SuperMat: Construction of a linked annotated dataset from superconductors-related publications

Luca Foppiano, Sae Dieb, Akira Suzuki, Pedro Baptista de Castro, Suguru Iwasaki, Azusa Uzuki, Miren Garbine Esparza Echevarria, Yan Meng, Kensei Terashima, Laurent Romary, et al.

SuperMat: Construction of a linked annotated dataset from superconductors-related publications

Luca Foppiano[1*], Sae Dieb[1], Akira Suzuki[1], Pedro Baptista de Castro[2], Suguru Iwasaki[2], Azusa Uzuki[2], Miren Garbine Esparza Echevarria[2], Yan Meng[2], Kensei Terashima[2], Laurent Romary[3], Yoshihiko Takano[2], and Masashi Ishii[1*]

[1]Material Database Group, MaDIS, NIMS, Tsukuba, 305-0044, Japan
[2]Nano Frontier Superconducting Materials Group, MANA, NIMS, Tsukuba, 305-0047, Japan
[3]ALMAnaCH, Inria, Paris, 75012, France
[*]corresponding authors: Luca Foppiano (FOPPIANO.Luca@nims.go.jp), Masashi Ishii (ISHII.Masashi@nims.go.jp)

January 28, 2021

Abstract

A growing number of papers are published in the area of superconducting materials science. However, novel text and data mining (TDM) processes are still needed to efficiently access and exploit this accumulated knowledge, paving the way towards data-driven materials design. Herein, we present SuperMat (Superconductor Materials), an annotated corpus of linked data derived from scientific publications on superconductors, which comprises 142 articles, 16052 entities, and 1398 links that are characterised into six categories: the names, classes, and properties of materials; links to their respective superconducting critical temperature (T_c); and parametric conditions such as applied pressure or measurement methods. The construction of SuperMat resulted from a fruitful collaboration between computer scientists and material scientists, and its high quality is ensured through validation by domain experts. The quality of the annotation guidelines was ensured by satisfactory Inter Annotator Agreement (IAA) between the annotators and the domain experts. SuperMat includes the dataset, annotation guidelines, and annotation support tools that use automatic suggestions to help minimise human errors.

Introduction

The vast majority of scientific knowledge exists as published articles [9, 16, 35, 1]. These publications are presented mainly as text, which is challenging to be

used as a machine-readable structure. Meanwhile, as a part of the text and data mining (TDM) discipline, computer-assisted information collection from the literature has become a supportive asset for scientific research [34]. In the past decades, new TDM processes were developed for several natural science disciplines to achieve automatic document processing such as information retrieval, entity extraction, and clustering. TDM has been applied in biology for identifing interactions between agents (e.g. bacteria, viruses, genes, and proteins) [12, 24, 23] to support the research on serious diseases including cancer [26]. In chemistry, it was used for the disambiguation of chemical compounds names, synthesis extraction, and retrieval [11]. In both domains, the application of TDM was based on manually curated datasets (corpora) that functioned as infrastructures. Examples are the BioCreative IV CHEMDNER corpus [25] in chemistry, and Genia [17] and GENETAG [38, 33] in biology. Such datasets are crucial for developing, training, and evaluating TDM systems.

In comparison, such resources in the materials science domain are rather limited. Reported cases include NaDev [4] on nanocrystal devices research, SC-CoMIcs [39] in the superconductors domain, and a corpus for extracting synthesis recipes [21]. To address this shortage of infrastructure, experimental data is extracted manually [8], or ab-initio calculations are used [14] but they might not accurately describe the real system. Several challenges still hinder the data-driven exploration of materials (also called Materials Informatics (MI)), namely: the lack of data standard, infant stage of the data-driven culture, a wide variety of conflicting stakeholders, and missing incentives for researchers to contribute to large collaborative initiatives [13]. To bridge these gaps, it is necessary to create infrastructural resources to support TDM processes in materials science through the automatic construction of databases for materials and their properties. Such application can minimise the need for humans to read the new papers and extract the key information therein. Equally importantly, it enables scientists to focus and leverage computing power and human resources to find deeper relationships between superficially unrelated information. Other applications include providing semantically enriched search engines that accept fine-grain queries [29] to reduce the time needed to access specific information. These processes cannot be established without essential resources such as dictionaries, lexicons, and datasets.

