and network maps, it also allows the downloading of all underlying data necessary to generate static maps. This feature of the application is about to include more adjustable parameters, interactive linked graphs, and a vector graphics output. In forthcoming developments, NETSCITY will also give access to the Netmap visualisation interface[1] allowing the simultaneous exploration of the geographic and the network maps: by clicking on an item on the graph, the item and its relations are highlighted on the map and vice versa (Maisonobe & Jégou, 2018). As demonstrated in Andurand et al. (2015) and Bach et al. (2015), combining several types of graphic representation enriches the data exploration and analysis experience.

In what follows, we explain how to use the NETSCITY geoweb application and we showcase results obtained by applying NETSCITY to a set of references extracted from the Web of Science.

Features implemented in NETSCITY

In this section, we provide a description of the geoweb application from the data import and processing step to the visualization step. We also discuss technical implementations and directions for improvements throughout the section.

NETSCITY is designed for different information needs and types of users:

- The researchers interested in mapping their own field of research or their own publication record.
- The librarians interested in organizing sets of references and mapping the geographic relations between scientific references.
- The policy officers or international experts interested in scientific evaluation.
- The students or journalists interested in mapping a field of knowledge or the production of a specific institution, author, or place.
- The science studies researchers interested — as we are — in mapping the geography of science.

The home page of NETCITY, presenting the project, is available following this web link: https://irit.fr/netscity

The main hub features four types of actions are available: 1) import a set of references extracted from the Web of Science or Scopus online websites, or personal CSV records, 2) explore the results (lists or data-visualizations) 3) export the results 4) visit the help pages.

[1] Source code of the prototype: https://gitlab.com/ljegou/netmap
Demo: http://www.geotests.net/test/net_map/science_mep.html

Importing a dataset from Web of Science, Scopus or else

In the current version, users can only import one type of Web of Science file. From the Web of Science web interface, they must download a set of references using the following method: 1) Search for a list of records to export and select "Save to Other File Format", 2) Export all records returned (up to 500) by entering "1" to "[the total number of records]", 3) Select "Full Record and Cited References" to include addresses and times cited information, 4) Select "Tab-Delimitated (Mac/Win, UTF-8)" from the "save" drop-down menu. Similarly, for a Scopus file, the procedure is the following: 1) Click on the Export icon[2], 2) From the pop-up menu, select "CSV" format for your file format and "Citation" as well as "Bibliographical information" as the information to export, 3) Click "Export", select "CSV - Only the first 2,000 documents" and click "Export" again.

It is also possible to import a personal CSV file. To do so, users must specify the name of the ID column and the Year column. The name of the columns storing the spatial information is also needed. Several options are possible since the spatial information can be stored:
- In entire postal addresses' strings,
- In three pre-processed columns corresponding to the triples: "city", "province", "country",
- Both in entire addresses' strings (including the institutions) and in pre-processed columns.

The text file can be compressed before the upload, to speed up the transfer. Once one file is uploaded in NETSCITY, it is possible to import another file without deleting the previously uploaded one. This allows us to deal with a bigger dataset and adding the data progressively, coming from various bibliometric sources. According to users' needs, we are eager to improve the import options and open it to other types of sources and formats.

