Geospatial Discovery in Collections of Text

Dan C. Baciu[1,2], Sunit Kajarekar[1], and Anna Abramova[3]

[1]Architektur Studio Bellerive, Bern, Switzerland
[2]TU Delft, South Holland, Netherlands
[3]University of California Berkeley, Berkeley, USA

Geospatial discovery enables people to search for "hotels nearby," "museums nearby," or perform other similar geospatial queries. We extend this concept to collections of text. Our present article comes with a prototype of a geospatial discovery tool for text that facilitates queries such as "news that deals with the area around this museum" and "news that connects the area near the museum with the one near this hotel." Through these and similar examples, we introduce geospatial discovery in collections of text, discussing its broad relevance. Furthermore, we also demonstrate how geospatial queries in collections of text can be used to formulate metrics for the assessment of cultural connectivity in urban spaces. In particular, we coin the terms "geospatial iso-lex" and "cultural avenues." They are discussed in more detail in the article.

1 Geospatial Discovery, Isolexes, and Cultural Avenues

Geospatial discovery is becoming increasingly important (Burrough et al., 2015, Shekhar and Chawla, 2003, Wang and Goodchild, 2018). Many search engines support it. For example, geospatial discovery is used by people to explore where to go shopping, what events are taking place nearby, or where their upcoming vacations could take them. Typical queries include "bars near me," "events near the Eiffel Tower," or "hotels in Rome." Geospatial discovery goes beyond analysis or exploration; it helps people find what they are looking for by supporting geospatial queries and selection.

This present article builds on a series of presentations and preprints in which we have introduced a new idea: Geospatial Discovery in Collections of Text (Baciu, 2019, 2020b,c, 2021b, Baciu and Kajarekar, 2022, Baciu et al., 2023). Our innovation is to expand geospatial discovery to textual data such as news, books, or any other data written in natural language. Note that we are primarily interested in the content of the texts. For example, if a book mentions the "Eiffel Tower" somewhere in the text, we enrich the text with the geolocation of the Eiffel Tower. This information is then used towards geospatial discovery.

To expand geospatial discovery to textual data, we introduce two types of geospatial query: (1) queries for texts that deal with one contiguous search area, e.g., "news that deals with the area near the Eiffel Tower" and (2) queries for texts that deal with multiple independent search areas, e.g., “news that deals both with the area near the Eiffel Tower AND the area near the Louvre.” At DH Benelux 2023, we provided a fully functional prototype of a tool that can perform these queries. Multiple version are available at Baciu and Kajarekar (2022, 2023a,b).

Our tool responds to the two queries mentioned above by displaying a list of texts that fulfill the search criteria. Additionally, a map is displayed to geographically visualize the search results. This map makes our interface multimodal, combining text and map. Before our work, many library interfaces were primarily text based, yet we believe that multimodality is an added quality. Specifically, the geospatial discovery tool that we showed at DH Benelux integrates geospatial queries with string-matching, thematic, and multifaceted queries.

To understand the workings of our tool, it is relevant to grasp the idea that each text that one reads builds not only a thematic but also a geospatial narrative. Together the geospatial narratives of multiple texts connect different parts of the world, often in unexpected ways. Building on this insight, we visualize how texts connect the world, and we use this information to support geospatial discovery with the queries that we have already mentioned. For the first type of query, the map that is returned by our interface shows how texts connect the search area with the rest of the world. For the second type of query, the map shows how texts connect the two search areas with each other and with the rest of the world.

Our **G**eospatial **D**iscovery Tool for **T**extual Data (GDT) represents an innovation of significant practical value, as illustrated by the following concrete examples:

1. With the geospatial discovery capabilities that we have developed, any library patron can formulate geospatial queries. This capability is crucial, for example, for scholars who believe in the importance of diversifying their research and studying geographical areas that have long been understudied. With a geospatial discovery tool, such scholars can both identify which areas have been understudied (geospatial analysis) and direct their attention towards texts that have dealt with these areas (geospatial discovery).

2. Geospatial discovery is useful to many different groups of both scholars and practitioners. A group of practitioners that benefits from our tool is architects. The work of architects is often dedicated to specific urban areas. With a GDT, architects can now easily find books, news, or social media posts that deal with the area that they are presently working on. Finding and reading this news and other textual data will provide architects with cultural insights that would be difficult to obtain otherwise, and yet are relevant in making design decisions. In this context, see also Kuneva (2024). Our tool makes it easy to design a building not only in response to the built fabric around it, but also considering the soft, cultural dimensions of urban space.

3. GDTs are beneficial not only for scholars and practitioners, but also for the geographers themselves. One task of geography and environmental science is to study how pollution is distributed, see for example Abramova (2023). In this context, the question arises of how environmental awareness responds to pollution. A study of how people lived with pollution and how environmental awareness emerged in response is found in Baciu and Abramova (2020). Such a study could benefit from a GDT. Through geospatial discovery, geographers gain new capabilities when they explore how social media or news discusses pollution. For example, one can now directly search for news about areas in which the level

of contamination peaks, and one can study whether this news accurately reflects the type and degree of contamination present in the environment.

Our idea to expand geospatial discovery to text does not come out of thin air. In traditional library settings, people occasionally used traditional string-matching or thematic searches as a workaround to indirectly perform geospatial discovery. For example, a full-text search for "Paris" returned texts that mentioned this city. Yet, technically speaking, this query is not a geospatial query. When one searched for "Paris" or "Montmartre" through string-matching, one found texts that literally mentioned "Paris" and "Montmartre," but one did not also find texts that mentioned the "Moulin de la Galette," although the latter is located in Montmartre, Paris. What was missing was the ability to formulate actual geospatial search criteria, such as "texts that deal with the area I have drawn on the map." Combined with conventional string-matching and thematic search engines, the advantage of a GDT is that it allows all user groups (students, scholars, researchers, practitioners, consumers, lay audiences, etc.) to perform geospatial queries consciously and directly.

Since we first proposed the idea of a GDT, some libraries have started to implement maps on their websites. We feel that this is a partial success. Nevertheless, the success is not complete. Library websites that show maps still use text metadata for mapping purposes. For example, they tell you where a text was published, and visualize this on a map. However, we feel that it is even more important to (1) provide geospatial discovery capabilities, i.e., let people formulate geospatial queries, (2) provide these discovery capabilities on the content of the texts, not only on metadata; there is a difference between a text that was published in Paris and one that actually discusses Paris, (3) visualize the geospatial narrative of each text towards analysis, exploration, and discovery.

Beyond its practical value, our tool is opening new research directions. Specifically, we hope that our ideas will raise awareness of the importance of the cultural dimensions of urban space. To illustrate this, we have introduced two quantitative measures, which we shall refer to as "geospatial isolex" and "cultural avenue."

