OpenCSMap: A System for Geolocating Computer Science Publications

Felipe Manen and Aidan Hogan

DCC, University of Chile; IMFD; Santiago, Chile;
{fmanen,ahogan}@dcc.uchile.cl

Abstract. Although academic search engines are important tools for researchers, they typically support limited (if any) geographical features. In this demo we present a system that allows researchers to search for a specific Computer Science research topic and visualize in a map which affiliations have publications matching their search. The dataset is based on DBLP, using Entity Linking (OpenTapioca) over author affiliations to find geographic metadata for publications from Wikidata.
Demo link: `http://opencsmap.dcc.uchile.cl/`

1 Introduction

Academic search engines help researchers to find relevant literature for their topics of interest and to understand the impact of publications, venues, and authors. However, we are not aware of any available search engine that can return information about how many publications on a given topic are associated with a specific city or region. A system with such characteristics could help researchers to find new collaborators that work close by on similar topics; to select a place for organizing a conference or for continuing their career; to identify and include researchers from under-represented regions when organizing conferences; to understand networks of collaboration between cities and countries; etc.

Addressing this gap, we propose a system called OpenCSMap, which allows users to search over publications by keyword, and then aggregates all matching publications geographically (based on affiliation) in a map visualization. The current version of the system focuses on Computer Science publications, and is based on a snapshot of DBLP [3], extracting meta-data for all the conference and journal papers it describes. Geographic features are enabled by linking the authors' affiliations to Wikidata [6] using the OpenTapioca Entity Linking system [1]. SPARQL queries over the Wikidata Query Service are then used to obtain geographical information (coordinates, city, country) for those affiliations.

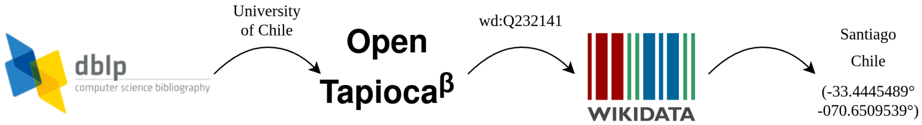


Fig. 1. Overview of how georeferenced data is acquired for University of Chile.

2 OpenCSMap

OpenCSMap can be divided into two main parts, relating first to the dataset preparation, and second to the search system itself. Semantic Web tools and resources (Wikidata, OpenTapioca) are used for the dataset preparation, while a NoSQL back-end (ElasticSearch) is used for the search system. Code for the system is available from `https://github.com/fmanen/OpenCSMap`.

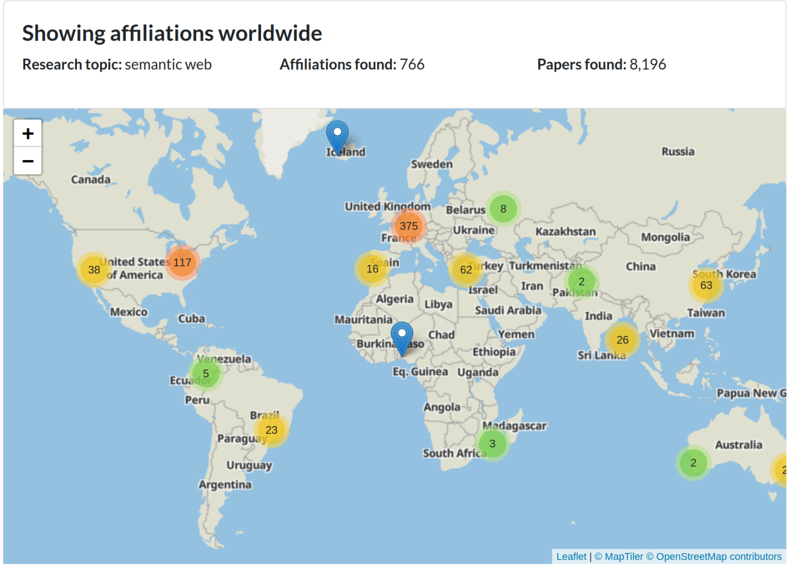
Dataset: We use DBLP as the core of our dataset. For OpenCSMap, we consider journal articles and conference (including workshop) papers, totaling 4,976,740 publications. The publication metadata we extract are the title, authors, year of publication, the name of the journal or proceedings in which it is published, and the DOI. Affiliations are not associated with publications, but rather authors. Hence it is unclear which affiliation an author had when publishing a particular work. For this reason, and to reduce noise, we design a conservative solution to select the most common affiliation across all authors, assigning one affiliation to nearly 50% of the articles. While more detailed affiliation information is available in commercial search engines such as Google Scholar, these datasets are not open.