Geocoding and using the error-correction tool

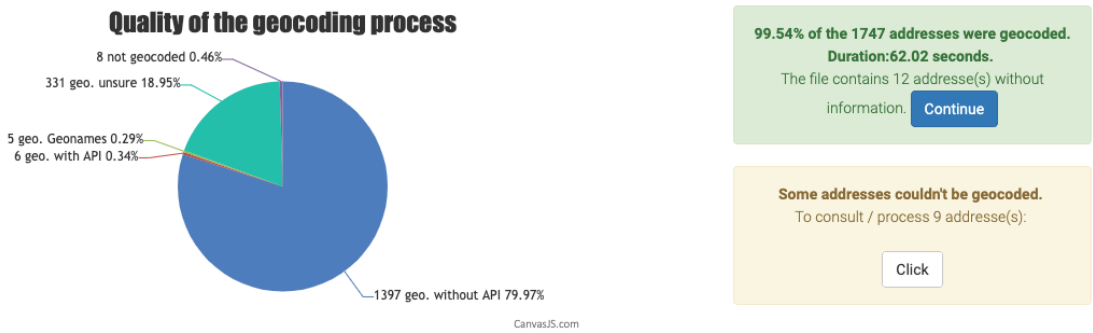

Figure 1. The result of the geocoding process showing a breakdown of the methods used to geocode the addresses from the uploaded addresses

After the import step, the user gets a report detailing the quality of the geocoding process. In Figure 1, the geocoding process has been applied to a homemade CSV file. The bibliometric records included in this file are coming from different sources: Web of Science, Scopus and a French open archive (HAL). As a result, the quality of the geocoding process is poorer than when applied on Web of Science or Scopus files only. However, nearly all the 1,747 addresses included in this file were geocoded. The majority were geocoded with no recourse to an API[3], which means that our internal geocoding knowledge base was sufficient to resolve the

[2] Depending on how you use Scopus, this might be displayed according to your export preference. To select a different export method than what is displayed, click on the drop-down arrow

[3] An external service providing geocoding information.

geographic coordinates. The remaining addresses were geocoded using either a state-of-the-art gazetteer (Geonames), integrated in the application, or an online geocoding API (LocationIQ service).

Identifiant	Adresse	City	Province	Country	Latitude	Longitude	
WOS:000426976100062	VTU, BELAGAVI, INDIA	BELAGAVI		INDIA			Save
WOS:000387985000007	SAVOIE TECHNOLAC, F-73373 LE BOURGET DU LAC, FRANCE	LE-BOURGET-DU-LAC	LAC	FRANCE			
WOS:000366848100332	GRAD SCH ENGN, KAGAWA 7610396, PEOPLES R CHINA			CHINA			
WOS:000374164900016	EXOMARS PROJECT, NL-2201 AZ NOORDWIJK, NETHERLANDS	NOORDWIJK	AZ	NETHERLANDS			
WOS:000380485500366	ENVIRONM OCCUPAT & AGEING PHYSIOL LAB, IXELLE, BELGIUM			BELGIUM			
WOS:000380550000310	DEPT INFORMAT & SISTEMAS, LAS PALMAS DE GC 35017, SPAIN	LAS-PALMAS-DE-GC	GC	SPAIN			
WOS:000367486600022	UJI, CASTELLN 12006, SPAIN			SPAIN			
WOS:000366848100332	FAC ENGN, KAGAWA 7610396, PEOPLES R CHINA			CHINA			

Figure 2. The error-correction interface allowing users to amend the results of the geocoding process

As for non-geocoded addresses, NETSCITY offers the possibility to type some missing spatial information or corrections to complete the geocoding process (Figure 2). Verifying or filling the content of the fields in red (i.e., “city”, “province” and “country”) enhances the geocoding process. If the geocoding fails despite these complements or corrections, it is still possible to add manually the geographical coordinates corresponding to the non-geocoded addresses. In a future version of NETSCITY, we plan to let users modify the result of the geocoding process for geocoded addresses to thwart false positives.

After this first step, users can export the result of the geocoding process, add another file or directly go to the data exploration tools. When choosing to export the results of the geocoding process, NETSCITY generates a list of addresses together with the eventually corrected name of the corresponding cities, provinces and countries (these are sometimes misspelled), the publications ID and the geographic coordinates in a CSV file.

Exploring the dataset through statistical tables and maps

Four types of data exploration are available. The users are presented with:

- Production data through a map and a table listing the number of publications per spatial entity (urban area/country)
- Collaboration data through a map and a table listing the number of collaborations between all pairs of spatial entities (urban areas/countries)

When accessing these different views, the users can change the counting method and unit so that they can immediately see the effects of their choice on the ranking (table view) and on the visualisation (map view).

