Automatic construction of an accessible linked dataset from scientific literature for superconducting materials discovery

Luca Foppiano, Sae Dieb, Akira Suzuki, Pedro Baptista de Castro, Suguru Iwasaki, Yan Meng, Kensei Terashima, Yoshihiko Takano, Masashi Ishii

Automatic construction of an accessible linked dataset from scientific literature for superconducting materials discovery

Luca Foppiano[1], Sae Dieb[1], Akira Suzuki[1], Pedro Baptista de Castro[2], Suguru Iwasaki[2], Yan Meng[2], Kensei Terashima[2], Yoshihiko Takano[2], Masashi Ishii[1]

[1] ***Materials Database Group, MaDIS, NIMS, Japan***

[2] ***Nano Frontier Superconducting Materials Group, MANA, NIMS, Japan***

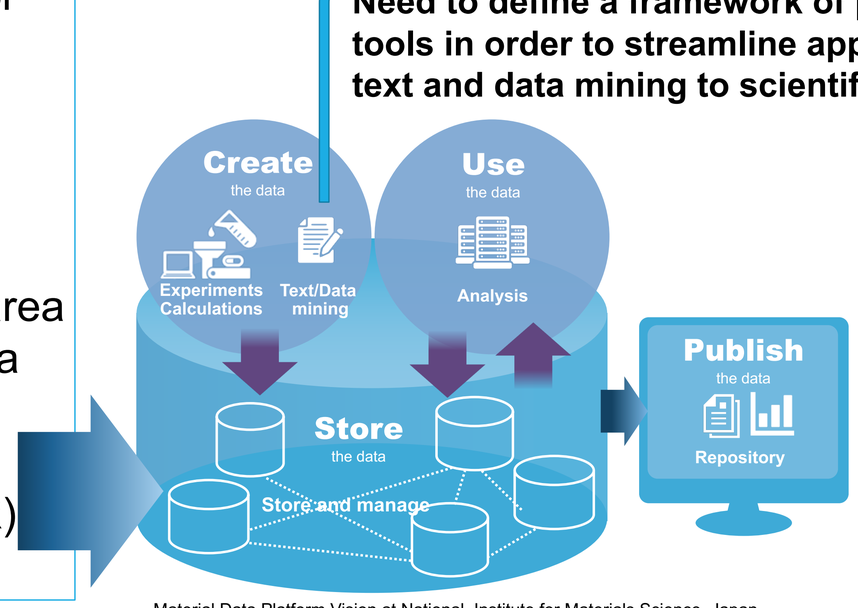
Introduction and Motivation

Structured data is key ingredient in machine learning for materials research.

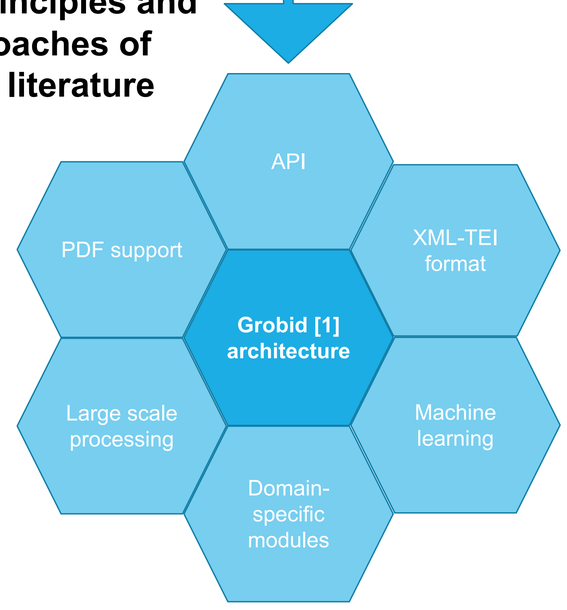
Scientific publications provide large scale unstructured data, however:
- Text-mining is challenging
- Solutions can be reused within the same scientific area
- Machine learning requires availability of training data

Needs of FAIR infrastructure for data and tools:
- Findable, Accessible, Interoperable Reusable (FAIR) data principles

Need to define a framework of principles and tools in order to streamline approaches of text and data mining to scientific literature



Material Data Platform Vision at National Institute for Materials Science, Japan
Published in NIMS NOW, 19 (1), 2019