An isolex is the cultural equivalent of an isochrone. The difference is that isochrones are used as a measure for urban mobility, whereas isolexes can serve as a measure for cultural connectivity in urban space. Isochrones and isolexes are computed in the following manner: A 15-minute isochrone from the Eiffel Tower is the answer to the question how far people can get from the Eiffel Tower by taking all modes of transportation 15 minutes in all directions. By contrast, a 15-sentence isolex from the Eiffel Tower is the answer to the question how far people can get from the Eiffel Tower by opening all available books at the spot where they mention the Eiffel Tower and reading 15 sentences. Let us define the isolex as the set of geolocations that can be extracted from these sentences. As mentioned, isochrones give information about mobility, whereas isolexes give information about cultural connectivity. Isochrones are about traveling, whereas isolexes are about the places that people can get to virtually, by reading. The GDT that we have developed can automatically compute isolexes.

In principle, isolexes can be defined and visualized in many alternative ways. One can speak of a 15-sentence isolex, as mentioned before. One can also speak of abstract-length isolexes, to give another example. An abstract-length isolex from the Eiffel Tower selects all abstracts that mention the Eiffel Tower and collects all geolocations that can be extracted from these textual data. It is also possible to define isolexes in terms of time. A 1-minute isolex from the Eiffel Tower would open all books to the spot where they mention the Eiffel Tower, and instead of reading a fixed number of sentences, it reads a fixed amount of time in each book. This approach combines textual and temporal analysis. Perhaps the result would be aptly called an isoclex, combining the Greek words "chronos" for time, as in chronometer or chronograph, with the word "lex" for speaking, as in "lexicon."

Multiple isolexes can also be combined into a joint query. We have already mentioned that our tool can perform queries of the type "texts that deal with the area near the Eiffel Tower AND the area near the Louvre." The result of such a query unites two isolexes with an AND operation. An important outcome of this query is that it displays cultural connections between the two areas. We call these connections "cultural avenues."

Discovering "cultural avenues" in urban space may be of particular importance for urban planners and architects as well as urban activists. Planners, activists, and architects often encounter segregation. As Thomas Schelling has demonstrated, segregation is often undesirable. People do not wish to be segregated, but segregation nevertheless takes place. It can be described as a problem of large-scale coordination in large populations of individuals (Baciu, 2015, 2023, Baciu et al., 2022, Shelling, 1971). Given this problem of large-scale coordination, it is common that planners, activists, and architects are entrusted with the goal of reducing segregation. Typically, segregation can be reduced by strengthening the physical connections between segregated areas. For example, if there is a street or avenue that connects the two areas, one can help reduce segregation by embellishing and enhancing it. With the tool that we have developed, it is now possible to go beyond the enhancement of streets and avenues. With a mouse click, one can suddenly discover which cultural avenues connect segregated areas. A click may sometimes not even be needed, as the process of discovery can be automated, letting a computer calculate cultural avenues and automatically detect cultural segregation. The same computer can then make suggestions on how to alleviate the segregation it has discovered. The role of the human in this process becomes to evaluate the computer's automated suggestions. Planners, activists, and architects can thus rely on early-warning systems to discover cultural segregation more easily and strengthen both physical and the cultural avenues between segregated areas.

To summarize, we have developed a geospatial discovery tool for text. The tool is of practical value for libraries and people, and it has also led us to formulate new fundamental research concepts. To complement the existing concept of "isochrones," which is used in urban mobility, we have formulated the concepts of "geospatial isolexes" and "cultural avenues," which can be used to study cultural connectivity in urban space.

Now just think big! Geospatial discovery for textual data together with the ability to compute isolexes and cultural avenues will open countless new research opportunities. Everything that has been done with isochrones can now also be done with isolexes, and the results that are obtained show a new, cultural perspective on cities.

Cities are not only about hard physical objects such as streets and buildings. Above all, cities are about people and their cultures. Devoid of people, cities are worthless hazard zones. Thus, it is very important to be able to study urban cultures. Our tool and the research concepts that we have developed provide researchers with new capabilities and with a quantitative approach to many soft, cultural questions about urban space.

In the spirit of developing a digital tool of practical utility and explaining theories through practical applications, the rest of this article is written in an object-oriented style. The following sections are structured as follows: First we formulate the problem of geospatial discovery for textual data. Then, we explain our motivation and approach to solving the problem. Third, we demonstrate how the tool works, and we do this based on three concrete examples. Finally, we explain how the tool has led us to formulate the two concepts of "geospatial isolex" and "cultural avenues," which we visualize with an additional example. The article concludes with additional technical considerations, a final discussion, and our offer to provide support to implement geospatial discovery for text for any library, company, or interest group that reaches out to us.

To facilitate interactive engagement with our article, we have created a GPT-chatbot in our "Conversing with Science" series. The chatbot can be accessed at https://doi.org/10.25496/W2301K

2 Problem statement and solution

If one visits a library with the intention of reading a book or an article, one is rarely asked what areas around the world one is interested in, and even if one comes with a particular geospatial interest, it is very hard to let it flow into library queries. Our point of departure is the desire to change this situation. We provide a geospatial discovery tool for textual data. To support patrons, our tool considers both metadata and the actual written text.

The geospatial discovery tool has in particular the following capabilities:

1. It can read text and geolocate the places, institutions, well-known infrastructures, buildings, monuments, and the like, that are mentioned in the text.

2. The geospatial information that is extracted has fine-grained (street-level) resolution. This makes the tool useful for people who wish to study particular areas of cities or landscapes, especially areas that do not have a name, but have easily definable geographical ranges. For example, the "1km square area around the Eiffel tower" is an area that is easy to draw on a map, but it has no commonly known name.

3. Our tool solves problems of polysemy and homonymy. For example, New York can refer to both the city and the state, which is ambiguous, and it can also be written as NY or NYC, which are homonyms when used to refer to the city. Our solution in this context builds on the outstanding work of Biemann et al. (2007), Cheng and Roth (2013), Mihalcea and Csomai (2007), Pilehvar et al. (2013), Roth et al. (2014), Sil et al. (2018).

4. We provide an interactive, multimodal, online interface that allows people to define geospatial queries, combining geospatial search criteria with a multi-faceted search. Typically, library queries are text-based, but we felt motivated to develop a multimodal query that uses text in combination with maps, to make the searching experience more engaging to users.

5. The interface visualizes the geographical selection criteria and the selected texts on the map. Each text is represented as a geospatial narrative—a curved gradient line that passes through all geolocations that can be extracted from the text.

Typically, the end-goal of geospatial discovery is to identify a location such as that of a bar, grocery, supermarket, hotel, or event. In our application, the end-goal is different. The GDT is not looking for a book's physical geolocation and does not help users locate a physical copy of the book. Instead, the tool uses geospatial search criteria to discover textual data that mention entities such as infrastructures, monuments, or institutions that fulfill certain geospatial search criteria. In this sense, the end-goal is not to find a geolocation, but to use geospatial search criteria to find a text.

3 Motivation and approach

Since the early 2010s, libraries have become increasingly digitized. Due to this transition, texts written in books have become increasingly available in digital format. However, libraries have been rather slow in evaluating these digitized texts and using textual analysis to help patrons search for new readings. We pressed changing this situation early on, when few people understood why textual analysis matters. In 2015, we proposed algorithms for libraries to process and analyze their text Baciu (2016). Meanwhile, algorithms similar to the one that we proposed have been implemented in many a library. We would now like to look ahead and make another step forward by proposing to facilitate multimodal searches that unite text with non-textual modes of representation. The present tool unites text and geospatial discovery.

