Supercuration: A machine-assisted data curation tool for rapid database construction for materials informatics

Luca Foppiano*, Masashi Ishii*
***Material Database Group, MaDIS (NIMS)**
E-mail: FOPPIANO.Luca@nims.go.jp

The availability of abundant publications in materials science and the needs for large structured datasets in Materials Informatics (MI) naturally require the establishment of techniques for automatic dataset construction from publications, namely Text and Data Mining (TDM) processes. However, although a lot of efforts have been paid for the establishment, full automatic TDM is still challenging. In this study, we propose a practical application of TDM, as a machine-assisted curation system.

The National Institute for Materials Science (NIMS) is constructing several databases for MI, and *SuperCon* is a hopeful data source in superconductor domain. To accelerate the data curation for *SuperCon*, we developed a machine-assisted curation system, Supercuration. Supercuration is based on an ML framework [1] using SuperMat[2] as training data.

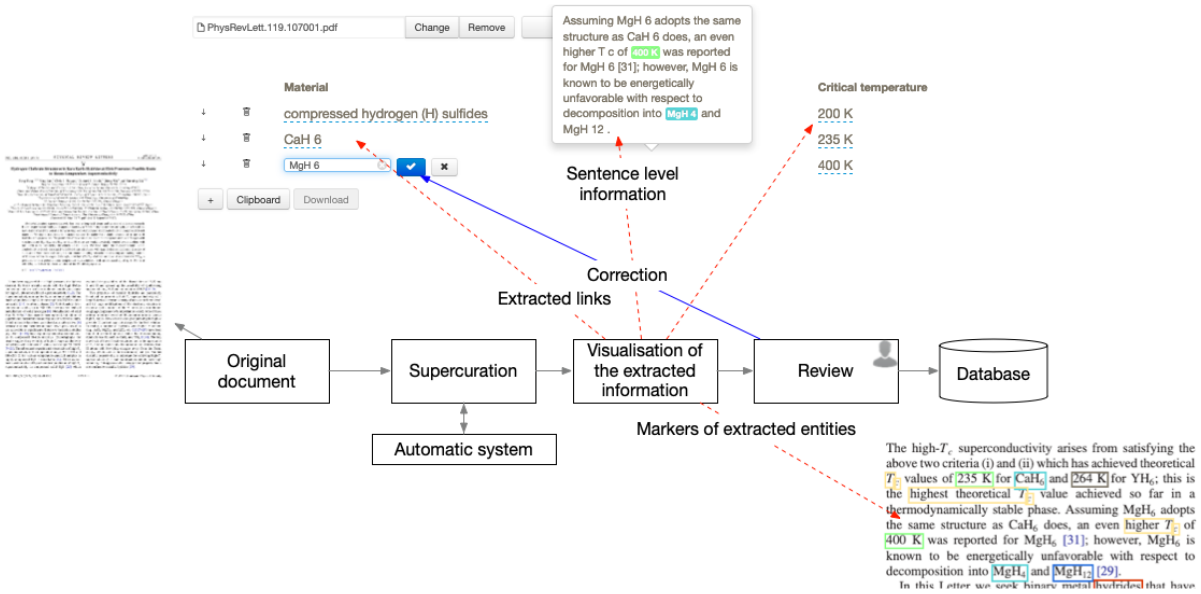

Figure 1: Supercuration workflow overview

Supercuration is an application that interfaces with the automatic system and allows users to process, visualise and correct the extracted data from a document. The obtained linked information (material - properties) are presented as a table that can be edited by the user, while all the entities are marked on a layer on top of the original document, indicating their type, value and other extracted information. The user can review the data and export it as standard formats, such as XML-TEI, XML-RDF, and JSON.

We plan to use this interface as a document viewer in other projects to exploit the functionality where each extracted information can be linked back to the original document.

1. Foppiano et al., Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature. IEICE 2019
2. Foppiano et al., SuperMat: Corpus for Extraction of Superconductor Materials Data, JSAP Spring, 2020