Open Knowledge Maps: Creating a Visual Interface to the World's Scientific Knowledge Based on Natural Language Processing

Peter Kraker[1], Christopher Kittel[2], and Asura Enkhbayar[3]

[1] *Know-Center Graz,*
pkraker@know-center.at
[2] *University of Graz,*
contact@christopherkittel.eu
[3] *University of Vienna,*
asura.enkhbayar@gmail.com

Abstract

The goal of Open Knowledge Maps is to create a visual interface to the world's scientific knowledge. The base for this visual interface consists of so-called knowledge maps, which enable the exploration of existing knowledge and the discovery of new knowledge. Our open source knowledge mapping software applies a mixture of summarization techniques and similarity measures on article metadata, which are iteratively chained together. After processing, the representation is saved in a database for use in a web visualization. In the future, we want to create a space for collective knowledge mapping that brings together individuals and communities involved in exploration and discovery. We want to enable people to guide each other in their discovery by collaboratively annotating and modifying the automatically created maps.

Das Ziel von Open Knowledge Map ist es, ein visuelles Interface zum wissenschaftlichen Wissen der Welt bereitzustellen. Die Basis für die dieses Interface sind sogenannte "knowledge maps", zu deutsch Wissenslandkarten. Wissenslandkarten ermöglichen die Exploration bestehenden Wissens und die Entdeckung neuen Wissens. Unsere Open Source Software wendet für die Erstellung der Wissenslandkarten eine Reihe von Text Mining Verfahren iterativ auf die Metadaten wissenschaftlicher Artikel an. Die daraus resultierende Repräsentation wird in einer Datenbank für die Anzeige in einer Web-Visualisierung abgespeichert. In Zukunft wollen wir einen Raum für das kollektive Erstellen von Wissenslandkarten schaffen, der die Personen und Communities, welche sich mit der Exploration und Entdeckung wissenschaftlichen Wissens beschäftigen, zusammenbringt. Wir wollen es den NutzerInnen ermöglichen, einander in der Literatursuche durch kollaboratives Annotieren und Modifizieren von automatisch erstellten Wissenslandkarten zu unterstützen.

1 Introduction

In the recent past, humanity has started to open up large amounts of scientific knowledge. Today, we can read over 60 million scientific articles on the web[1]. But the tools for exploring

[1] Bielefeld Academic Search Engine (BASE) indexes over 100 million articles from more than 4,600 sources. They estimate that about 60% of their articles are open access. Source: https://www.base-search.net/about/en/faq.php?#openaccess.

and discovering this massive amount of content are still lacking. Most people rely on list-based search engines, where they have to examine articles and their relationships by hand in order to acquire new knowledge. If one wants to gain an overview of a research field, it takes weeks if not months to process all the necessary information, which is usually scattered over thousands of scholarly articles.

Keeping track of a scientific field is therefore a challenging task, even for researchers who often have a community of peers to support them. People outside academia are usually on their own and therefore often lost. Take the example of patients who would like to learn about the newest research on their illness. In the worst case, they don't discover a life-saving treatment because the paper describing it was buried far down the results list.

Open Knowledge Maps is an attempt to solve these challenges by providing an open exploration and discovery tool that leverages the emerging digital open-science ecosystem.

2 Open Knowledge Maps

The goal of Open Knowledge Maps is to create a visual interface to the world's scientific knowledge. The base for this visual interface consists of so-called knowledge maps, a powerful tool for the exploration of a research field. Let's consider the example map of heart diseases depicted in Figure 1. Knowledge maps show the main areas of the field at a glance, represented in the example by circles, as well as papers related to each area. Areas that are more similar are positioned closer to each other than those dissimilar in subject.

By overlaying further connections between papers, e.g. references (see Figure 2), one can also highlight relationships between areas. In our example map, this shows us that a certain type of heart disease (A) is caused by a risk factor (B). We can also see that a certain drug (C) moderates this risk factor. This enables us to infer a connection between drug (C) and disease (A) which may have been previously unknown. This literature-based discovery was made popular by Swanson (1988). Knowledge maps thus enable the exploration of existing knowledge, and the discovery of new knowledge.