Research on superconducting materials has been growing rapidly towards both fundamental science as well as practical applications. Superconductors display many intriguing phenomena including zero-resistivity, the ability to host a high magnetic field, quantisation of the magnetic flux, and vortex pinning. Current applications of superconductors include medical instruments, high-speed trains, quantum computers, and the Linear Hadron Collider (LHC) [30, 18, 2]. However, discovering a new superconductor is a challenging task, as only 3% of candidate materials were found to be superconductors [20]. The National Institute for Materials Science (NIMS) in Japan has been manually constructing databases to support material research, and SuperCon (`http://supercon.nims.go.jp`) is a manually curated data source for the superconductor domain. These databases would help researchers design new superconducting materials with a higher superconducting critical temperature (T_c) (ideally up to room temperature) [10, 37]. However, the current resources are very limited and not dynamic enough to incorporate the information from new publications in a timely manner. In this paper, we present SuperMat (Superconductors Ma-

terials), an annotated linked corpus for superconducting material information. This dataset contains 142 documents with 16052 (7166 unique) entities, and 1398 links that can serve as an infrastructural data for TDM processes in the domain of superconducting materials. SuperMat is different from SC-CoMIcs based on the following reasons: (a) it provides full papers instead of abstracts which contain more detailed information about the research on superconducting materials, and (b) it contains linked entities. We also describe the construction guidelines for SuperMat, in the hope of supporting researchers to systematically create annotated data. Furthermore, the unique feature of links between entities in SuperMat will allow the development of more precise methodologies to associate a particular material with its properties.

Methods

Content acquisition

SuperMat originates from PDF documents of scientific articles related to superconductor research. The PDF format is the most widely used format for scientific publications [15]. The original documents were collected from the following sources: (a) the Open Access (OA) version of peer-reviewed articles referenced in the SuperCon database records; (b) articles provided by domain experts containing suitable items and potential links of material names, T_c values, measurement methods, and pressures; (c) articles from "condensed matter" category of arXiv (`https://arxiv.org/archive/cond-mat`) selected using the search terms of "superconductor", "critical temperature", and "superconductivity".

Pre-print versions of peer-reviewed articles were obtained using a lookup service for bibliographic data called biblio-glutton (`https://github.com/kermitt2/biblio-glutton`) that aggregates data from various sources: the Crossref (`https://www.crossref.org/`) bibliographic database, the unPaywall (`http://unpaywall.org`) service, the PubMed Central repository (`https://pubmed.ncbi.nlm.nih.gov/`), and mappings to other databases. We queried *biblio-glutton* using the bibliographic data of each article referenced in Supercon; subsequently, we downloaded the pre-print article associated with the retrieved record, if available. Although the published version may be different from the pre-print version of a document, the differences measured by comparing pre-print and peer-reviewed articles in biology [3] measured objective differences to be around 5%.

Preliminary annotation study

Preliminary annotation study was carried out to assess the effort required from the annotators to reach an acceptable Inter Annotation Agreement (IAA ¿ 0.7) . We annotated two randomly selected OA papers, by using a preliminary version of the guidelines with a limited tag-set of four labels: `<material>`, `<tc>` (expression describing the presence or absence of superconductivity), `<tcValue>` (value of T_c), and `<doping>` (amount of substitution, such as stochiometric values, usually expressed as functions of x or y). The process was iterated multiple times. Each iteration ended with computing the IAA using the Krippendorff's alpha coefficient [27, 41], while annotators discussed the disagreements, and

updated the guidelines.

Based on the results in Table 1, IAA reached a satisfactory level (0.9) after the third iteration. In the second iteration, although the average IAA reached 0.7 on three of the four labels, the average agreement was not satisfactory. When analysing the disagreement, we noticed that the low score in the `<doping>` label was caused by a heavy overlap with the `<material>` label, which required more precise definition in the guidelines.

Based on this preliminary study, the following changes were implemented. (a) The label `<doping>` was merged under the `<material>` because, even with detailed documentation it was too difficult for humans to annotate them in a consistent way. (b) Three more labels were added: measurement methods and pressure (described as parametric conditions in relation to T_c), and class of materials.

Tag-set design

The tag set (also referred to as *labels*) represents the classes of entities and the type of links between them, which were designed to be extracted from the text (Figure 1).