In our continuing work, we have gained extensive experience with geographical information retrieval from text. We have used a supercomputer to process more than 50,000,000 pages of books and periodicals that mention the term "Chicago school" Baciu (2016, 2017, 2018a,b, 2021b). We have also processed roughly 200,000 news items and 1,000,000 social media posts about science and humanities Baciu (2020a), Liu et al. (2022), as well as other material Baciu (2021a), Baciu and Cellucci (2022), Baciu et al. (2023). The information obtained from processing these data was used for example to map how the idea of the Chicago school has spread over the world, and how the humanities and science are perceived in the United States and abroad. We have developed the present tool in parallel to such research. Previous versions of our tool can be found in our previous articles.

4 How does the tool work?

Most libraries allow patrons to do some kind of multi-faceted search: patrons can narrow their queries by focusing on specific publication years and places, they can specify whether they want to search books or articles, and they can perform full-text queries on the textual content. All of the queries just mentioned are text-based.

What we now introduce is a map. On the map, each book is represented as a gradient line. The first vertex on this line is the publication place of the book. All other vertices are things such as cities, institutions, buildings, or birthplaces of famous people that are mentioned in the text. These latter data are obtained through our method of geographical information retrieval first published in Baciu (2020a).

The main geospatial discovery capability that we provide consists of giving patrons the possibility to narrow down their search based on geospatial criteria. In addition to other criteria they use in their search, they can now specify a geographical search window, and they can use this window to further narrow down their search.

In theory, the geographical search window can have any shape. To make our tool easy to work with but also powerful, we let users create up to two squares. This choice is arbitrary and easy to change.

To further advance the idea of multimodal searches, our tool can also be applied to allow users to search texts on a timeline that reflects the age of the entities that are mentioned in the text. We shall call this capability “historical discovery in collections of text.” It is also a new idea that we are introducing. An application of this type of analysis and discovery is found in Baciu (2020a)

5 Examples

Figures 1, 2 and 3 illustrate a practical demonstration of our tool, each displaying a distinct collection of texts acquired through previous collaborations. In all figures, our geospatial discovery tool represents each text as a gradient line on the map. We have chosen the colors of the gradient lines as follows: each line starts in white, after which it becomes green, yellow, and red. White is chosen for the publisher location. The gradient from green to red is for the textual content. We represent the content in reading direction. The gradient line is green where the text starts and red where it ends. The longer the text, the more saturated is the red hue. The illustrations demonstrate how geospatial queries are used to narrow search results.

6 New methods and concepts

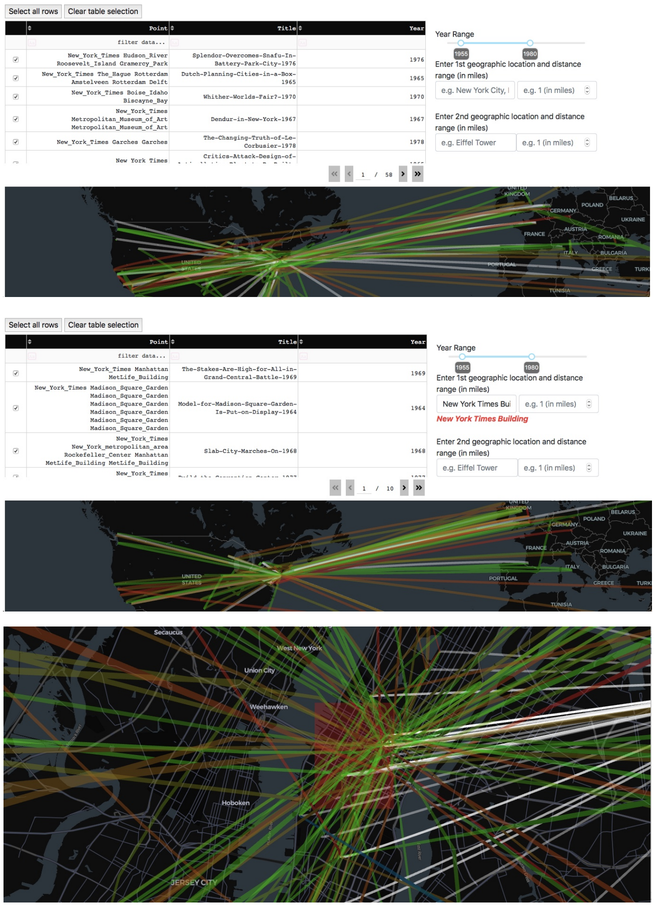
The geospatial discovery tool that we developed not only empowers people to perform multimodal and geospatial queries; it also opens new research directions. In particular, new research concepts emerge in the study of urbanization and cultural geography. The maps shown in Figure 3 can already help us develop two new concepts.

In Figure 3, when we selected “Chicago” and “Washington DC,” we created two visuals. Actually, these visuals are more than just visuals; they are a type of analysis. Let us explain the value of this analysis by comparing it to another well-known type of analysis—the isochrones analysis.

An isochrone shows how far one can get from a given point in space, given a fixed amount of traveling time, for example, 5 minutes. Imagine someone starts in the geographical center of Chicago. A 5-minute isochrone shows how far one can travel in this amount of time. Of course, one can specify what type of transportation is used for the analysis. One could choose to use private or public transportation or both.

Our analysis performs the equivalent of the isochrones analysis, but with text. The difference is that we are not using a transportation system. Instead, we evaluate how far one can get from the center of Chicago by reading textual data. In the given example in Figure 2, the maximum distance is one page, while in Figure 3, the maximum distance one is reading is an abstract, and one reads each abstract in a corpus of dissertations. Thus the transportation system is replaced in our example by a collection of texts. The traveling time is replaced by a distance in text. Other distances for our analysis could be 5 sentences, 5 paragraphs, 1 chapter, etc. Let us call this type of analysis a “geospatial isolex.” We use “iso-“ as in isochrone and “lex” as in lexicon. Isolexes of this kind are analyses in the family of isochrones, isotherms, isonomes, or isolines more generally. In Figure 4 we discuss several sample isolexes visualized with our tool.

