Geospatial discovery in collections of written text

Authors: **Dan C. Baciu & Sunit Kajarekar**

Affiliations: TU Delft, Delft, Netherlands; Architektur Studio Bellerive, Bern, Switzerland.

Keywords: Natural Language Processing; Geographic Information Retrieval; Geospatial Analysis; Geospatial Discovery; Libraries.

Abstract

When you travel around the world and search for potential hotels to stay at, the search engine that you use may first ask you about your destination, 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 fast accustomed to comparing your options on a map. The map thus becomes a tool for geospatial discovery. A different situation is encountered when 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 Seattle, our tool can help you find books in which famous buildings, institutions, and inhabitants have been mentioned that have a tie to this area. At DHB2023, our audience will be encouraged to test the tool during our presentation (DHB2023, Mont des Arts 28, 1000, Brussels, Belgium, June 2. 2023, 14:15-15:30).

Problem statement

When you want to read a book or a journal article, and you visit a library, you are rarely asked what areas around the world your book or article should cover, and even if you come with a particular geospatial interest, it is very hard to let it flow into your library searches. We would like to change this situation by providing a spatial discovery tool for textual data. Our tool evaluates both metadata and the actual written text.

Our tool has in particular the following capabilities:

1) It can read text and extract geographic information for place names, institution names, famous buildings, famous people, and the like.

2) The geographical information that is extracted has 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.

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 is a case of multiple terms that refer to the same entity (Roth et al. 2014, Sil et al. 2018).

Our approach

Since the early 2010s, libraries have become increasingly digitized. The texts written in books are increasingly available in digital format. However, libraries have been rather slow in evaluating this text and using textual analysis to help patrons find their books. We have pressed on changing this situation early on, when few people understood why textual analysis matters. In 2015, we were rejected from multiple conferences when we proposed algorithms that libraries could use 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 go a step further.

In our continuing work, we have gained extensive experience in geographical information retrieval from text. We have used a supercomputer to process 50,000,000 pages of books and periodicals that mentioned the term "Chicago school" (Baciu 2017, 2018, 2019, 2022). Later we processed roughly 200,000 news items and 1,000,000 social media posts about science and humanities (Baciu 2020, Liu et al. 2022), as well as other material (Baciu 2021, Baciu and Cellucci 2022). We mapped how the idea of the Chicago school spread over the world, and how the humanities and science are perceived in a globalized world. This working experience helped us develop the present tool. Previous versions of our tool can be found in our previous articles.

How does the tool work?

Most libraries allow you to do some kind of multi-faceted search: you can search publication years and places; you can specify whether you search books or articles; and you can use full-text search on the textual content.

What we now introduce is a map. On the map, each book is represented as a gradient line that connects many points. The first point is the publication place of the book. All other points are things such as cities, institution, buildings, or birthplaces of famous people that are mentioned in the text. These latter data are collected through our method of geographical information retrieval with street-level resolution.

The main geospatial discovery capability that we provide consists in giving you the possibility to narrow down your search based on geospatial criteria: in addition to other criteria you use in your search, you can specify a geographical search window, and you can use this window to narrow down your search.

Example

Figure 1 shows a practical example. In this visual, we allow you to perform a search in a collection of news articles written by architectural critic Ada Louise Huxtable. Our geospatial discovery tool represents each news article 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 red, purple, green, and grey. This sequence of colors advances in reading direction of the text.

Figure 1A shows us narrowing in the geographical window, such that only news articles are shown that mention at least one item that falls in the search window 30 to 70 degrees of latitude and -25 to -90 degrees of longitude. Figure 1B, shows us further narrowing down the window to 30 to 40 degrees of latitude and -65 to -90 degrees of longitude. Note how the number of articles is greatly diminished, as the search is narrowed in. We can then click on a line, which highlights an article in our list, or we can click on a row in the list, which highlights the article on the map. This is shown if Figure 1C & D.