Entities

Entities (also referred as Named Entities, mentions, or surface forms) are chunks of texts that represent an information of interest, as follow:

- Class (tag: `<class>`) represents a group of materials defined by certain characteristics. Superconducting materials can be classified according to different criteria such as the composition and magnetic properties. Among publications collected for this study, the domain experts identified three types of classes based on: (a) the composition and crystal structure, (b) material phenomena (e.g. "I-type" and "II-type superconductivity", "BCS superconductors", "nematic", and "conventional/unconventional superconductivity"), and (c) high/low T_c value (e.g. "high-tc" superconductors).

  In this work, we only considered the (a) classes, mainly because the material composition and crystal structure do not change with time. For example, a cuprate from 1998 is still called a cuprate today. In comparison, many material phenomena used for (b) are not robust enough, and can be biased by the viewpoint of the author(s) or research group, or the measurement methods. Finally, the definition of "high-tc" superconductors (c) is completely relative; i.e., with the progress of research, materials once considered "high-tc" might not be so anymore.

- Material (tag: `<material>`) identifies the name of one or more materials. This label is used to collect the following types of information:

  - Chemical formula indicating the material by its general or stochiometric formula (e.g. $LaFe_{1-x}O_7$, WB_2),
  - Compositional name (e.g. `magnesium diboride`) or abbreviations (e.g. `YBCO`),

  - The material's shape (e.g. wire, powder, thin film) or form of material (e.g. single/poly crystal),
  - Modification by a dopant (`Zn-doped`, `Si-doped`) or by percentage of doping (`2%-doped`). We also considered qualitative expressions such as *overdoped*, *lightly doped*, and *pure* as valid information,
  - Substrate information (e.g. `grown on MgO(100) film`) when it was adjacent to the material name or formula, in the text,
  - Additional information about the sample (e.g. `as-grown`, `untwinned`, `single-layer`) when it was adjacent to the material name or formula, in the text.

- Superconducting critical temperature (tag: `<tc>`) identifies expressions related to the phenomenon of superconductivity. Any temperature mentioned in the text is not necessarily the T_c. Rather, it could refer to the temperature for other processes/events such as annealing/sintering temperature, specific measurements, and structural changes. This label identifies the presence or absence of superconductivity at a given temperature (*showing/not showing superconductivity at this* T_c). In addition, modifiers of this information (increasing/descreasing T_c) are also retained.

- Superconducting critical temperature value (tag: `<tcValue>`) represents the temperature at which the superconducting phenomenon occurs. It can be defined by different experimental criteria, such as the onset, mid-point of resistivity drop, or zero resistivity. This value also considers boundary conditions, such as the *onset of superconductivity*, *zero resistance*.

- Applied pressure (tag: `<pressure>`) indicates the applied pressure corresponding to a measured T_c.

- The measurement method (tag: `<me_method>`) indicates the method used to measure or calculate the presence of superconductivity. Here, we considered the following categories: resistivity, magnetic susceptibility, specific heat, and theoretical calculations.

Links

The links connects entities of materials or samples to their corresponding properties, conditions, and results. The links are non-directional, and there are no restrictions on the number of links for each entity. We defined three types of links:

- material-tc: linking materials to their T_c values.

- tc-pressure: connecting T_c and the applied pressure under which it was obtained.

- tc-me_method: linking T_c and the corresponding measurement method.

Annotation guidelines

Annotation guidelines include the principles and the rules that describe what constitutes as desired information for the SuperMat dataset and how to annotate it. They include detailed description of the specific rules that have been defined for each type of information to be annotated, with one or more definitions and examples illustrating what to annotate in different cases, exceptions, and references. We used an online system to track the discussions and decisions when a question or a comment was raised, and provided a link to such issues in the respective description or example. In addition, the guidelines include *linking rules* that provide information on how to correctly connect the entities in a relationship. The guidelines were built using a dynamic markup language (called RestructuredText) and stored in a git (`https://git-scm.com/`) version control system repository. We deployed them as HTML files via web, which were updated automatically after each modification.

Annotation support tools

The task of annotating documents is tedious and requires both attention and subject knowledge from the annotators. Annotation support tools aim to maximise the efficiency of annotators and minimise human mistakes. They are composed of a web-based collaborative annotation tool, automatic annotation suggestions, and automatic corpus analysis.