Now let us attract your attention to Figure 3, below. This figure is also not just a visual but a type of analysis. In the example given in the figure, we query two geographical selection windows at once: “Chicago AND Washington.” Only texts are shown that pass through both windows. What is interesting about this analysis



	Point	Title	Year
	filter data...		
☑	New_York_Times Hudson_River Roosevelt_Island Gramercy_Park	Splendor-Overcomes-Snafu-In-Battery-Park-City-1976	1976
☑	New_York_Times The_Hague Rotterdam Amstelveen Rotterdam Delft	Dutch-Planning-Cities-in-a-Box-1965	1965
☑	New_York_Times Boise_Idaho Biscayne_Bay	Whither-Worlds-Fair?-1970	1970
☑	New_York_Times Metropolitan_Museum_of_Art Metropolitan_Museum_of_Art	Dendur-in-New-York-1967	1967
☑	New_York_Times Garches Garches	The-Changing-Truth-of-Le-Corbusier-1978	1978
	New York Times	Critics-Attack-Design-of-	

Year Range: 1955 – 1980

Enter 1st geographic location and distance range (in miles): e.g. New York City, | e.g. 1 (in miles)

Enter 2nd geographic location and distance range (in miles): e.g. Eiffel Tower | e.g. 1 (in miles)

	Point	Title	Year
	filter data...		
☑	New_York_Times Manhattan MetLife_Building	The-Stakes-Are-High-for-All-in-Grand-Central-Battle-1969	1969
☑	New_York_Times Madison_Square_Garden Madison_Square_Garden Madison_Square_Garden Madison_Square_Garden Madison_Square_Garden Madison_Square_Garden	Model-for-Madison-Square-Garden-Is-Put-on-Display-1964	1964
☑	New_York_Times New_York_metropolitan_area Rockefeller_Center Manhattan MetLife_Building MetLife_Building	Slab-City-Marches-On-1968	1968
	New_York_Times		

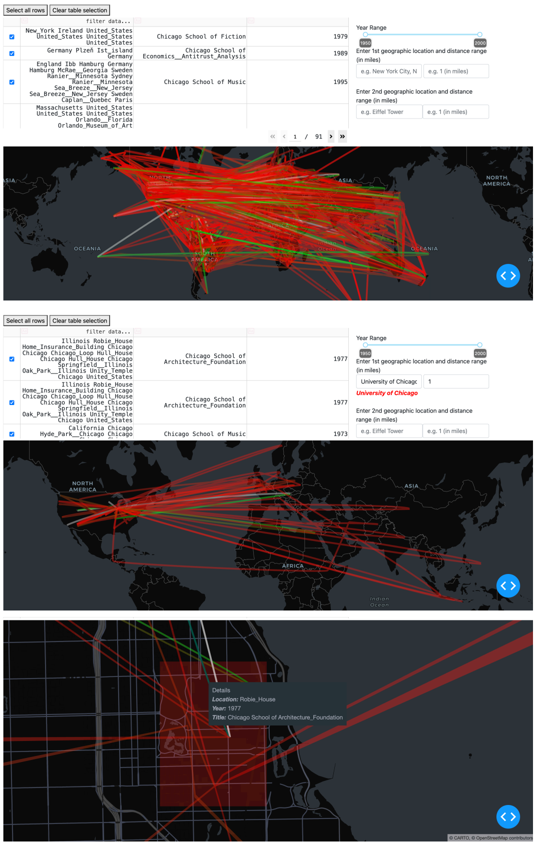
Year Range: 1955 – 1980

Enter 1st geographic location and distance range (in miles): New York Times Bui | e.g. 1 (in miles)

New York Times Building

Enter 2nd geographic location and distance range (in miles): e.g. Eiffel Tower | e.g. 1 (in miles)

Figure 1: This figure shows a collection of some 500 news articles written by Ada Louise Huxtable (compiled by staff of the Getty Research Institute in Los Angeles, which holds the Ada Louise Huxtable Papers, 1859–2013, 2013.M.9, https://www.getty.edu/research/collections/collection/113YNR). We begin by showing all articles, after which we narrow the search window onto a square area roughly 1 mile around the New York Times building. We also show a zoomed-in detail in New York. Note that the number of articles is diminished as the search is narrowed. Given that our geographic information retrieval method has high resolution, one can choose small geographical search windows, such as shown here. In addition, one can also combine geospatial queries with other conventional text-based queries, and one can select individual records from the table of records. The visuals shown in Figure 1 are interactively available at https://doi.org/10.25496/W26P4J Geospatial Discovery Tool for Ada Louise Huxtable's articles, Baciu and Kajarekar (2023a)



Select all rows | Clear table selection

	filter data...		
☑	New_York Ireland United_States United_States United_States United_States	Chicago School of Fiction	1979
☑	Germany Plzeň Ist_island Germany	Chicago School of Economics__Antitrust_Analysis	1989
☑	England Ibb Hamburg Germany Hamburg McRae__Georgia Sweden Ranier__Minnesota Sydney Ranier__Minnesota Sea_Breeze__New_Jersey Sea_Breeze__New_Jersey Sweden Caplan__Quebec Paris	Chicago School of Music	1995
	Massachusetts United_States United_States United_States Orlando__Florida Orlando_Museum_of_Art		

1 / 91

Year Range: 1950 – 2000

Enter 1st geographic location and distance range (in miles): e.g. New York City, N | e.g. 1 (in miles)

Enter 2nd geographic location and distance range (in miles): e.g. Eiffel Tower | e.g. 1 (in miles)

Select all rows | Clear table selection

	filter data...		
☑	Illinois Robie_House Home_Insurance_Building Chicago Chicago Chicago_Loop Hull_House Chicago Hull_House Chicago Springfield__Illinois Oak_Park__Illinois Unity_Temple Chicago United_States	Chicago School of Architecture_Foundation	1977
☑	Illinois Robie_House Home_Insurance_Building Chicago Chicago Chicago_Loop Hull_House Chicago Hull_House Chicago Springfield__Illinois Oak_Park__Illinois Unity_Temple Chicago United_States	Chicago School of Architecture_Foundation	1977
☑	California Chicago Hyde_Park__Chicago Chicago	Chicago School of Music	1973

Year Range: 1950 – 2000

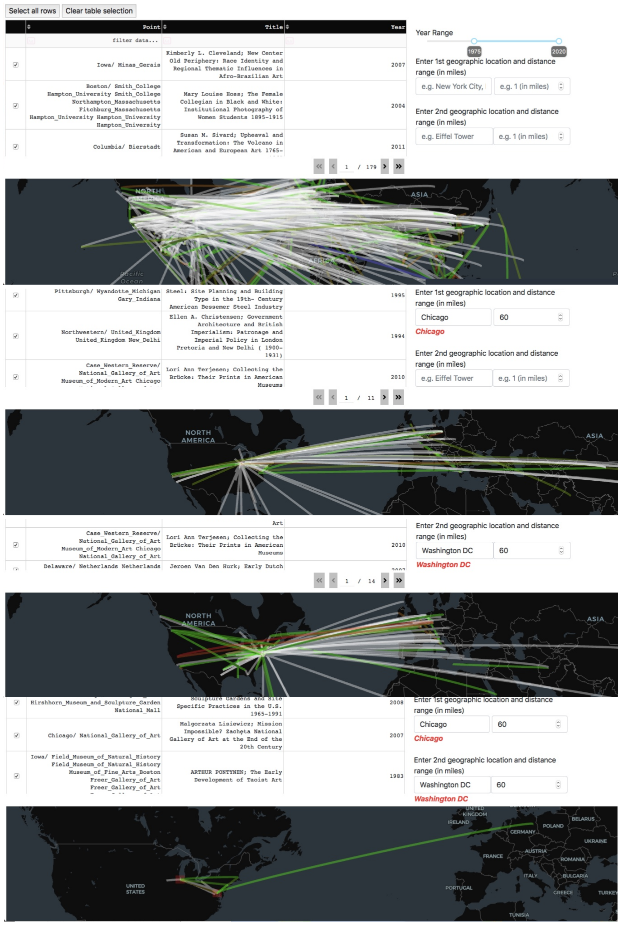
Enter 1st geographic location and distance range (in miles): University of Chicagc | 1