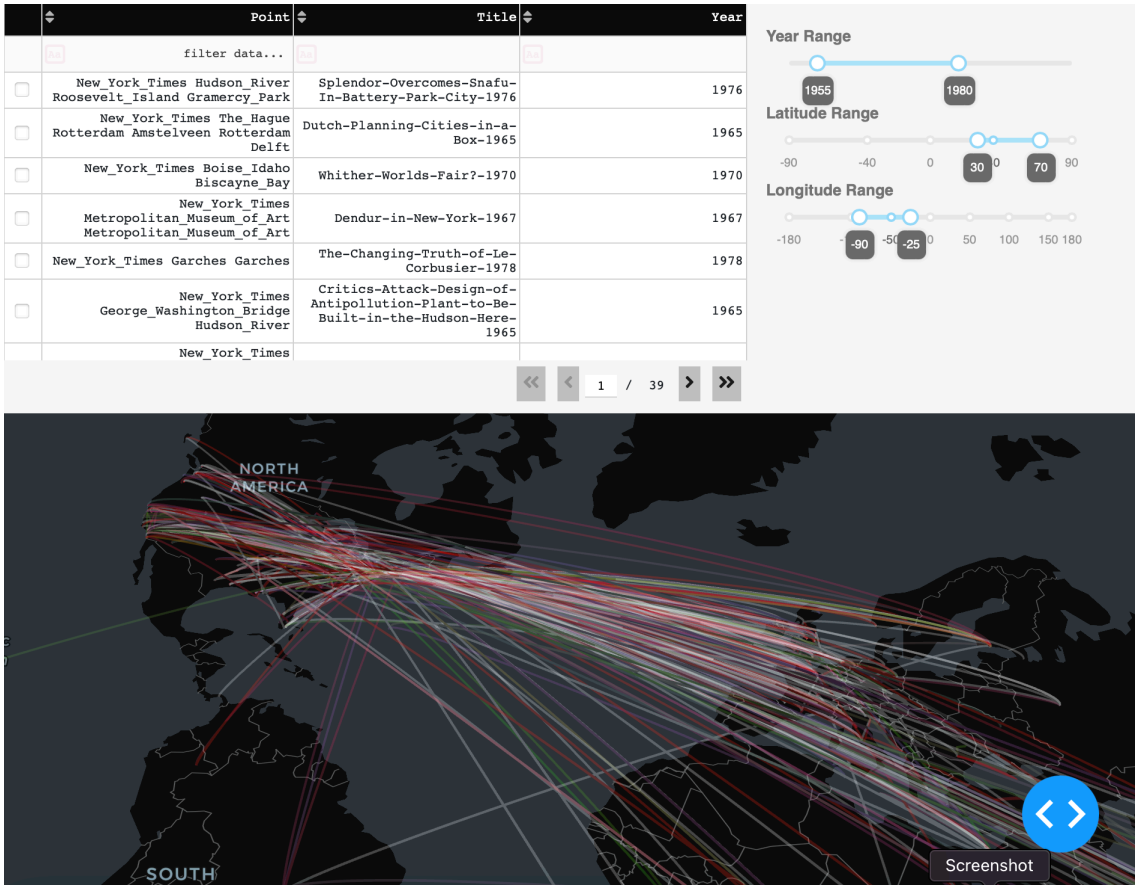
Of course, we can narrow in the search using any of the search facets given in the table. With respect to the geographical search window, the use of degrees of latitude and longitude is one of multiple options. In another version of this app, we narrow in by distance in miles from a location such as "Eiffel tower". Given that our geographic information retrieval method has street-level resolution, one can choose very small geographical search windows, such as "1 mile from the Eiffel tower" or "100 m from Bahnhofstrasse 10, Zurich, Switzerland".