Web-based collaborative annotation tool: INCEpTION

The annotation tool is the platform used for creating, correcting and linking annotations. After evaluating several tools, we selected INCEpTION [19, 5], a web-based multi-user platform for machine-assisted rapid dataset annotation construction. INCEpTION provides supportive functionalities that include:

- Multi-layer annotation sheets allow different annotation schemas over the same documents,
- Two annotation steps: annotation consists of manually correcting pre-imported documents, while curation allows another user to validate the annotations (Figure 5).
- On-the-fly automatic suggestions based on active learning and string matching (Figure 5),
- Bulk annotation corrections, and
- Being open-source (Apache 2.0 license), and under active development at the time of this paper (`https://inception-project.github.io/`).

Annotation suggestions

Previous works have demonstrated that annotation suggestions improve the quality of the output [7, 32, 28]. We provide two types of annotations suggestions. (i) *Machine-based annotated data* that were assigned to the documents before loading into the annotation tool. Here, we use a machine learning (ML)-based system from a previously implemented prototype [6] to support our

tag-set. (ii). *Active learning recommendations* provided by INCEpTION are assigned on-the-fly based on previous annotations. The active-learning recommendations are less precise since they aim to increase the recall, and therefore they need to be explicitly accepted by the annotator.

Automatic corpus analysis

Automatic corpus analysis is a set of scripts designed to run after the validation step. These scripts automatically find inconsistencies in the links and entities, while extracting the statistics of the corpus. We calculated the inconsistencies by examining every annotated entity and computing the frequency of the same text being annotated with different labels. The script outputs a summary table by visualising each annotation value, as well as their labels and frequencies. We visually inspected this table, because the reported inconsistencies can be either obvious mistakes (Table 2) or arise from ambiguities (Table 3); therefore their context should be verified.

Although the links are conceptually non-directed, we have defined a practical convention to maintain their consistency. For example, *material-tc* is always represented as a link between `<tcValue>` and `<material>` entities. The script also computes the statistics (Table 4) for the number of entities (total, unique, by class), the number of links (total, intra- and inter-paragraph, between paragraphs), and other statistical information.

Annotation process

The annotation workflow (Figure 2) was designed following the *MATTER* (Model, Annotate, Train, Test, Evaluate, and Revise) schema[36] and other related work [4, 25]. The workflow is composed of five steps (Figure 2): *data-preparation*, *correction*, *validation*, *testing and evaluation*, *revision*. This workflow involves three main actors: the automatic process, computer scientists, and the domain experts.

The first step of the annotation process involves preparing the machine-based annotated data from the source PDF documents. The PDF files are converted to an XML-based format, and annotation is automatically applied. This is followed by four more steps:

- Annotation: The human annotator can select a document and manually add, remove, or modify each entity based on rules defined in the guidelines. Once the annotation is complete, the document is marked "ready" for the validation.

- Validation/Curation by domain experts: Annotations from different users are validated and merged into a final document (Figure 5). The domain expert ("curator"), can compare the different annotated versions, and select the best combination of annotations, or add new ones. This step ensures that the annotations are cross-checked and that the document is validated by domain experts.

- Automatic consistency checks and statistical analysis: This step aims to discover obvious mistakes such as mislabelling or incorrect linking. A

sequence labelling model is trained and evaluated using 10-fold cross-validation. The evaluation provides precision, recall, and f-score metrics for all the labels. The resulting model is used for producing machine-based annotated data in the following iteration.

- Review: Retrospective analysis of the past iteration, where unclear cases are discussed and documented in the annotation guidelines.

Data transformation

There are two processes of data transformation (Figure 3): (a) from the source document (PDF) to the dataset format representation (XML-based), and (b) from the dataset format representation to the annotation tool exchange formats (`https://inception-project.github.io/releases/0.16.1/docs/user-guide.html\#sect_formats`) and vice-versa.

- PDF to XML-based: This step converts the PDF source document to the dataset format representation in XML following the Text Encoding Initiative (TEI, `https://tei-c.org/`) format guidelines. Such transformation is performed by leveraging the functionalities provided by GROBID (`https://github.com/kermitt2/grobid`).

  We developed a customised process for collecting a subset of information from the source PDF document. The process extracts the title, keywords, and abstract from the header; and paragraphs, sections. and figure and table captions from the body. All the callouts to references, tables, and figures are ignored. The resulting structured document is then encoded in XML as will be described below.

- XML to the annotation tool exchange formats: We transform our XML-formatted data into an INCEpTIONS compatible import format, such as the Webanno TSV 3.2 (`https://inception-project.github.io/releases/0.17.0/docs/user-guide.html\#sect_formats_webannotsv3`), and vice-versa using a set of Python scripts. The Webanno TSV 3.2 format is an extension of the CONLL (`https://www.signll.org/conll/`) format, with additions of the header and column representation.

