Efficient workflow for automatic database creation from large scale scientific articles

Luca Foppiano[*], Pedro Baptista de Castro[+], Kensei Terashima[+], Yoshihiko Takano[+], Masashi Ishii[*]

[*]Material Database Group, MaDIS (NIMS)

[+]Nano Frontier Superconducting Materials Group, MANA (NIMS)

E-mail: FOPPIANO.Luca@nims.go.jp

The creation of automatically extracted databases of materials and properties from the scientific literature is the building block for data-driven materials science (Materials Informatics). Major related work in material science consists of extraction of materials and specific properties[1], synthesis recipes[2], capture of phase transitions for magnetic and superconducting materials[3], and collection of other superconductors properties from abstracts[4].

In the past, we have discussed our approaches to the extraction of linked named entities (NER) of materials and properties [5] [6] [7] which has been the core task of the project. However, creating databases automatically on large data requires high throughput, complete monitoring, data versioning, and punctual data reprocessing. Unfortunately, all these requirements cannot be addressed without designing a hybrid solution integrated with an existing workflow engine. We have developed a service combining Apache Airflow (workflow engine) and custom made tasks to use our superconductors extractor service for building a database of superconductors materials and properties from PDF documents of scientific articles. The process contains two main steps (Figure 1). (a) ”Extraction” of the document-based annotations supporting the storage of multiple versions (e.g. when reprocessing documents). (b) ”Aggregation” into tabular format, where each annotation transforms into various rows for each linked material-properties combination. We have processed about 250000 PDF documents from journals in materials science from various publishers, including APS, IOP and Elsevier and obtained a database of nearly 12000 entries of superconductors materials and linked properties. Furthermore, we plan to integrate manual curation into the aggregated information to correct mistakes in the automatic extraction.

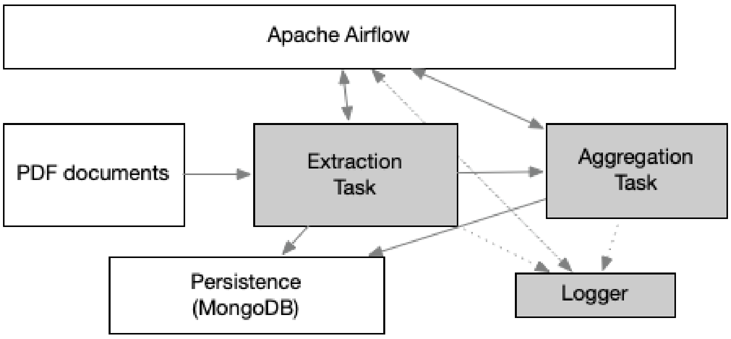

Figure 1: Workflow schema overview

[1] Court, C., Cole, J. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. Sci Data 5, 180111 (2018).

[2] Kononova, O., Huo, H., He, T. et al. Text-mined dataset of inorganic materials synthesis recipes. Sci Data 6, 203 (2019).

[3] Court, C., Cole, J. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. npj Comput Mater 6, 18 (2020).

[4] Yamaguchi, Kyosuke, Ryoji Asahi, and Yutaka Sasaki. ”SC-CoMIcs: A Superconductivity Corpus for Materials Informatics.” Proceedings of The 12th Language Resources and Evaluation Conference. 2020.

[5] Foppiano et al. SuperMat: Construction of a linked annotated dataset from superconductors-related publications. 2021.

[6] Foppiano et al. Supercuration: A machine-assisted data curation tool for rapid database construction for materials informatics, 2020

[7] Foppiano et al. Toward full automatic identification of superconducting materials and their properties in original papers: ambitious scope and current status, 2020