University of Chicago

Enter 2nd geographic location and distance range (in miles): e.g. Eiffel Tower | e.g. 1 (in miles)

Details
Location: Robie_House
Year: 1977
Title: Chicago School of Architecture_Foundation

Figure 2: This figure shows a collection of several thousand records that mention Chicago schools of architecture, fiction, theology, art, music, and several other fields. The collection is taken from Baciu (2017). We demonstrate how geospatial discovery works, following the same procedure as in Figure 1. We begin by showing all texts, after which we narrow the search window onto a square area roughly 1 mile around the University of Chicago. We also show a zoomed-in detail in Chicago. Note that the number of texts is diminished as the search is narrowed down. Also, one can hover with the cursor over a line, which identifies the textual record for this particular line, here a line leading to the Robie House, located near the University of Chicago, which has served as center for the area used in the geospatial query. The texts in this collection are much longer than in either the Ada Louise Huxtable or the dissertation corpus, which is why the lines are predominantly red.



Select all rows | Clear table selection

	Point	Title	Year
	filter data...		
☑	Iowa/ Minas_Gerais	Kimberly L. Cleveland; New Center Old Periphery: Race Identity and Regional Thematic Influences in Afro-Brazilian Art	2007
☑	Boston/ Smith_College Hampton_University Smith_College Northampton_Massachusetts Fitchburg_Massachusetts Hampton_University Hampton_University Hampton_University	Mary Louise Hoss; The Female Collegian in Black and White: Institutional Photography of Women Students 1895-1915	2004
☑	Columbia/ Bierstadt	Susan M. Sivard; Upheaval and Transformation: The Volcano in American and European Art 1765-	2011

« ‹ 1 / 179 › »

Year Range 1975 2020

Enter 1st geographic location and distance range (in miles)

e.g. New York City, | e.g. 1 (in miles)

Enter 2nd geographic location and distance range (in miles)

e.g. Eiffel Tower | e.g. 1 (in miles)

	Point	Title	Year
☑	Pittsburgh/ Wyandotte_Michigan Gary_Indiana	Steel: Site Planning and Building Type in the 19th- Century American Bessemer Steel Industry	1995
☑	Northwestern/ United_Kingdom United_Kingdom New_Delhi	Ellen A. Christensen; Government Architecture and British Imperialism: Patronage and Imperial Policy in London Pretoria and New Delhi (1900-1931)	1994
☑	Case_Western_Reserve/ National_Gallery_of_Art Museum_of_Modern_Art Chicago	Lori Ann Terjesen; Collecting the Brücke: Their Prints in American Museums	2010

« ‹ 1 / 11 › »

Enter 1st geographic location and distance range (in miles)

Chicago | 60

Chicago

Enter 2nd geographic location and distance range (in miles)

e.g. Eiffel Tower | e.g. 1 (in miles)

	Point	Title	Year
		Art	
☑	Case_Western_Reserve/ National_Gallery_of_Art Museum_of_Modern_Art Chicago National_Gallery_of_Art	Lori Ann Terjesen; Collecting the Brücke: Their Prints in American Museums	2010
☑	Delaware/ Netherlands Netherlands	Jeroen Van Den Hurk; Early Dutch	2007

« ‹ 1 / 14 › »

Enter 2nd geographic location and distance range (in miles)

Washington DC | 60

Washington DC

	Point	Title	Year
☑	Hirshhorn_Museum_and_Sculpture_Garden National_Mall	Sculpture Gardens and Site Specific Practices in the U.S. 1965-1991	2008
☑	Chicago/ National_Gallery_of_Art	Malgorzata Lisiewicz; Mission Impossible? Zachęta National Gallery of Art at the End of the 20th Century	2007
☑	Iowa/ Field_Museum_of_Natural_History Field_Museum_of_Natural_History Museum_of_Fine_Arts_Boston Freer_Gallery_of_Art Freer_Gallery_of_Art	ARTHUR PONTYNEN; The Early Development of Taoist Art	1983

Enter 1st geographic location and distance range (in miles)

Chicago | 60

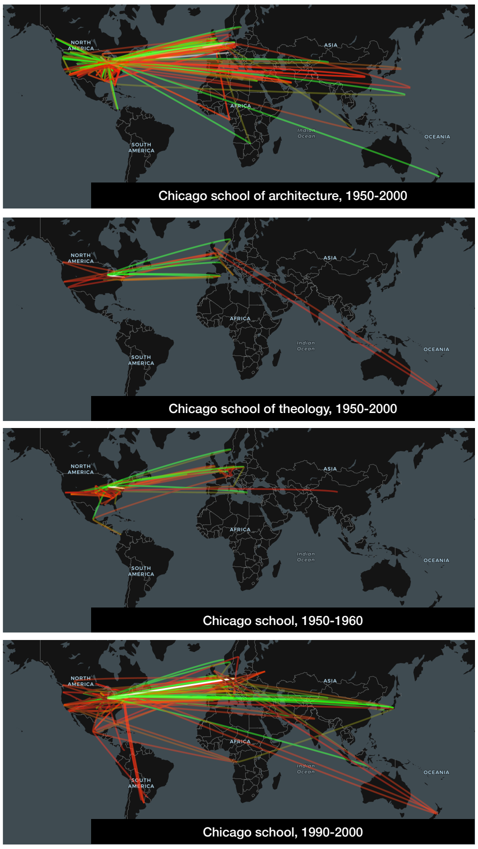
Chicago

Enter 2nd geographic location and distance range (in miles)

Washington DC | 60

Washington DC

Figure 3: This figure shows a collection of art history dissertation abstracts, initially collected and provided by Adams and Lucarelli (2021), and further enhanced by staff of the Getty Research Institute in Los Angeles (Um, 2020, Um and Hagen, 2021). We begin again by showing the mapping results for all texts. The search window is then narrowed, first onto "Chicago," then onto "Washington DC," and finally onto both "Chicago AND Washington," together. The square query areas of roughly 60 miles (as selected by the user) are highlighted in red on the map. This corpus has only abstracts, hence the predominance of white and green, compared to the previous visual that has longer texts. The visuals shown in Figure 3 are interactively available at https://doi.org/10.25496/W2BC7T Geospatial Discovery Tool for Art History Dissertation Abstracts, Baciu and Kajarekar (2023b)


Chicago school of architecture, 1950-2000

Chicago school of theology, 1950-2000

Chicago school, 1950-1960

Chicago school, 1990-2000

Figure 4: Comparison of multiple isolexes. The figure shows four 10-step isolexes, in a corpus about the Chicago school that we created as part of our earlier research Baciu (2017, 2021b). The isolexes all start at University of Chicago and consider the ten following geolocatable entities. As before, the lines start green and end red. The publishers are again drawn in white. The comparison of the two isolexes above shows that texts that mentioned the Chicago School of Architecture have connected the globe more broadly than those that mentioned the Chicago School of Theology. The isolexes below show that the Chicago School overall, supported more cultural connectivity in 1990-2000 than it did in 1950-1960, in the corpus under consideration. A similar analysis can also be performed with social networks, rather than with text, calculating how far one can get through several degrees of friendship (also known as small-world-problem, discussed for example by Watts (1999). We shall call the analysis through degrees of social networks in geographical space an isonex analysis, to differentiate it from isolexes.

is that it shows cultural connections between the two areas. We shall call this set of connections a "cultural avenue."