In the current version, two levels of counting units are available: the users can opt for 1) counting the number of addresses per spatial entity (urban area/country) or 2) counting the number of urban areas or countries. Then, the users are free to choose between full and fractional counting methods. Applied to collaboration data at the urban area level, the fractional counting method implies that by adding the weights of all the links, we obtain the total number of co-written publications of the dataset. It implies weighting the links according to the number of interurban links per co-publication. For instance, if a given publication stems from three different urban areas, each inter-urban link receives 1/3 as a weight for this publication. More generally, if a publication is co-signed from n urban areas, each pair of urban areas (A, B) with A < B is assigned a value l equals to: $1/n(n-1)/2 = 2/(n(n-1))$.
When exploring the numerical data online, all temporal information is overlooked. Longitudinal analysis is available at the export stage, however: the option "with the evolution by years" computes the different variables (number of publications and number of collaborations) on an annual basis. Then, the export final step allows, on the one hand, opting for full or fractional counting and, on the other hand, selecting between CSV or JSON formats.

Availability of NETSCITY as an open source application

The source code is to be released in an online open-source forge (like GitHub) with a suitable license (like one compatible with the well-used GNU-GPL), after several enhancements and another pass of testing and debugging. The project is still ongoing, and we are eager to collect feedback and ideas to expand its capabilities, to tune it to new needs, and to enhance interoperability with other analysis applications such as the CorTexT project[4] or VOSViewer. Together with several bibliometrics tools, the geocoding tool and the aggregation tool included in NETSCITY could be available as off-the-shelf modules.

Geographical underpinning of a set of bibliographic references: a case study using NETSCITY

To test NETSCITY on a specific set of publications, we decided to focus on the publications referring in their titles, abstracts or keywords to the use of a common device: ROV or AUV, which is a submarine robot. Therefore, we issued the following topic search in the *Web of Science Core Collection*: TS= ("ROV" OR "AUV") AND "robot*" AND "underwater", all publication years combined. This query returned 1,004 records published between 1991 and 2017 (Figure 3). We consider this WOS result in the running example that follows.

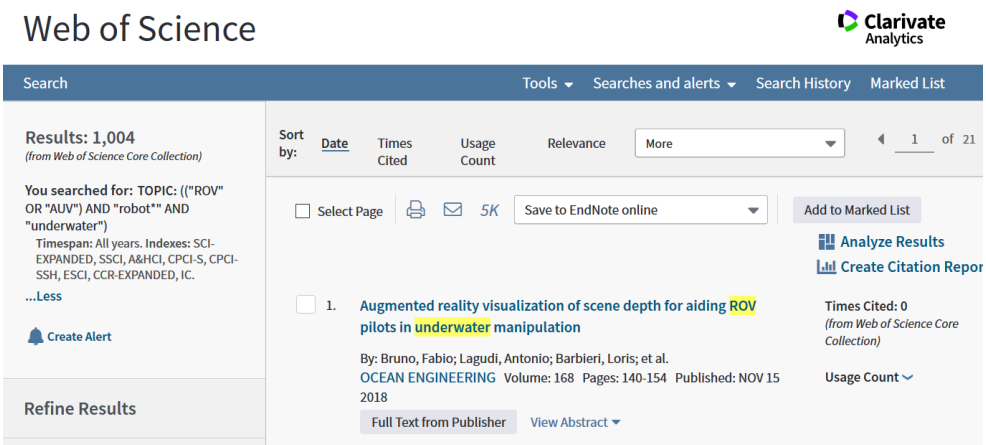

Figure 3. Result page for a sample Web of Science query with 1,004 records

[4] https://www.cortext.net

Although the *Web of Science* interface cannot reveal the leading urban areas producing knowledge on the matter under consideration, NETSCITY lists the top 10 urban areas which publish the most about ROV/AUV (Table 1) at once.

