Visualization of Materials Science Topics in Publications of Institutional Repository using Natural Language Processing

Sae Dieb‡, Keitaro Sodeyama‡, Mikiko Tanifuji‡

‡ National Institute for Materials Science (NIMS), Tsukuba, Japan

Corresponding author: Sae Dieb (dieb.sae@nims.go.jp)

Abstract

SAMURAI (NIMS 2022), a directory service of the National Institute for Materials Science (NIMS) researchers in Japan was launched in 2009 following the development of NIMS institutional repository (Tanifuji et al. 2019). The concept is to synchronize between profile information of researchers and their publications which are self-archived in the repository system. The SAMURAI was renewed in 2017 with interoperable functions with ORCID. SAMURAI supports various links to not only individual articles and patents, but also to databases such as KAKEN (Database of Grants-in-Aid for Scientific Research by NII). The service has yielded fully identified authors of journal articles from research members of NIMS by implementing a unique ResearcherID. Through this directory, NIMS is promoting materials research, supporting management of its researchers activities, and introducing NIMS researchers and their work to the public.

In this work, we present an application to describe each researcher's output topics automatically from the archived research papers in the repository, by implementing materials science specific natural language processing developed in our study (Dieb et al. 2021) that visualizes the research trend of each SAMURAI researchers. The approach can maximize information absorbance for general audience and fully corresponds to open science policy.

A list of publications' digital object identifiers (DOIs DOI 2022) for each researcher was constructed from his profile in SAMURAI. (In SAMURAI, the DOIs are stored in a PostgreSQL database). Using the DOI, recent publications were retrieved from NIMS text data mining platform (TDM-PF) in their XML format which were mainly available from 2003. Representative topic terms from their research publications that are related to materials science and engineering were extracted. We utilize term frequency analysis and automatic extraction for materials names to extract these necessary informative terms. Additionally, domain knowledge resources such as dictionaries were used. Data was preprocessed using noise reduction such as removing general English language stop words and physical

units filtering. Such words do not have significance on their own. Word cloud approach was used for visualization (Fig. 1).

This work brings us an opportunity to apply our NLP experience to mine information from research papers for public knowledge as a step towards data-driven materials science.

Keywords

FAIR data, open access repository, topic map

Presenting author

Sae Dieb

Presented at

First International Conference on FAIR Digital Objects, presentation

Hosting institution

Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Japan.

Conflicts of interest

The authors declare no conflict of interest.

References

- Dieb S, Amano K, Tanabe K, Sato D, Ishii M, Tanifuji M (2021) Creating research topic map for NIMS SAMURAI database using natural language processing approach. Science and Technology of Advanced Materials: Methods 1 (1): 2-11,. https://doi.org/10.1080/27660400.2021.1899426
- DOI (Ed.) (2022) DOI. https://www.doi.org/. Accessed on: 2022-8-22.
- NIMS (Ed.) (2022) SAMURAI. https://samurai.nims.go.jp/i. Accessed on: 2022-7-06.
- Tanifuji M, Matsuda A, Yoshikawa H (2019) Materials Data Platform- a FAIR System for Data-Driven Materials Science. 8th International Congress on Advanced Applied Informatics (IIAI-AAI). pp. 1021–1022 pp. https://doi.org/10.1109/IIAI-AAI.2019.00206

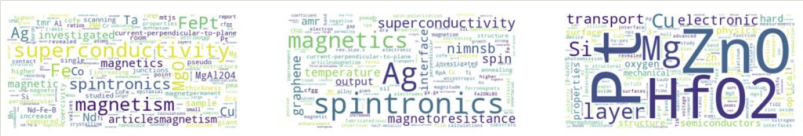

Figure 1.

Word cloud visualization of research topics for 3 researchers in SAMURAI repository. Representative terms were extracted using natural language processing approach.