Short description of algorithmic classification of PubMed at KI

2022-12-07 Peter Sjögårde

The implementation of the PubMed-based algorithmic classification in the KI bibliometric system is described in this document. The classification was created using clustering (or community detection) in a citation network of PubMed publications. The Leiden algorithm was used for clustering (Traag et al., 2019), basically using the procedure in Waltman and van Eck (2012).

Citation relations from the NIH Open Citation Collection was used to create the classification (iCite 2019). All articles and reviews from 1995 until the date of data extraction (December 2022 at the time of writing) are included, about 20 million publications with about 500 million relations.

The classification is hierarchical and has 4 levels. The most granular level corresponds to research topics. Topics are grouped into research specialties, specialties into disciplines and disciplines to broad research areas. Each publication belongs to exactly one class at each level. Granularities at the level of topics and specialties have been set to approximately the size of classes obtained in Sjögårde and Ahlgren (2018) and Sjögårde and Ahlgren (2020) respectively.

Labels have been created by extracting noun phrases from titles, medical subject headings (MeSH), journals and author addresses. The three most relevant terms have been selected by a term-weighting approach and concatenated into a label. The labeling approach is described in Sjögårde et al. (2021).

## Base map

A base map is available including all publications in PubMed from 2020-2022. The map can be used to create overlay maps. The visualization methodology is described in Sjögårde (2022).

Chart, bubble chart

Description automatically generated

<https://petersjogarde.github.io/pm_classification/2022oct/research_areas/index.html>

## Example of classification of a publication

“Comparison of balance and stabilizing trainings on balance indices in patients suffering from nonspecific chronic low back pain.”

Hosseinifar M, Akbari A, Mahdavi M, Rahmati M.

J Adv Pharm Technol Res. 2018 Apr-Jun;9(2):44-50. doi: 10.4103/japtr.JAPTR\_130\_18.

PMID: 30131936 Free PMC article.

Dicipline: pain; low; anesthesiology

Specialty: ergonomics; posture; vibration

Topic: abdominal muscle; transversus abdominis; core stabilityReferences

## Data availability

The classification is openly available for bulk downloads in Figshare:

<https://figshare.com/collections/PubMed_Classification/5610971>

## References

iCite, Hutchins, B. I., & Santangelo, G. (2019). *ICite Database Snapshots (NIH Open Citation Collection)*. The NIH Figshare Archive. Collection. https://doi.org/10.35092/yhjc.c.4586573

Sjögårde, P. (2022). Improving overlay maps of science: Combining overview and detail. *Quantitative Science Studies*, 1–40. https://doi.org/10.1162/qss\_a\_00216

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006

Sjögårde, P., & Ahlgren, P. (2020). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Science Studies*, *1*(1), 207–238. https://doi.org/10.1162/qss\_a\_00004

Sjögårde, P., Ahlgren, P., & Waltman, L. (2021). Algorithmic labeling in hierarchical classifications of publications: Evaluation of bibliographic fields and term weighting approaches. *Journal of the Association for Information Science and Technology*, *72*(7), 853–869. https://doi.org/10.1002/asi.24452

Traag, V. A., Waltman, L., & Eck, N. J. van. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748