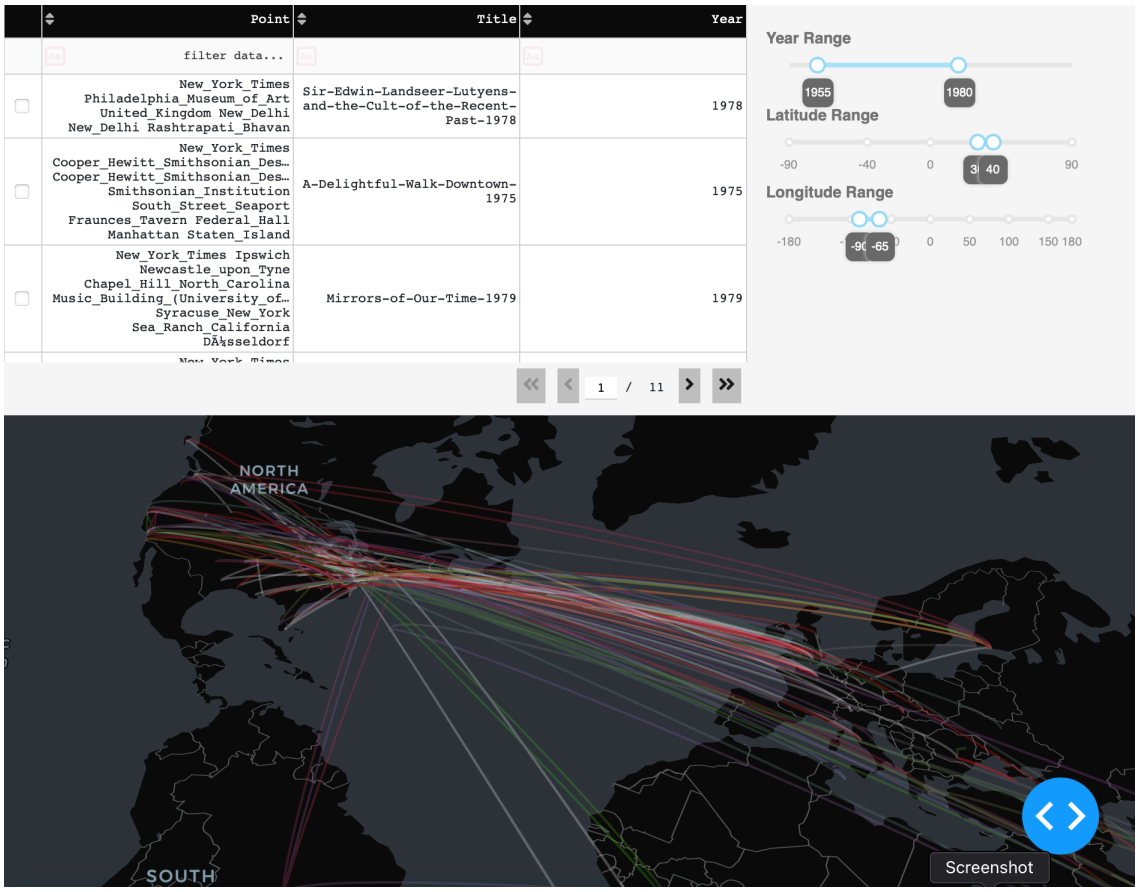
Figure 1 A & B, C & D. (next two pages)

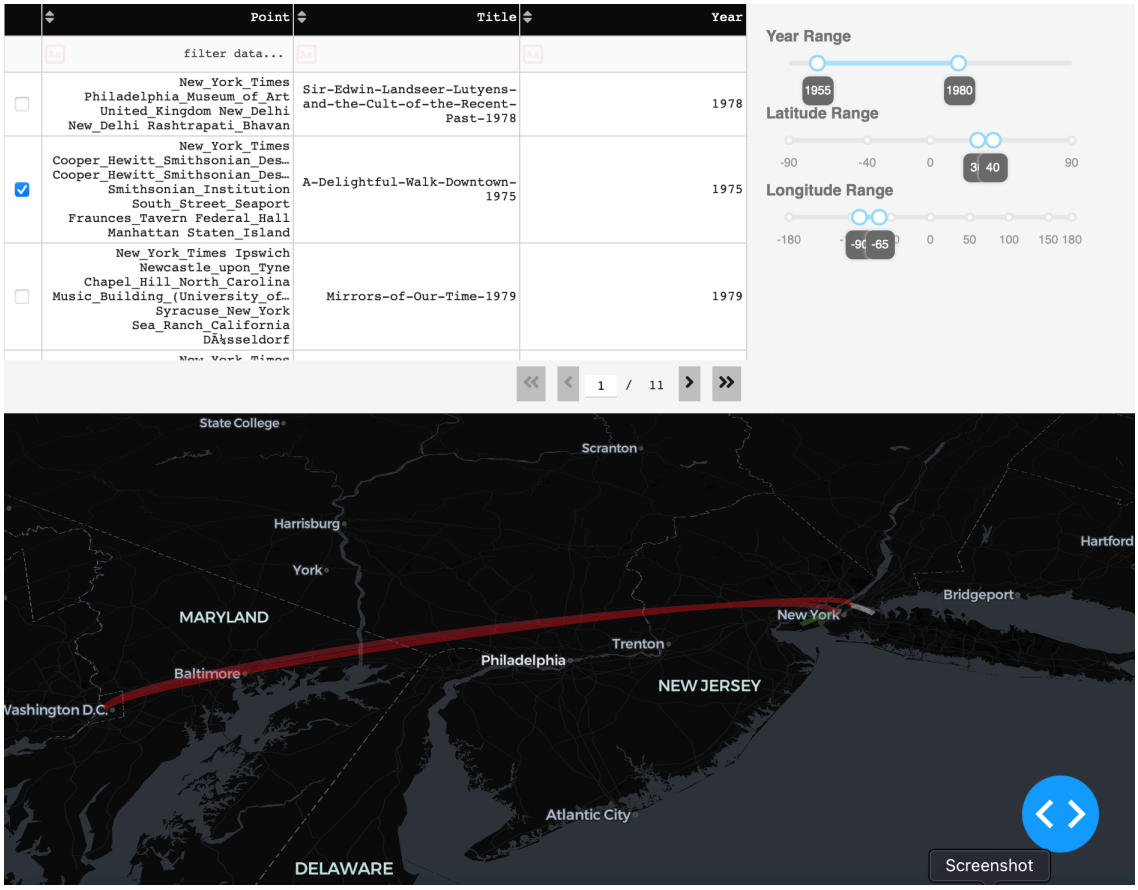
This figure is an example of our spatial discovery tool. We let you search in a collection of news article written by architectural critic Ada Louis Huxtable. Each news article is represented as a gradient line, starting white, and advancing to red, green, and grey, in reading direction.