We then focus on obtaining the georeferenced information of DBLP's textual affiliations. An overview of the process is shown in Figure 1. We performed initial experiments using a number of Entity Linking (EL) tools with online APIs – specifically DBpedia Spotlight [4], TAGME [2] and OpenTapioca [1] – for linking 100 randomly sampled affiliations to entities in DBpedia and Wikidata. Based on manual evaluation, OpenTapioca provided the best precision (0.84), followed by TAGME (0.81) and DBpedia Spotlight (0.48). We thus opted to proceed with OpenTapioca. For each Wikidata ID, we pose SPARQL queries using Wikidata's Query Service to filter the entities and extract the city and country of the affiliation (if available) and its respective coordinates. We provide the mapping of DBLP publications to Wikidata identifiers for their affiliations online at `https://zenodo.org/record/5038583`.

Search system: The search system uses Elasticsearch for storing and querying data; inverted indexes are built over the publication and geographical metadata. The web application is built using Django, Bootstrap, and Leaflet (for maps).

3 Demo

On the landing page, the user is presented with a keyword search dialog that can be used to search for a topic of interest (for example, "semantic web"). The



Showing affiliations worldwide

Research topic: semantic web **Affiliations found:** 766 **Papers found:** 8,196

Fig. 2. Map visualization of affiliations publishing works matching “semantic web”

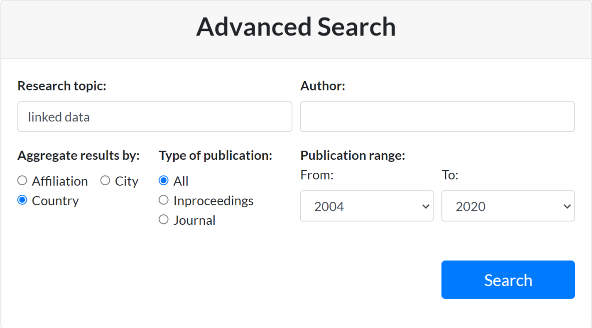
search is matched on paper and/or conference/journal titles. When the search button is clicked a map visualization is rendered as shown in Figure 2. The map shows all of the affiliations that have at least one publication in the results. If there are multiple affiliations nearby in the map, a cluster will be displayed with the number of affiliations in the cluster. As the user zooms in on a particular region, the clusters will become more fine-grained until markers are associated with only one affiliation. If such a marker is clicked, the number of matching publications and a list of publications are shown for that affiliation.

Aside from direct keyword searches, the system also supports an advanced search feature for restricting and aggregating results in a custom way. The interface for advanced search is shown in Figure 3. It enables users to search by topic, by author, by publication type, and by year of publication. It also allows for aggregating publications at the level of affiliation, city or country.