The analysis of cultural avenues may be particularly relevant for reconnecting urban areas that are segregated. Segregation is a common phenomenon in urbanism. Often, it emerges on its own, with detrimental effects for the populations of the segregated areas. Although the majority of the population of segregated areas may collectively favor living without segregation, the process of segregation is hard to stop or reverse without some kind of coordination (Baciu et al., 2022, Shelling, 1971). Architects and urban planners are often entrusted with the task of supporting such coordination. They may try to stop segregation by revitalizing streets and avenues that connect the segregated areas.

A cultural avenue analysis may now provide new ways to reconnect areas. Through cultural avenue analysis, it is possible to identify cultural connections between two areas. Once the connections are known, one can try to strengthen them, as one would do it with physical connections, in order to reconnect segregated areas.

Together, isolexes and cultural avenues are examples of new research concepts that our geospatial discovery tool has inspired. We hope for many more such concepts to come.

7 Implementation for libraries and online services

Let us ask: Should geospatial discovery become a common feature in libraries? We say "Yes!" Compare just how attractive the front pages of "Google Explore" and "Google Things To Do" look compared to "Google Books" or "Google News." We feel that the latter pages could also provide geospatial discovery (Figure 5).

When presenting the idea of geospatial discovery, we have proceeded in an object-based manner, often talking about libraries and patrons. Yet the geospatial discovery tool presented here may prove most useful when used by companies online. We would like to encourage them to further develop our idea and to contact us to support product development.

We are equally happy to help forward-looking or traditional libraries implement our tool or develop new versions of it. We'll try to support anyone who contacts us. The requirement is that the textual data are available in digitized format, or that we may contact the data provider to arrange an agreement. In previous work, we have developed safe practices to work with copyrighted data and may not even need access to the data ourselves. For this approach see also Organisciak and Downie (2021).

8 Technical considerations

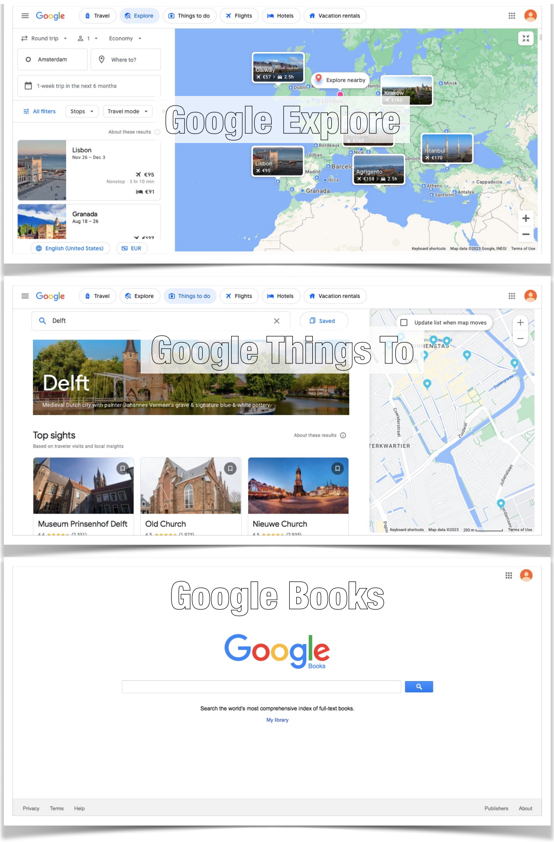
To create a tool for geospatial discovery in collections of text, as we propose, the following components are necessary:

1. Access to a collection of text.

2. Geocoding, to provide geolocations for entities that are mentioned in the text.

3. Storage of the processed data or parts thereof.

4. An algorithm or machine that can perform queries.



Medieval Dutch city with painter Johannes Vermeer's grave & signature blue-&-white pottery

Search the world's most comprehensive index of full-text books.

Figure 5: Comparison of the front pages of Google Explore, Google Things to do, and Google Books. The first two make excellent use of geospatial discovery. We would like to suggest and support the implementation of geospatial discovery for books and news, too.

5. An interactive interface that lets users formulate queries and receive the results from their queries.

6. Or, instead of this interface, a tool that can formulate the queries on behalf of a user and let the query results serve the user with or without an interface. (For example an AI chatbot can execute geospatial discovery on behalf of a user, or a search for "Culture near me" could display not only cultural venues but also cultural news articles on the map, even when a query for news has not been consciously formulated.)

The prototype that we have developed performs geocoding through Natural Language Processing with enhanced use of geodata from a large knowledge base (Wikipedia). The accuracy of the Natural Language Processing tasks is discussed by Roth et al. (2014), Tsai et al. (2016). Given this setup, old place names and institution names are considered, as are alternative names or abbreviations. We pioneered this approach in Baciu (2020a). The Natural Language Processing identifies named entities based on both the name that is mentioned in the text (surface shape), and the context in which this name is mentioned. It is also possible to do the geocoding with standard methods of geographic information retrieval, relying solely on gazetteers, without Natural Language Processing, or to combine both types of methods.

If the geocoding accuracy achieved on a corpus is not sufficient (for example because the text quality is low), we recommend considering whether it helps displaying items only if they have more than one entity located in the query window. This leads to an increase in accuracy. For example, if the code has a 0.1 error rate for each geocoded item, the probability that two items are both inaccurately geocoded, independently, is 0.1 times 0.1 = 0.01. Thus, if one lists query results by their relevance for the selected area, the first results are overall correct even if the geocoding accuracy is low due to poor text quality.

From a technical perspective, we expect that future GDT implementations will be available on much larger datasets. In addition, an interesting area of future research will be the application of continual and multilingual learning, as discussed in Praharaj and Matveeva (2023), to keep the code behind geospatial discovery up-to-date and multicultural.

9 Definitions

Our article introduces the idea of geospatial discovery for textual data, and it suggests two quantitative measures for urban connectivity. To facilitate communication about these measures, we suggest to use the terms with the meaning listed below.

Geospatial discovery: We use here the term "geospatial" because we let users formulate geospatial queries with geospatial search criteria such as geospatial distances from a given location in geographical space. We use the term "discovery" because through the queries, users can discover material. This goes beyond analysis or exploration. An example of geospatial discovery is to support users to search and select bars near the Eiffel Tower. Geospatial exploration and analysis differ from geospatial discovery. Compared to geospatial discovery, geospatial exploration does not need to let users perform queries and selection. Geospatial exploration is performed in any map that visualizes geospatial information for users to explore. Geospatial exploration becomes geospatial discovery when geospatial queries are introduced. Geospatial analysis, on the other hand, is about analyzing, not discovering. An example of geospatial analysis is the question how many of the bars near the Eiffel Tower provide access to the internet, and whether the geospatial distribution of these bars reflects knowledge of designers that the Eiffel Tower has long served as an antenna. Like geospatial exploration, geospatial analysis is not focused around geospatial queries and selection.