Data Record

The dataset is composed of 142 PDF documents, of which 92% (130) are OA (Figure 4a). To comply with copyright restriction, few articles from our dataset are not publicly available in our repository. The top three publishers represented in the corpus are American Physical Society (APS), Elsevier, and IOP Publishing (Figure 4b). Figure 4c illustrate the distribution by publication date. We summarise SuperMat's content in Table 4, with the statistics of documents, entities, and links given separately. In particular, this dataset contains 16052 (7166 unique) entities spread over six labels and 1398 links.

Each document is encoded according to the XML TEI guidelines, which is a rich format for document representation. We have carried out no specific customisation, in order to remain fully compliant with the general TEI schema. A TEI document has two main parts: the header (within the `<teiHeader>`

tags) containing all the document metadata, and the body (within the section delimited by the <text> tag). The transformed data has the following structure:

```
<TEI xml:lang="en" xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>[...]</title>
            </titleStmt>
            <publicationStmt>
                <publisher>[...]</publisher>
            </publicationStmt>
        </fileDesc>
        <encodingDesc/>
            <abstract>
                <p>[...]</p>
                <ab type="keywords">[...]</ab>
            </abstract>
        <profileDesc>
        </profileDesc>
    </teiHeader>
    <text>
        <body>
            <p>[...]</p>
            <ab type="tableCaption"> [...] </ab>
            <p> [...] </p>
            <ab type="figureCaption"> [...] </ab>
        </body>
    </text>
</TEI>
```

We transformed the source documents into these TEI-compliant structures using a simplified representation for specific content types. The general objective is to flatten the content into a generic structure where priority is given to the annotations. For instance, the keywords section, which groups together the key terms defined by the author(s) of the paper, is encoded using the generic tag `<ab type="keywords">` as free text, instead of the dedicated `<keywords>` element that would typically be part of the header. For both the abstract and the article body, the text is segmented in paragraphs (by means of the `<p>` element). The text is annotated with the generic `<rs>` (referencing string) element adorned with three attributes: `@type` (the entity type), `@corresp` (to provide a link to another annotation such as from *material* to T_c), and `@xml:id` (to uniquely identify the annotation for referencing or linking purposes).

Because only the captions of tables and figures are retained from the original source, a simplified encoding was defined by means of the `<ab>` element characterised by a `@type` attribute; that is, `<ab type="figureCaption">` for figure captions and `<ab type="tableCaption">` for table captions. Here is an example:

```
<p>
    The electron-doped high-<rs type="tc">transition-
```

```
    temperature</rs> (<rs type="tc">Tc</rs>) <rs
    type="class">iron-based pnictide</rs>
    superconductor <rs type="material"
    xml:id="m6">LaFeAsO1-xHx</rs> has a unique
    phase diagram: Superconducting (SC) double domes are
    sandwiched by antiferromagnetic phases at ambient
    pressure and they turn into a single dome with
    a maximum <rs type="tc">Tc</rs> that
    <rs type="tcValue" xml:id="m7"
    corresp="#m6,#9">exceeds 45K</rs>
    at a pressure of <rs type="pressure"
    corresp="#m7">3.0 GPa</rs>.
    [...]
</p>
```

In the above snippet, the entities *"3.0 GPa"*, *"exceed 45K"* and *"LaFeAsO1-xHx"* are linked together via the pairs `@corresp, @xml:id`. This schema supports multiple annotations to any part of the document. For example, the entity *exceed 45K* has a second link with the corresponding identifier (*"#9"*) to an annotation outside this paragraph.

Applications

SuperMat is constructed as a resource for TDM applications in superconducting materials. It can be used as data source in several complementary tasks: (1) creation of an automatic information extraction system for dataset creation, (2) articles classification, (3) named entity extraction (for example, automatic dictionary construction), (4) clustering and document synthesis, (5) training of machine learning (ML) algorithms, (6) evaluation of rule-based or ML-based algorithms, and (7) development of downstream processes, such as material name parser, or quantity normalisation.

Reusability

The data structure employed in this study (classes of materials, materials names, and related properties), is similar to that used in other domains in materials science. Therefore, SuperMat can be reused to facilitate or bootstrap the creation of new TDM processes in areas of materials research besides superconductors. SuperMat could be used as a feature for a machine learning model for NER or EL systems in materials science such as magnetocaloric, piezoelectric, and thermoelectric domains.