Table 1. Top ten urban areas which publish the most in the ROV/AUV field (Web of Science)

Urban area	*Country*	*Full number of publications*
Tokyo	Japan	44
Boston	United-States	41
Genoa	Italy	41
Shanghai	China	39
Girona	Spain	37
Pisa	Italy	30
Florence	Italy	28
Woods-Hole	United-States	26
Edinburgh	United-Kingdom	24
Singapore	Singapore	23

Then, we can identify the top 10 most intense interurban collaborations about ROV/AUV according to a different counting method (Table 2).

Table 2. Top ten most intense interurban collaboration in the ROV/AUV field (Web of Science)

Urban area 1	*Country 1*	*Urban area 2*	*Country 2*	*Fractional number of collaborations*
Florence	Italy	Pisa	Italy	30.0
Waterloo-Guelph	Canada	Shanghai	China	16.0
Genoa	Italy	Pisa	Italy	14.0
Florence	Italy	Genoa	Italy	9.5
Geelong	Australia	San-Antonio	United-States	9.0
Oslo	Norway	Trondheim	Norway	8.5
Kitakyushu	Japan	Tokyo	Japan	8.0
Brest	France	Shanghai	China	7.5
Lisbon	Portugal	Porto	Portugal	7.0
Woods-Hole	United-States	Baltimore	United-States	6.5

In addition, we can interactively explore the world production map and the world collaboration map per urban area and per country using different counting methods. Figure 4 and Figure 5 are screenshots taken from NETSCITY cartographic zoomable views.

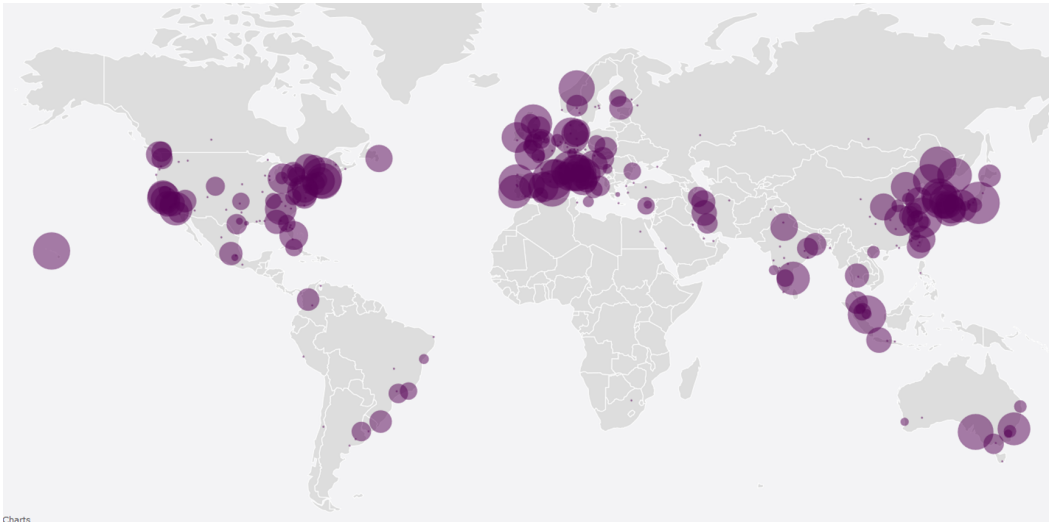

Figure 4. Production map for the 1,004 WOS records of the running example, after geocoding and aggregation are performed. Fractional counts at the urban area level.

From these four views (we could also have chosen to discuss the networks views available in NETSCITY), we can notably observe that the ROV/AUV topic involves publications coming from all over the world, but mainly from maritime cities (Figure 5). We can also notice an important Italian cluster of scientific collaborations on this research issue (Table 2). It is worth noting here that we obtained all these results with a few clicks only. NETSCITY constitutes an effective application to efficiently and quickly visualise and explore the geography of a bibliometric dataset according to different levels of analysis and counting methods.