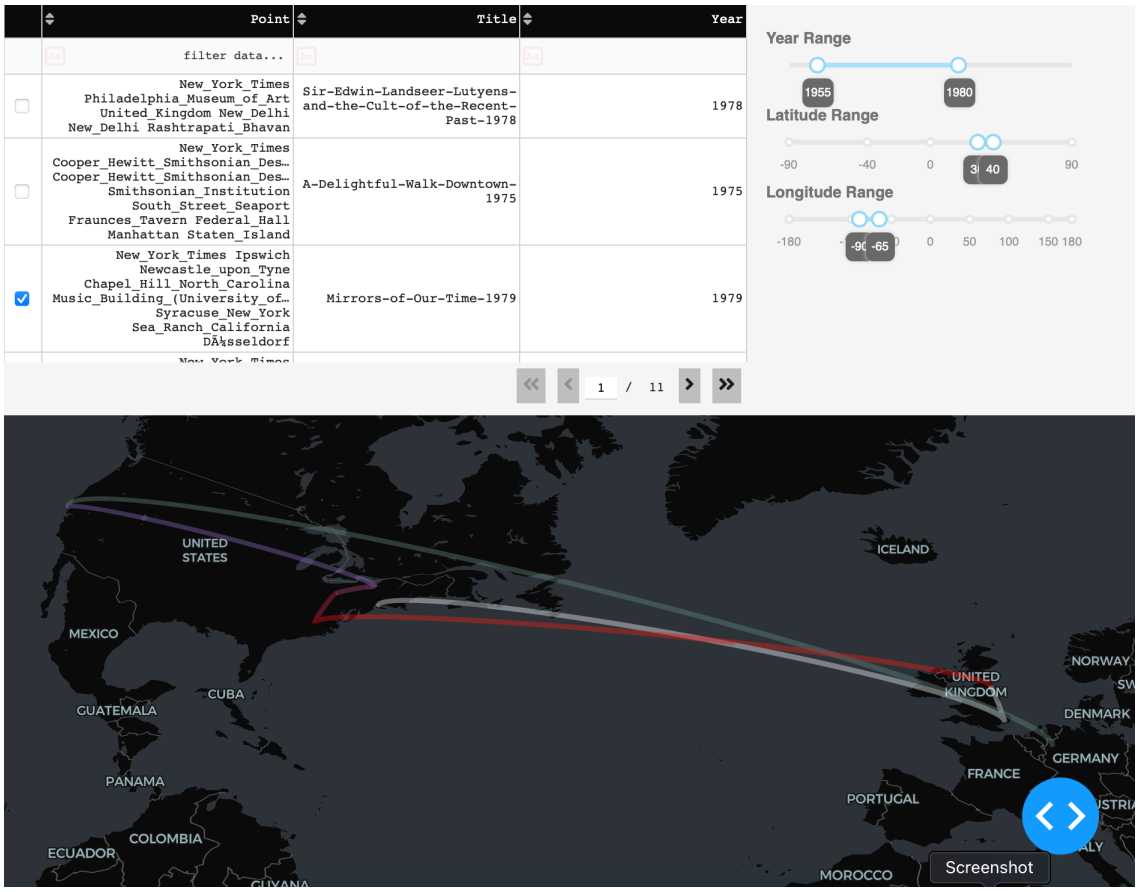
Figure 1 A and B (next page above and below) shows us narrowing down the geographical search window to filter for news articles that mention at least one entity in the given search window.