Practical applications

Such a dataset may benefit several types of possible applications:

- Evaluation tasks: This corpus can be used for evaluation tasks on automatic extraction. In particular, we can envision two popular tasks in superconducting materials science, namely: (a) NER and (b) EL methods. EL techniques have been mainly designed and studied using text

from Wikipedia and newswires services which represent most of the available data. To the best of our knowledge, however, there is no application within materials science.

- Automatic information extraction for superconducting materials: This dataset can be used as training data for such a purpose. Automatic information extraction using ML and text mining techniques can accelerate the construction of databases for superconducting materials.

- Document retrieval: Information retrieval is a key application helping researchers overcome information overload. One way is through query expansion to cover multiple expressions of the same term. By collecting and clustering all expressions under the same concept, it would be possible to retrieve documents when, for example, the resistivity measurement is described by a phrase other than "resistivity". Furthermore, the assigned labels can be used to boost documents where a certain term belongs on a specific label. For example, *cobalt oxide* can appear as either `<material>` or `<class>` depending on the context, while a user would like to obtain documents where *cobalt oxide* appears as `<material>`.

- Weighted-clustering: Scientific document clustering has recently gained growing attention because of its potential capacity for finding additional relevant documents of interest. For example, clustering can help locating similar experimental settings in a large collection of documents. However, clustering documents based on their general content might not be optimal for finding such detailed similarities. Annotation can be leveraged to tilt the clustering algorithm toward entity similarity, which may provide a more focused clustering towards a specific type of information.

Technical Validation

The following measures were employed to ensure the creation of a high-quality dataset:

- Each document was revised and validated by domain experts,

- The workflow begins by assigning machine-based annotated data. This has demonstrated to improve the annotation task over several aspects, namely: time consumption, error rate, and annotation agreement [7, 32, 28].

- On-the-fly automatic annotation recommendations, which provide fresh suggestions based on online decisions made by the annotators.

- The annotators have rapid access to changes in the annotation guidelines.

- The discussions were documented and linked in the guidelines.

- Reviews are discussed and approved collaboratively between domain experts and other annotators.

These guidelines are a vital piece of this work since they contain knowledge accumulated from these activities. However, measuring the completeness of the guidelines is challenging. Assuming that the documents validated by domain experts represent the ground truth, we conducted IAA analysis between different annotators against the ground truth, using the Krippendorf's Alpha metric [27]. Table 5 shows the average IAA which is satisfying with a value of approximately 0.9. The highest score is obtained in the `<material>` entities, while the lowest one is obtained in `<pressure>`, which appears less frequently in the papers. The disagreement in `<tcValue>` can appear to be too low as compared with other labels such as `<class>`, which is, at first look, more ambiguous. We analysed the different cases and identified three reasons why this happens. First, `<tcValue>` may depend heavily on the context that requires more human attention, and it is therefore more prone to errors. Second, our suggestions system is challenged in its ability to disambiguate critical temperatures from other temperature data, leading to incorrect or invalid suggestions. Finally, the presence of mathematical symbols (e.g. "~", "<", and ">") or other modifiers ("`up to`", "`exceeds`", etc.) before the `<tcValue>` could generate small disagreements that accumulate in the average score.

To more precisely isolate the impact of the guidelines, we grouped the IAA results by level of domain experience. Table 6 displays the IAA between the validated data and the data corrected by (a) domain experts (researchers who conduct superconducting development experiments), (b) non-domain-experts (researchers with no experience with superconducting materials), and (c) novices (students in materials science with limited domain experience). Obviously, the domain experts have the highest agreement and the IAA value (around 0.95) is 0.06 higher on average than that of non-domain experts. Thus, superconducting materials is a complex domain that requires knowledge in materials science to produce high-quality data, while crowdsourcing initiatives such as the Amazon Mechanical Turk might not work well.

Furthermore, we measured the reliability of the guidelines by observing how quickly novices could reach a satisfying agreement with the validation of the domain experts, without any previous training on the guidelines. From Table 6, the novices can attain high IAA results by only using the guidelines and our annotation support tools. The average difference in agreement with domain experts (around 0.05) indicates that the guidelines are precise and complete, and that the annotations tools offer sufficient support.