4 Experiments

We performed some initial user experiments to understand the relative strengths and weaknesses of the system. First we measured the response times for search; for these purposes we extracted 5,000 randomly sampled keywords for Computer Science papers from the Open Academic Graph[1]. We first performed a search for affiliations with publications matching the keywords against the back-end, where

[1] https://www.microsoft.com/en-us/research/project/open-academic-graph/



Advanced Search

Research topic: linked data

Author:

Aggregate results by: ○ Affiliation ○ City ● Country

Type of publication: ● All ○ Inproceedings ○ Journal

Publication range: From: 2004 To: 2020

Fig. 3. Advanced search interface

Table 1. Questionnaire results with **L1**–**L5** indicating disagreement–agreement

Statement	**L1**	**L2**	**L3**	**L4**	**L5**
I understand the purpose of the system	1	0	2	4	3
The system is intuitive to use	1	0	1	7	1
The system provides novel features vs. other online systems	1	0	3	4	2
I would use this system or a system like this	2	1	2	3	2
The data presented are complete and precise	2	3	2	3	0
The response times are reasonable	1	0	0	2	7

the average time for search per keyword was 18.3 ± 11.9 ms (standard deviation), with each query returning on average 472.3 ± 383.3 matching affiliations. We also tested the front-end times for requesting the map visualization over HTTP, where each search took on average 147.2 ± 93.1 ms. Though the front-end and HTTP connections add some overhead, response times remain well below a second.

We also wished to evaluate the initial impression of users of our system. We prepared a short questionnaire of six statements on a 5-point Likert scale. We sent the questionnaire to 20 full time professors of the Computer Science Department (DCC) of the University of Chile. The respondents were asked to try some searches of their choosing and to explore the system before responding. No detailed guidance on how to use the system, or its purpose, was provided. We received 10 responses. The statements and results are shown in Table 1. Overall the responses generally lean positive. The most positive aspect related to the low response times of the system, while more mixed or neutral responses were seen in terms of whether or not the respondents would use such a system, and regarding the completeness and precision of the data they observed.

5 Limitations and future work

OpenCSMap is a work in progress, and has some limitations that we aim to address in future work. The main limitation of the current system relates to the sparsity of affiliation data in DBLP, where we only find affiliations for 50% of the publications. For future work we intend to assign an affiliation per authorship of a publication, and to take into account temporal aspects relating to affiliations changing. Not all affiliations can be geolocated through Wikidata. Though this will become less of an issue as Wikidata expands and improves, generic affiliations such as IBM may be too vague to be geolocated accurately.

Another limitation is the recall of the searches. Our approach currently matches the search keyword(s) against some of the fields of a publication (title, proceedings or journal). Although this tends to provide relevant results, it may miss results; for example, a paper mentioning "RDF" but not "semantic web" would not be included in results for the latter. We are thus currently investigating adding topic classifications to the dataset (based on CSO [5]).

Though informative, our evaluation of the usability and usefulness of the system – whose results are summarized in Table 1 – remains preliminary, where our survey features some high-level questions and a relatively low number of responses (10) from users who are all professors. Performing more detailed evaluation with a more diverse set of users would help us to gain further insights on how to improve the system, and to tailor it for specific preferences and use-cases.

Other interesting ideas for future work include additional filtering options to select publications of interest; integrating metrics for papers, conferences, journals, etc., in order to support filtering by the impact of the publication or the prestige of the venue; inclusion of additional details for the papers indexed, such as abstracts; better navigation of the hierarchical information (e.g., from region to country to city to institution to publication, and back); autocompletion of specific topics of interest when searching by keyword; and a temporal slider to dynamically restrict the publications displayed on the map.

Acknowledgements We would like to thank Henry Rosales and Daniel Diomedi for their help with the Entity Linking process. We also thank the reviewers for their suggestions on potential future improvements for the system. This work was supported by ANID – Millennium Science Initiative Program – Code ICN17_002 and by FONDECYT Grant No. 1181896.

References

1. Delpeuch, A.: OpenTapioca: Lightweight Entity Linking for Wikidata. In: Wikidata Workshop. CEUR-WS.org (2020)
2. Ferragina, P., Scaiella, U.: TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities). In: CIKM. pp. 1625–1628. ACM (2010)
3. Ley, M.: DBLP - some lessons learned. PVLDB **2**(2), 1493–1500 (2009)
4. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: shedding light on the web of documents. In: I-SEMANTICS. pp. 1–8. ACM (2011)
5. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E.: The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas. Data Intell. **2**(3), 379–416 (2020)
6. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)