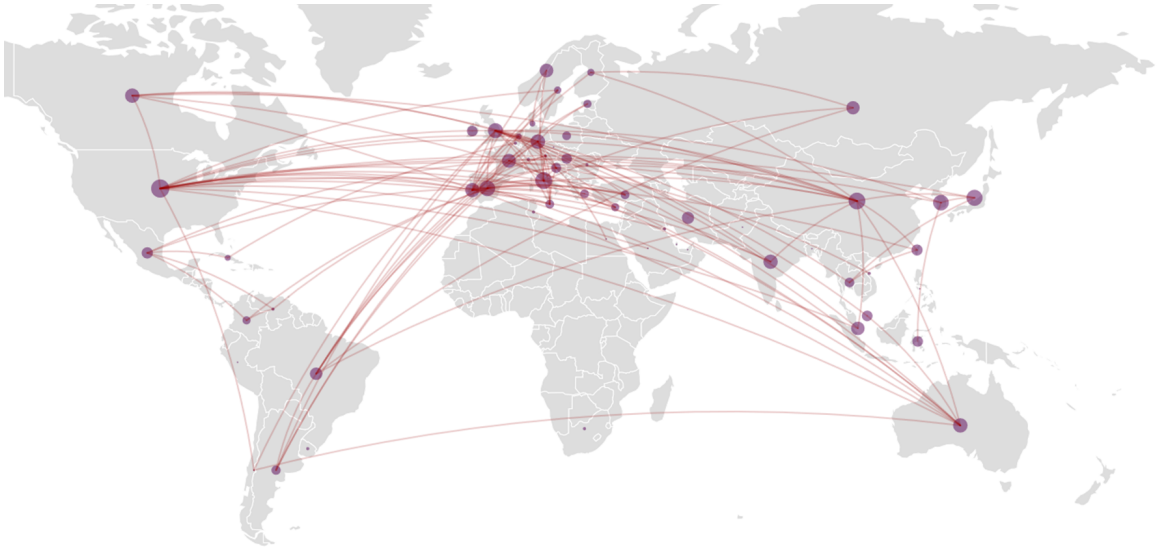

Figure 5. Collaboration map for the 1,004 WOS records of the running example, after geocoding and aggregation are performed. Fractional counts at the country level.

Conclusion

The field of bibliometrics already develops and provides a variety of software to process bibliometric data without requiring any programming skills. However, most existing software fail to account for the geographical complexity of bibliometric data, which leads to misleading results. In addition, many applications generate graph-based maps of locations but, to the best of our knowledge, there is no fully integrated solution to parse and clean the affiliations recorded in a set of references, geolocalising them, clustering them at the requested level of

analysis, and mapping them both on a world base map and in a relational space. The goal of NETSCITY is to offer and promote such a solution.
Concisely, NETSCITY is a new application designed for the general public to perform spatial bibliometrics at the click of a mouse. It computes metrics and produces maps abiding by the state-of-the-art methods of geography of science. We wish that NETSCITY contributes to overcome the pitfalls of spatial data and better inform policy makers regarding territorialised policies worldwide.

Acknowledgments

We would like to thank our entire geography of science research team, and especially Béatrice Milard, Michel Grossetti, and Denis Eckert. This research is supported by LABEX SMS (ANR-11-LABX-0066) under project codenamed Netscience.

References

Andurand, A., Jégou, L., Maisonobe, M., & Sigrist, R. (2015). Les mondes savants et leur visualisation, de l'Antiquité à aujourd'hui. *Histoire et Informatique*, 18/19, 59–94.

Bach, B., Riche, N. H., Fernandez, R., Giannisakis, E., Lee, B., & Fekete, J.-D. (2015). NetworkCube: bringing dynamic network visualizations to domain scientists. In *Posters of the Conference on Information Visualization (InfoVis)*.

Börner, K. Atlas of Science: Visualizing What We Know. Cambridge, MA: MIT Press.