Figure 1 C and D (in two pages, above and below) shows us picking articles on the map, or picking them in the table of search results, which shows them highlighted on the map.


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 George_Washington_Bridge Hudson_River
Critics-Attack-Design-of-Antipollution-Plant-to-Be-Built-in-the-Hudson-Here-1965
1965
New_York_Times
1 / 39
Year Range
1955
1980
Latitude Range
-90
-40
0
30
70
90
Longitude Range
-180
-90
-25
50
100
150
180
NORTH AMERICA
SOUTH
Screenshot


Point
Title
Year
filter data...
New_York_Times Philadelphia_Museum_of_Art United_Kingdom New_Delhi New_Delhi Rashtrapati_Bhavan
Sir-Edwin-Landseer-Lutyens-and-the-Cult-of-the-Recent-Past-1978
1978
New_York_Times Cooper_Hewitt_Smithsonian_Des… Cooper_Hewitt_Smithsonian_Des… Smithsonian_Institution South_Street_Seaport Fraunces_Tavern Federal_Hall Manhattan Staten_Island
A-Delightful-Walk-Downtown-1975
1975
New_York_Times Ipswich Newcastle_upon_Tyne Chapel_Hill_North_Carolina Music_Building_(University_of… Syracuse_New_York Sea_Ranch_California DÃ¼sseldorf
Mirrors-of-Our-Time-1979
1979
1 / 11
Year Range
1955
1980
Latitude Range
-90
-40
0
40
90
Longitude Range
-180
-65
0
50
100
150
180
NORTH AMERICA
SOUTH
Screenshot


Point
Title
Year
filter data...
New_York_Times Philadelphia_Museum_of_Art United_Kingdom New_Delhi New_Delhi Rashtrapati_Bhavan
Sir-Edwin-Landseer-Lutyens-and-the-Cult-of-the-Recent-Past-1978
1978
New_York_Times Cooper_Hewitt_Smithsonian_Des… Cooper_Hewitt_Smithsonian_Des… Smithsonian_Institution South_Street_Seaport Fraunces_Tavern Federal_Hall Manhattan Staten_Island
A-Delightful-Walk-Downtown-1975
1975
New_York_Times Ipswich Newcastle_upon_Tyne Chapel_Hill_North_Carolina Music_Building_(University_of… Syracuse_New_York Sea_Ranch_California DÃ¼sseldorf
Mirrors-of-Our-Time-1979
1979
1 / 11
Year Range
1955
1980
Latitude Range
-90
-40
0
40
90
Longitude Range
-180
-65
0
50
100
150 180
State College
Scranton
Harrisburg
York
Hartford
MARYLAND
Bridgeport
New York
Trenton
Philadelphia
Baltimore
NEW JERSEY
Washington D.C.
Atlantic City
DELAWARE
Screenshot


Point
Title
Year
filter data...
New_York_Times Philadelphia_Museum_of_Art United_Kingdom New_Delhi New_Delhi Rashtrapati_Bhavan
Sir-Edwin-Landseer-Lutyens-and-the-Cult-of-the-Recent-Past-1978
1978
New_York_Times Cooper_Hewitt_Smithsonian_Des… Cooper_Hewitt_Smithsonian_Des… Smithsonian_Institution South_Street_Seaport Fraunces_Tavern Federal_Hall Manhattan Staten_Island
A-Delightful-Walk-Downtown-1975
1975
New_York_Times Ipswich Newcastle_upon_Tyne Chapel_Hill_North_Carolina Music_Building_(University_of… Syracuse_New_York Sea_Ranch_California DÃ¼sseldorf
Mirrors-of-Our-Time-1979
1979
1 / 11
Year Range
1955
1980
Latitude Range
-90
-40
0
40
90
Longitude Range
-180
-65
0
50
100
150 180
UNITED STATES
ICELAND
MEXICO
NORWAY
UNITED KINGDOM
CUBA
GUATEMALA
DENMARK
GERMANY
PANAMA
FRANCE
PORTUGAL
COLOMBIA
ECUADOR
MOROCCO
Screenshot