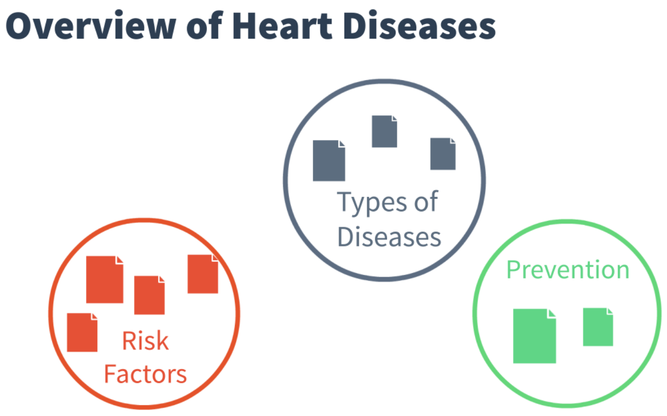


Figure 1: Example knowledge map of heart diseases. Circles represent research areas, paper icons represent important articles within an area.

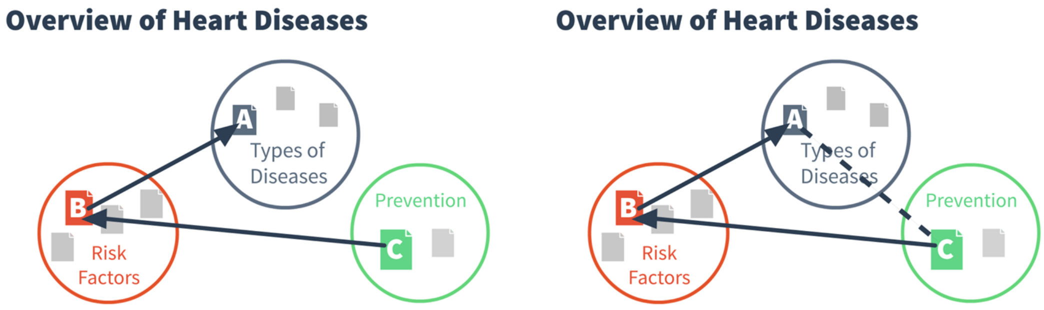
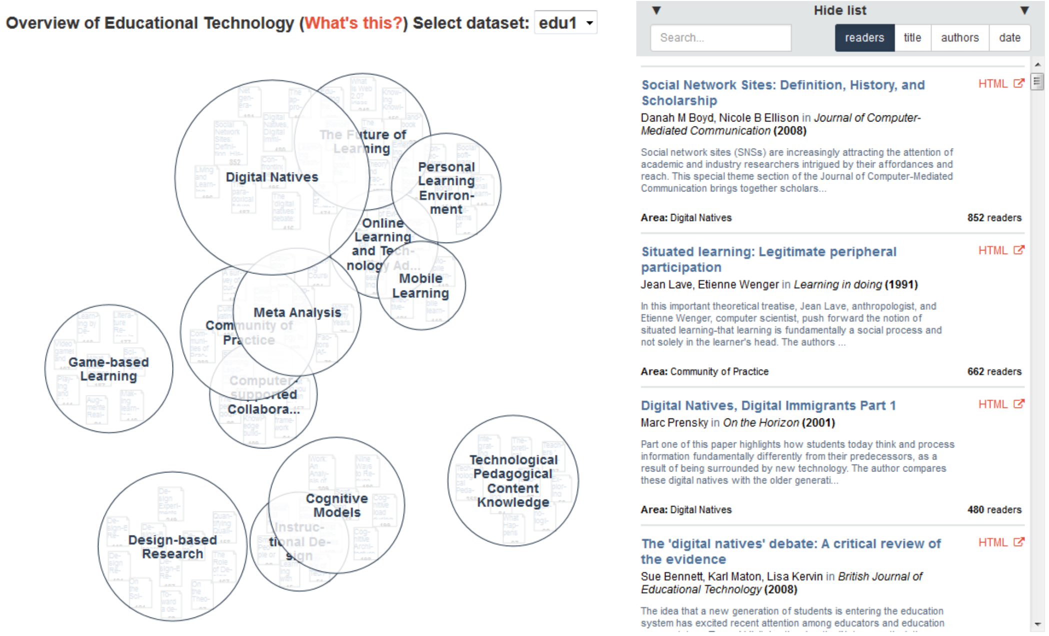


Figure 2: Literature-based discovery using knowledge maps. Continuous lines represent known connections, dashed lines represent unknown connections.

We aim to provide a large-scale system of open, interactive and interlinked knowledge maps for all fields of research that can be used by anyone. The existing service on http://openknowledgemaps.org automatically creates a knowledge map for any search term. It is based on the open-source, web-based knowledge mapping software Head Start[2], which is capable of producing knowledge maps from a variety of data, including text, metadata and references (see example in Figure 3 below). Head Start presents you with the main areas in the field, and lets you zoom into the most important publications within each area. Users can even read full papers within the same interface, provided that they are open access.

Currently, Open Knowledge Maps is based on the Public Library of Science (PLOS), but we are working on connecting it to PubMed and the Directory of Open Access Journals (DOAJ) to provide a wider coverage.