Bornmann, L., & de Moya-Anegón, F. (2018). Spatial bibliometrics on the city level. *Journal of Information Science*, 0165551518806119. https://doi.org/10.1177/0165551518806119

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351–374. https://doi.org/10.1007/s11192-005-0255-6

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Chen, C. (2016). *CiteSpace: A Practical Guide for Mapping Scientific Literature*. Nova Publishers.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, Analysis, and Cooperative Study Among Tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402.

Cobo, M. J., Dehdarirad, T., García-Sánchez, P., & Moral-Munoz, J. A. (2018). Quantifying the reproducibility of scientometric analyses: a case study. In *STI 2018 Conference Proceedings* (pp. 925–933). Leiden: Centre for Science and Technology Studies (CWTS).

Csomós, G. (2018). A spatial scientometric analysis of the publication output of cities worldwide. *Journal of Informetrics*, 12(2), 547–566. https://doi.org/10.1016/j.joi.2018.05.003

Dodge, M., McDerby, M., Turner, M., (2011). Geographic Visualization: concepts, Tools and Applications, Wiley.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Science of Science: Conceptualizations and Models of Science*, *3*(3), 222–232. https://doi.org/10.1016/j.joi.2009.03.005

Frenken, K., & Hoekman, J. (2014). Spatial Scientometrics and Scholarly Impact: A Review of Recent Studies, Tools, and Methods. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact* (pp. 127–146). Cham: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-10377-8_6

Gauffriau, M., Larsen, P., Maye, I., Roulin-Perriard, A., & Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147–176. https://doi.org/10.1007/s11192-007-1934-2

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*(5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012

Jégou, L. (2014). Toward spatially referenced academic data at global scale: the full geocoding of Wos-Datasets, methods and results. Presented at the 2nd Geography of Innovation International Conference, Utrecht.

Leydesdorff, L., & Persson, O. (2010). Mapping the Geography of Science: Distribution Patterns and Networks of Relations among Cities and Institutes. *Journal of the American Society for Information Science and Technology*, *61*(8), 1622–1634. https://doi.org/10.1002/asi.21347

Leydesdorff, L., & Park, H. W. (2017). Full and Fractional Counting in Bibliometric Networks. *Journal of Informetrics*, *11*(1), 117–120.

Maisonobe, M., Jégou, L., & Eckert, D. (2018). Delineating urban agglomerations across the world: a dataset for studying the spatial distribution of academic research at city level. *Cybergeo: European Journal of Geography*, e871. https://doi.org/10.4000/cybergeo.29637

Maisonobe, M. & Jégou, L., (2018). Explorer les réseaux mondiaux : proposition d'outil interactif combinant graphe (diagramme nœuds-liens) et carte de flux. Presented at the CIST2018 - Représenter les territoires / Representing territories, Rouen.

Matthiessen, C. W., Schwarz, A. W., & Find, S. (2010). World Cities of Scientific Knowledge: Systems, Networks and Potential Dynamics. An Analysis Based on Bibliometric Indicators. *Urban Studies*, *47*(9), 1879–1897. https://doi.org/10.1177/0042098010372683

Nomaler, Ö., Frenken, K., & Heimeriks, G. (2014). On Scaling of Scientific Knowledge Production in U.S. Metropolitan Areas. *PLOS ONE*, *9*(10), e110805. https://doi.org/10.1371/journal.pone.0110805

Pan, R. K., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, *2*. https://doi.org/10.1038/srep00902

van Eck, N. J., & Waltman, L. (2009). How to Normalize Co-occurrence Data? An Analysis of Some Well-Known Similarity Measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651. https://doi.org/10.1002/asi.21075

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Infometrics*, *4*(4), 629–635. https://doi.org/10.1016/j.joi.2010.07.002

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

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, *48*(10), 944–945. http://doi.org/d9f6jz

Waltman, L., Tijssen, R. J. W., & Eck, N. J. van. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, *5*(4), 574–582. https://doi.org/10.1016/j.joi.2011.05.003