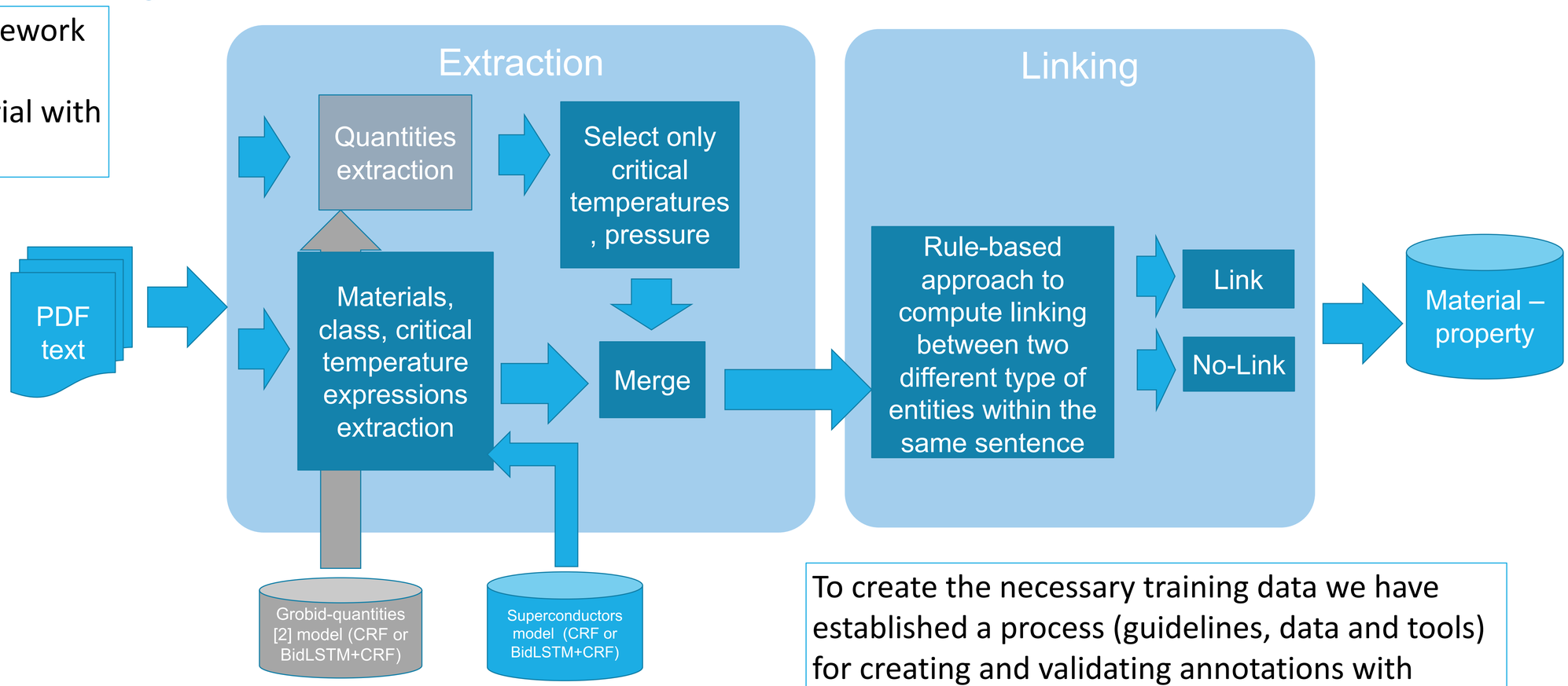


[1] Grobid: http://github.com/kermitt2/grobid

Superconductors case study

Implementation of our framework within the superconductors domain for extracting material with their respective properties.

Our tool combines ML-based entity ”Extraction” through sequence labelling and heuristic ”Linking” of the identified materials with their respective physical properties.



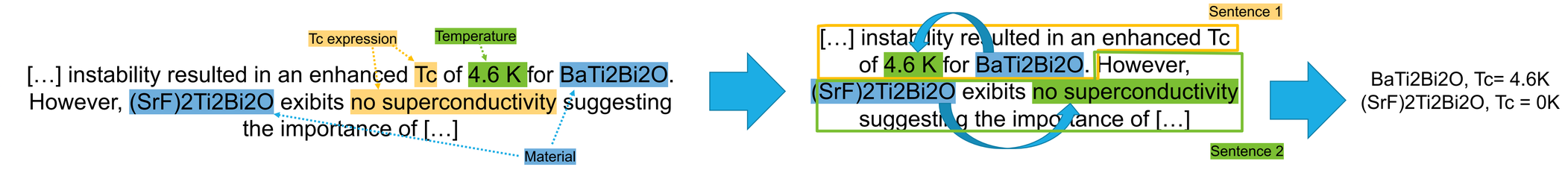
To create the necessary training data we have established a process (guidelines, data and tools) for creating and validating annotations with domain experts.

[2] Grobid-quantities is a module of Grobid focusing on Measurement extraction: http://github.com/kermitt2/grobid-quantities

[3] excerpt from article “Synthesis and Physical Properties of the New Oxybismuthides BaTi2Bi2O and (SrF)2Ti2Bi2O with a d1 Square Net, Takeshi et all, 2012, JPS”

Example

[…] instability resulted in an enhanced Tc of 4.6 K for BaTi2Bi2O. However, (SrF)2Ti2Bi2O exibits no superconductivity suggesting the importance of […]



Evaluation

Extraction

Corpus of 114 PDF papers evaluated using 10-Fold cross validation

Label	Precision	Recall	F1-Score
Class	81.66	72.36	76.64
Material	81.89	80.09	80.96
Measurement method	73.57	71.24	72.26
Critical pressure	55.27	34.4	41.54
Critical temperature expression	82.3	77.61	79.83
Critical temperature value	74.64	61.8	67.56
All (micro avg)	**80.41**	**75.89**	**78.07**

End to end Evaluation (Extraction + Linking)

Corpus of 500 PDF papers from American Institute of Physics (AIP), American Physical Society (APS) and Institute of Physics (IOP). Results calculated using manual evaluation

Total	Correct	Wrong	Precision	Recall	F1-score
597	441	156	**73.86**	**66.33**	**69.90**

Future work

- Linking improvement with probabilistic approach
- Increase the training data and extract more information (e.g. magnetic field)

Contact information: FOPPIANO.Luca@nims.go.jp