Geospatial discovery for textual data: We speak of geospatial discovery for textual data when geospatial discovery is facilitated for textual data. Geospatial discovery for textual data should not stop at providing information where a physical copy of a book is located or where the book was published. Instead, geospatial discovery for textual data should support users to find texts that have been written about user-defined, geographical areas of interest. A text that says "There was a celebration near the Eiffel Tower" is a text about the Eiffel Tower. The author may be located in New York and the text may have been published in the San Francisco Bay Area. At the same time, the text is about an entity located in France.

Geospatial isolex: Given a collection of textual data, given a certain area of departure (for example a circular area of 500 meters around the Eiffel Tower), and given a textual distance (for example 10 sentences), an isolex answers to the question how far one can get from the area of departure by mining the collection of textual data for items that are located in the area of departure, reading the given textual distance in the texts, and collecting the geocodes of all entities found in these textual data. Technically, an isolex is a set of connected points and/or polygons. It is easy to see that these points can be represented as network, as points, or, using a scale of analysis, as density map, or, analyzing equal density zones in the density map, the same data can be represented with isolines. We call the analysis an isolex independent of how it is visualized. Let us suggest terminology with respect to the various types of visualization: We suggest to use the terms "geospatial isolex network" when visualizing the network with points and lines, as we have done it in this article, "geospatial isolex density map" when visualizing the same data as a density map, "geospatial isolex heat map" when visualizing the data as heat maps, etc. We suggest "geospatial isograph" for an isoline used to visualize isolex data. Please feel free to shorten these terms when using them.

Isoclex: An isolex defined by measuring reading distance in time (for example, a 1 min. isoclex). The term is derived from uniting isochrone and isolex.

Isonex: An analysis of geospatial connectivity through degrees of connection in social or other networks. As the isolex, an isonex is a set of connected points. It can be represented as network, as points, as density map, or with isolines.

Cultural avenue: Given a collection of textual data, and given multiple geographical areas, a cultural avenue is a set of textual connections between the two areas. Cultural avenues are a special case of isolex analysis. Cultural avenues can be represented as network, as points, and therefore also as density map, or with isolines.

Social avenue: The equivalent of a cultural avenue, but calculated on a social network, rather than through textual data. It can be represented as network, as points, as density map, or with isolines.

Geospatial isograph: We suggest to use this term for the isolines used to represent isolexes or isothemes. We derive the term from "iso-" and "-graph," with the latter as in the Greek word for writing.

Isotheme: Given a collection of textual data, given a theme or topic (for example textual fragments about physical science in this collection, or words pertaining to a topic of discourse), and given a particular geospatial scale of analysis, an isotheme is the equivalent of an ecological isonome, it is any line along which the topic is covered at roughly constant density. We derive the term from "iso-" and "-theme," with the latter as in the Greek word for subject or theme. Examples of isotheme analyses visualized as heat maps or as networks are found in Baciu (2020a), Baciu et al. (2022). Specifically, we have shown that most topics in the public discourse have distinct geospatial footprints or isothemes, representable as networks, heatmaps, density maps, or with geospatial isographs; examples are found in Baciu (2020a).

Historical discovery for textual data: This is the equivalent of geospatial discovery for textual data, but instead of formulating geospatial search criteria, one formulates historical search criteria. Examples for a historical query is “texts that have dealt with the period 1830-1869” or “texts that have dealt with entities that are have been created in the period 1830-1869” or “texts that have primarily dealt with entities in one of the years 1830-1889.” Our geospatial discovery tool facilitates this type of discovery.

10 Conclusion

We have developed a geospatial discovery tool for textual data. Library patrons can now find their readings in a multimodal way using text and maps. We believe that this practical advancement is significant, changing how people do science and how they read in a globalized world. Geographic information that previously remained hidden in the text is now visible on maps, which can help readers more easily direct their attention to geographical areas of their choice.

The tool that we have developed can be applied in standard library settings as much as it is relevant online or for chatbots. News and social media platforms as well as AI application providers could use our tool or self-made GDTs either for similar purposes as libraries or integrated in automated recommendation systems. Furthermore, our tool (or derivate versions of our idea) can be used for geospatial discovery for text in any imaginable setting as well as in languages other than English.

As many people desire, it is now possible to actively direct research efforts and public attention to areas around the world that are understudied or receive insufficient public attention. This capability changes an entire perspective towards human cultural production and its geospatial discoverability. In addition, we believe that the new tool will also lead researchers to work with new concepts such as geospatial isolexes and cultural avenues to reconnect segregated areas and support mutual understanding among people who live in different geographical areas.

11 Summary

If you are planning to travel around the world and are searching for potential hotels to stay at, you can often expect that the search engine that you may use asks you about your destination right away, only later continuing to other travel details such as arrival and departure dates. Even before those latter questions are answered, you are presented with hotel options, and you get very quickly accustomed to comparing your options on a map. The map thus becomes a tool for geospatial discovery. A different situation is encountered if you want to read books from around the world. Unlike the tourism industry, libraries have not yet implemented geospatial discovery. In this article, we introduce a first tool for geospatial discovery for textual data. Our tool reads and interprets textual data, maps it, and helps you find your books on the map. If you search books about, say, a particular area of San Francisco, our tool can help you find books in which well-known buildings, institutions, and inhabitants have been mentioned that have a tie to this area. In a globalized world, we hope that geospatial discovery for text will support a new consciousness for the geospatial dimensions of cultures. At DHB2023, our audience was encouraged to test the tool during our presentation held in the Royal Library of Belgium, Brussels, Belgium, June 2, 2023 (Baciu and Kajarekar, 2022)).

Besides introducing new geospatial discovery capabilities, our tool also opens new research paths in urban and cultural geography. We hope that our quantitative measures for cultural connectivity will lead to new discoveries and geospatial concepts, especially among researchers who have been working in the field independently. Interesting examples of recent and ongoing work that have reached us through peer reviewers are Bodenhamer et al. (2013), Meijers and Peris (2019), Tongjing et al. (2023). There could be much more awareness of the cultural dimensions of the built and natural environments.

12 Contributions

Dan C. Baciu: first author. Sunit Kajarekar: equally contributing author, programmed the interface. Anna Abramova: support at revision stage. The support of the Getty Research Institute in Los Angeles and its staff during project development is thankfully acknowledged.

References

A. Abramova. Environmental assessment of coal mining practices in svalbard: Insights from soil, vegetation, and coal elemental analysis. In *International Multidisciplinary Scientific GeoConference: SGEM*, volume 23(5.1), pages 113–121, 2023. doi: 10.5593/sgem2023/5.1/s20.14.

C.D. Adams and C.J. Lucarelli. Art history dissertations and abstracts from North American institutions, 2021. URL `https://openpublishing.psu.edu/ahd/content/art-history-dissertations-and-abstracts-north-american-institutions`. Compiled dataset.