Overview of Educational Technology (What's this?) Select dataset: edu1
The Future of Learning
Digital Natives
Personal Learning Environment
Online Learning and Technology Ad...
Mobile Learning
Meta Analysis
Community of Practice
Game-based Learning
Computer-supported Collabora...
Technological Pedagogical Content Knowledge
Cognitive Models
Design-based Research
Instructional Design
Hide list
Search...
readers
title
authors
date
HTML

Social Network Sites: Definition, History, and Scholarship
Danah M Boyd, Nicole B Ellison in *Journal of Computer-Mediated Communication* **(2008)**
Social network sites (SNSs) are increasingly attracting the attention of academic and industry researchers intrigued by their affordances and reach. This special theme section of the Journal of Computer-Mediated Communication brings together scholars...
Area: Digital Natives **852** readers

Situated learning: Legitimate peripheral participation
Jean Lave, Etienne Wenger in *Learning in doing* **(1991)**
In this important theoretical treatise, Jean Lave, anthropologist, and Etienne Wenger, computer scientist, push forward the notion of situated learning-that learning is fundamentally a social process and not solely in the learner's head. The authors ...
Area: Community of Practice **662** readers

Digital Natives, Digital Immigrants Part 1
Marc Prensky in *On the Horizon* **(2001)**
Part one of this paper highlights how students today think and process information fundamentally differently from their predecessors, as a result of being surrounded by new technology. The author compares these digital natives with the older generati...
Area: Digital Natives **480** readers

The 'digital natives' debate: A critical review of the evidence
Sue Bennett, Karl Maton, Lisa Kervin in *British Journal of Educational Technology* **(2008)**
The idea that a new generation of students is entering the education system has excited recent attention among educators and education

Figure 3: Interface of Open Knowledge Maps.

[2] http://github.com/pkraker/Headstart

3 Natural Language Processing in Open Knowledge Maps

The goal of the service is to provide a high-level overview of research topics and their relatedness. For this reason we opted for a mixture of summarization techniques and similarity measures, which are iteratively chained together.

The back-end of the visualization is written in PHP and R. A processing component consisting of four stages is responsible for creating the data for the visualization, following the domain visualization process proposed by Börner, Chen, and Boyack (2003). In the first stage it connects to the user-selected API via an R script using packages provided by rOpenSci[3] to retrieve article metadata. Input queries of users are modeled according to the API specifications. For each query we retrieve the top 100 papers sorted by relevancy as well as metrics such as readership or citations.

In the second stage, we compile a bag-of-words corpus using article title, journal name, author names, subject keywords and the abstract. Several APIs also offer the possibility to retrieve full texts, but they are not available in all cases and also have a higher computing and bandwidth cost. Therefore, we have opted to use the metadata as a basis for creating the maps. Documents are preprocessed by removing punctuation, filtering stopwords, transforming to lower-case and stemming. Thereby, we reduce the dimensionality of the term-document matrix generated from this corpus. We then proceed to calculate the cosine similarity between papers using the R tm package[4].

Based on the similarity matrix, the spatial representation and the sub-areas are calculated using ordination and clustering techniques in the third stage, clusters are computed with Ward's method of minimum variance (Hair et al. 2010), which is known to join smaller clusters and to produce clusters of approximately the same size (Tan, Steinbach, and Kumar 2007). The number of clusters is determined by the *elbow* method, which chooses a number of clusters k that explains at least 80% of the model variance, and the increment is lower than 1% for $k+1$. We set a maximum of 15 clusters regardless of k.

In the fourth stage, a naming component determines the label for each cluster using keywords. For this, a Term Frequency - Inverse Document Frequency (TF-IDF) weighting is applied to the collection of subject keywords per cluster, and the top three keywords per cluster are selected. Some publication records are missing the "subject" field in the metadata, in which cases we heuristically fill the gaps by calculating the top-N words for these records.

After processing, the representation is saved in an SQLite database for use in the web visualization. The source code for the knowledge mapping-software called Head Start is available on Github[5] under the LGPL v3 open source license. With Head Start, visualizations can also be based on different similarity metrics, for example co-readership (Kraker et al. 2015).

[3] https://ropensci.org/

[4] https://cran.r-project.org/web/packages/tm/index.html

[5] http://github.com/pkraker/Headstart

4 Future Work

User feedback points to cases of papers being assigned a conceptually incorrect cluster (see Kraker et al. (2015)). We aim to improve clustering by tuning the features used to create the term-document matrix. Possible parameters include additional weights for, e.g., the title and keywords compared to the abstract.