Related concepts

Our mapping approach can also be used to map the diversity of textual coverage for any given area. We have demonstrated this in a previous article (Baciu et al. 2022). Using this approach, a diversity map can be created that estimates how high the diversity of textual coverage is for any given point in geographical space. Diversity can be estimated based on topics as well as based on the nature of the geographical connections.

Another three concepts that are related to the present tool are what we call "isolexes", "isothemes", and "isops". They are scientific visualizations in the family of isonomes, isothermes, and isochrones.

Isolex: Given a collection of textual data, given a certain word, term, or name ("Chicago school"), given a textual distance from this word (for example 10 sentences), and given a particular geospatial resolution (for example 10 miles), an isolex is the equivalent of an ecological isonome, it is any line along which the word and its context (all text that falls within the distance of any mention of the word) are covered at roughly constant density. Isolex is chosen from the antique term for word as in "lexicon."

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 an LDA topic), and given a particular geospatial resolution, an isotheme is the equivalent of an ecological isonome, it is any line along which the topic is covered at roughly constant density.

Isop: Given a textual collection, given a location in geographical space (for example the Eiffel tower, Paris, FR), and given a particular geospatial resolution, an isop is the equivalent of an isochrone, it is any line in geographical space such that the distance from the location to any point on the line, if measured as distance in the textual data (for example as distance in words) is constant. Isop is chosen from the ancient Greek word for speaking, as in "epos".

Discussion

We have developed a geospatial discovery tool for libraries. Patrons can now find their books on the map. We believe that this advancement will change 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 developed can be applied in standard library settings as much as it is relevant online. News and social media platforms could use our tool or self-made versions of it to help readers direct their attention to geographical areas of their choice, too. Our tool (or derivate versions of our idea) can be used for geospatial discovery for text in any imaginable setting.

As many people desire, research and general public attention can now be actively directed to cover 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.

Implementation for libraries and online

We are happy to implement our tool for anyone who contacts us. The requirement is that the textual data are available in digitized format. Copyright is not a problem. We have developed safe practices to work with the copyrighted data. We may not need direct access to the data to process them (Organisciak and Downie 2022).

Bibliography

Baciu DC (2016). Sigfried Giedion: Historiography and history of reception on a global stage. Ar(t)chitecture, Haifa: Technion, 40-52.

Baciu DC (2017). Chicago school: Evolving systems of value. Report, HathiTrust Research Center.

Baciu DC (2018). From everything called Chicago school towards the theory of varieties. Doctoral dissertation, Chicago: Illinois Institute of Technology.

Baciu DC (2019). The Chicago school: Large-scale dissemination and reception. Prometheus 2, 20-42.

Baciu DC (2020). Cultural life: Theory and empirical testing. BioSystems197, 104208.

Baciu DC (2021). Culture and people flow together. OSF preprints.

Baciu DC (2022). Mapping Chicago schools. OSF preprints.

Baciu DC, Cellucci V (2022). Paths of wind and of research. OSF preprints.

Baciu DC, Mi D, Birchall C, et al (2022). Mapping Diversity: From ecology and human geography to urbanism and culture. SNSS 2:136. https://doi.org/10.1007/s43545-022-00399-4

Liu A, Droge A, Kleinman S, Thomas L, Baciu DC, Douglass J (2022). What everyone says: public perceptions of the humanities in the media. Daedalus 151, 19-39.

Organisciak P, Downie JS (2021) Research access to in-copyright texts in the humanities. Information and knowledge organization in digital humanities: global perspectives. London: Routledge, 157-177.

Roth D, Ji H, Chang M, Cassidy T. Wikification and Beyond: The Challenges of Entity and Concept Grounding. ACL (Tutorial Abstracts) 2014: 7

Sil A, Ji H, Roth D, Cucerzan S. Multi-lingual Entity Discovery and Linking. ACL (5) 2018: 22-29