D.C. Baciu. Systemic analysis of cooperation: Architects, urban space and tourism. *JETK*, 4:27–30, 2015. URL `https://escholarship.org/uc/item/81h50276`.

D.C. Baciu. Sigfried Giedion: Historiography and history of reception on a global stage. In *Ar(t)chitecture*, pages 40–52. Technion Israel Institute of Technology Haifa, 2016. URL `https://escholarship.org/uc/item/01g1t2fw`.

D.C. Baciu. The Chicago School: Evolving systems of value, 2017. URL `https://escholarship.org/uc/item/5zt9859m`.

D.C. Baciu. From everything called Chicago school to the Theory of Varieties, 2018a. URL `https://escholarship.org/uc/item/19c227c3`.

D.C. Baciu. The Chicago School: Large-scale dissemination and reception. *Prometheus*, 2:20–43, 2018b. URL `https://escholarship.org/uc/item/22v9g5mn`.

D.C. Baciu. The geography of cultures: New methods for decoding, analysis, and synthesis, 2019. Spatial Tech Lunch Series 2019-2020, UCSB Department of Geography.

D.C. Baciu. Cultural Life: Theory and empirical testing. *BioSystems*, 197:104208, 2020a. URL https://doi.org/10.1016/j.biosystems.2020.104208.

D.C. Baciu. Cultural diversification processes and the culture radar, 2020b. URL https://escholarship.org/uc/item/4993b2vq.

D.C. Baciu, 2020c. Research proposal for geospatial discovery in collections of text. TU Delft.

D.C. Baciu. Culture and people flow together, 2021a. URL https://doi.org/10.31219/osf.io/u69s5. OSF preprints.

D.C. Baciu. Mapping Chicago Schools, 2021b. URL https://doi.org/10.31219/osf.io/zce2w. OSF preprints.

D.C. Baciu. Causal models, creativity, and diversity. *Humanities and Social Sciences Communications*, 10:134, 2023. doi: 10.1057/s41599-023-01540-1. URL https://doi.org/10.1057/s41599-023-01540-1.

D.C. Baciu and A. Abramova. Svalbard's Arctic settlements: From mining sites to urbanized environments. In *EGU General Assembly Conference Abstracts*, volume 22, page 2458, 2020. URL https://meetingorganizer.copernicus.org/EGU2020/EGU2020-2458.html.

D.C. Baciu and V. Cellucci. Paths of wind and of research, 2022. URL https://doi.org/10.31219/osf.io/h4g6p. In Exhibition: Atomic Reactions, TU Delft Library 2022. OSF preprints.

D.C. Baciu and S. Kajarekar. Geospatial discovery in collections of text, 2022. URL https://doi.org/10.31219/osf.io/ywqu9.

D.C. Baciu and S. Kajarekar. Geospatial discovery in collections of text, 2023a. URL https://doi.org/10.25496/W26P4J. DH Benelux.

D.C. Baciu and S. Kajarekar. Geospatial discovery in collections of text: Art history dissertation abstracts, 2023b. URL https://doi.org/10.25496/W2BC7T. DH Benelux.

D.C. Baciu, D. Mi, C. Birchall, D. Della Pietra, L. Loevezijn, and A. Nazou. Mapping diversity: From ecology and human geography to urbanism and culture. *SNSS*, 2: 136, 2022. URL https://doi.org/10.1007/s43545-022-00399-4.

D.C. Baciu, S. Kajarekar, and U. Gunes. Constructal flow of constructal thinking. In Bejan A. Lucia U. Grisolia G. Gunes, U. and A. Morega, editors, *Proceedings of the 12th Constructal Law Conference, Politecnico di Torino, 2023*. Yildiz Technical University, 2023. URL https://escholarship.org/uc/item/14479810.

C. Biemann, I. Matveeva, R. Mihalcea, and D. Radev. *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*. Association for Computational Linguistics, 2007. URL https://aclanthology.org/W07-0200.

D. Bodenhamer, J. Corrigan, and T.M. Harris. Deep mapping and the spatial humanities. *International Journal of Humanities and Arts Computing*, 7:170 – 175, 2013.

P.A. Burrough, R.A. McDonnell, and C.D. Lloyd. *Principles of Geographic Information Systems*. Oxford University Press, USA, 2015.

X. Cheng and D. Roth. Relational inference for wikification. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1787–1796, 2013. URL `https://api.semanticscholar.org/CorpusID:17784265`.

R. Kuneva. Space and culture: Mapping culture and culture resources. In *Foundation for Humanities and Social Research - Sofia*. 2024.

A. Liu, A. Droge, S. Kleinman, L. Thomas, D.C. Baciu, and J. Douglass. What everyone says: public perceptions of the humanities in the media. *Daedalus*, 151:19–39, 2022. URL `https://doi.org/10.1162/DAED_a_01926`.

E. Meijers and A. Peris. Using toponym co-occurrences to measure relationships between places: review, application and evaluation. *International Journal of Urban Sciences*, 23(2):246–268, 2019.

R. Mihalcea and A. Csomai. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, pages 233–242. ACM, 2007. doi: 978-1-59593-803-9/07/0011.

P. Organisciak and J.S. Downie. Research access to in-copyright texts in the humanities. In *Information and knowledge organization in digital humanities: global perspectives*, pages 157–177. Routledge, 2021. URL `https://doi.org/10.4324/9781003131816-8`.

M.T. Pilehvar, D. Jurgens, and R. Navigli. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351. Association for Computational Linguistics, 2013.

K. Praharaj and I. Matveeva. Multilingual continual learning approaches for text classification. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 864–870. INCOMA Ltd., Shoumen, Bulgaria, 2023. URL `https://aclanthology.org/2023.ranlp-1.93`.

D. Roth, H. Ji, M. Chang, and T. Cassidy. Wikification and beyond: The challenges of entity and concept grounding. In *ACL Tutorial Abstracts*, volume 7, 2014.

S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.

T.C. Shelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1: 143–186, 1971.

A. Sil, H. Ji, D. Roth, and S. Cucerzan. Multi-lingual entity discovery and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, volume 5, pages 22–29, 2018.

W. Tongjing, E. Meijers, Z. Bao, and H. Wang. Intercity networks and urban performance: a geographical text mining approach. *International Journal of Urban Sciences*, pages 1–22, 2023.

C.T. Tsai, S. Mayhew, and D. Roth. Cross-Lingual Named Entity Recognition via Wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 219–228. Association for Computational Linguistics, 2016.

N. Um. What do we know about the future of art history? part 1: Let's start by looking at its past, sixty years of dissertations. *CAA Reviews*, 2020. doi: 10.3202/caa.reviews.2020.74. URL `https://doi.org/10.3202/caa.reviews.2020.74`.

N. Um and E. Hagen. What do we know about the future of art history? part 2: Dissertations since 1980. *CAA Reviews*, 2021. doi: 10.3202/caa.reviews.2021.57. URL `https://doi.org/10.3202/caa.reviews.2021.57`.

S. Wang and M.F. Goodchild. *CyberGIS for Geospatial Discovery and Innovation*, volume 118 of *GeoJournal Library*. Springer, 2018. ISBN 978-3-319-96550-5.

D.J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999.