The keywords entitling each cluster act as labels for the underlying publications and are – besides the spatial layout – an important means of orientation to the user. It is our goal to further increase the representativity of the top keywords. We will explore a range of options from the field of automatic summarization (Nenkova and McKeown 2011), including adding n-grams and different weighting techniques other than TF-IDF. Missing subject keywords in the metadata are not uncommon, and since they are a key element in determining the cluster naming, we aim to improve the heuristics applied in those cases. Possible paths include using top-N bigrams from the metadata in addition to unigrams and, more experimentally, inferring keywords from a mixture of similar documents, metadata, and if available, taxonomies.

Regarding improving the interface, we will extend the map visualization to enable highlighting of contextual facts and to create additional links between the papers. This adds another layer of filtering and visual clues to reduce information load. For example, a researcher might want to highlight all papers that contain the same species, focus on recently published material, or view the citation links between papers. This will enable literature-based discovery as outlined above.

Researchers might also want to cluster the resources based on a metric other than keyword similarity, like readership, type of content (i.e. paper, data set, presentation, etc.) or funding source. This in turn raises the question of comparability between maps created from heterogeneous data sources. A possible answer could be data fusion by matrix factorization (DFMF) as described in Žitnik and Zupan (2015). By simultaneously factorizing data matrices that represent different views of the same map (e.g. maps based on readership, citations, downloads) it becomes possible to create a common ground. This approach can be extended to process any kind of information that is represented as a data matrix, possibly enabling the fusion of data coming from the users, their interaction with maps and the literature itself.

We will also add integration with existing tools in the open digital ecosystem, including the Open Science Framework, Zotero, and ORCiD, so that Open Knowledge Maps will fit seamlessly into researchers' current workflows. Open Knowledge Maps strives to be completely open, so we will also add functionality to export the map and the underlying data in various open formats, so that a researcher could embed a map on her personal website or download the data for further processing.

5 Automated Approaches and Collaborative Editing

Purely automated approaches are still limited in the sense that the context of complex scientific questions cannot be sufficiently captured by algorithms alone. As a result, a map created by unsupervised techniques is never perfect: Papers get misplaced, area labels may be non-descriptive and important papers are missing. We therefore plan to allow users to add their knowledge and contextualization to maps by enabling editing and sharing of knowledge maps.

This will require adaptations to the back-end database operations and the front-end user interaction. On the front-end, we will enable an edit mode that allows researchers to manually add content to the map, modify or add metadata to content, like tags, and create new clusters. The editing history will be preserved in a Wikipedia-like model to allow collaborative building of knowledge maps. The maps themselves will be saved at Open Knowledge Maps where they can be browsed by other researchers and can serve as a starting point for other researchers' exploration.

Enabling collaborative editing is one stepping stone towards creating a space for collective knowledge mapping that brings together individuals and communities involved in exploration and discovery: researchers, students, journalists, librarians, practitioners and citizens. We want to enable people to guide each other in getting to the knowledge that they need, by collaboratively annotating and modifying the automatically created maps. Layered overviews created with the perspectives of different epistemic cultures or geographic regions help achieving the contextualization for better tackling our increasingly complex scientific and social challenges.

6 Acknowledgements

We would like to thank Maxi Schramm for providing the images. The Know-Center is funded within the Austrian COMET program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Börner, K., Chen, C., and Boyack, K. W. (2003). Visualizing knowledge domains. In: *Annual Review of Information Science and Technology* 37.1, pp. 179–255. DOI: 10.1002/aris.1440370106.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2010). *Multivariate Data Analysis*. 7th ed. Pearson Prentice Hall.

Kraker, P., Schlögl, C., Jack, K., and Lindstaedt, S. (2015). Visualization of Co-Readership Patterns from an Online Reference Management System. In: *Journal of Informetrics* 9.1, pp. 169–182. DOI: 10.1016/j.joi.2014.12.003.

Nenkova, A. and McKeown, K. (2011). Automatic Summarization. In: *Foundations and Trends in Information Retrieval* 5.2–3, pp. 103–233. DOI: 10.1561/1500000015.

Swanson, D. (1988). Migraine and Magnesium: Eleven Neglected Connections. In: *Perspectives in Biology and Medicine* 31.4, pp. 526–557. DOI: 10.1353/pbm.1988.0009.

Tan, P.-N., Steinbach, M., and Kumar, V. (2007). *Introduction to Data Mining*. 1st ed. Addison-Wesley.

Žitnik, M. and Zupan, B. (2015). Data Fusion by Matrix Factorization. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.1, pp. 41–53. DOI: 10.1109/TPAMI.2014.2343973.