Data Availability

The references of the original papers and their bibliographical information, the annotation guidelines, and the developed code are available at the GitHub repository `https://github.com/lfoppiano/SuperMat`. The data transformation scripts were written in Python and can be run from the command line. They require BeautifulSoup (`https://www.crummy.com/software/BeautifulSoup/`), an open-source library for parsing XML and HTML formats. The data analysis scripts were developed as Jupyter notebooks (`https://jupyter.org/`) which can easily output results and graphs in the browser. The open source annotation tool is INCEpTION (`https://inception-project.github.io/`). The content was acquired using biblio-glutton (`https://www.github.com/kermitt2/`

biblio-glutton) and Grobid (https://www.github.com/kermitt2/grobid). We computed the IAA using the Java library DkPro statistics (https://dkpro.github.io/dkpro-statistics/) [31].

Conclusions

In this paper we described the construction of an annotated linked dataset from scientific publications on superconductors development. SuperMat aims to establish a solid infrastructure where to build or improve TDM processes in superconductor materials domain. We annotated 142 full-text articles where the data was automatically extracted from the PDF document and encoded through the XML TEI guidelines providing a basic structure of the original document. The dataset is validated by domain experts and provides 16052 entities of six categories, and 1398 links between materials, properties and conditions. This approach can be extended to other materials domains following similar methodology.

References

[1] Bo-Christer Björk, Annikki Roos, and Mari Lauri. Scientific journal publishing: yearly volume and open access availability. *Inf. Res.*, 14, 2009.

[2] Laura Cardani, Francesco Bellini, Nicola Casali, M. G. Casellano, Ivan Colantoni, Alessandro Coppolecchia, Carlo Cosmelli, Angelo Cruciani, A. D'Addabbo, Sergio Di Domizio, Mario Martinez, Carlos Tomei, and M Vignati. New application of superconductors: High sensitivity cryogenic light detectors. *Nuclear Instruments & Methods in Physics Research Section A-accelerators Spectrometers Detectors and Associated Equipment*, 845:338–341, 2017.

[3] Clarissa F. D. Carneiro, Victor G. S. Queiroz, Thiago C. Moulin, Carlos A. M. Carvalho, Clarissa B. Haas, Danielle Rayêe, David E. Henshall, Evandro A. De-Souza, Felippe E. Amorim, Flávia Z. Boos, Gerson D. Guercio, Igor R. Costa, Karina L. Hajdu, Lieve van Egmond, Martin Modrák, Pedro B. Tan, Richard J. Abdill, Steven J. Burgess, Sylvia F. S. Guerra, Vanessa T. Bortoluzzi, and Olavo B. Amaral. Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. *Research Integrity and Peer Review*, 5(1):16, December 2020.

[4] Thaer M. Dieb, Masaharu Yoshioka, and Shinjiro Hara. Nadev: An annotated corpus to support information extraction from research papers on nanocrystal devices. *Journal of Information Processing*, 24:554–564, jan 2016.

[5] Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[6] Luca Foppiano, Thaer M. Dieb, Akira Suzuki, and Ishii Masashi. Proposal for automatic extraction framework of superconductors related information from scientific literature. *THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS*, 2019.

[7] Karën Fort and Benoît Sagot. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[8] Michael W. Gaultois, Taylor D. Sparks, Christopher K. H. Borg, Ram Seshadri, William D. Bonificio, and David R. Clarke. Data-driven review of thermoelectric materials: Performance and resource considerations. *Chemistry of Materials*, 25(15):2911–2920, 2013.

[9] Vincas Grigas, Simona Juzeniene, and Jone Velickaite. 'just google it' - the scope of freely available information sources for doctoral thesis writing. *Inf. Res.*, 22, 2017.

[10] James J. Hamlin. Superconductivity near room temperature. *Nature*, 569:491–492, 2019.

[11] Lezan Hawizy, David M. Jessop, Nico Adams, and Peter Murray-Rust. Chemicaltagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, 3:17 – 17, 2011.

[12] Min He, Yi Wang, and Wei Li. Ppi finder: A mining tool for human protein-protein interactions. *PLOS ONE*, 4(2):1–6, 02 2009.

[13] J. Hill, Gregory J. Mulholland, K. Persson, R. Seshadri, C. Wolverton, and B. Meredig. Materials science with large-scale data and informatics: Unlocking new opportunities. *Mrs Bulletin*, 41:399–409, 2016.

[14] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[15] Duff Johnson. Pdf statistics – the universe of electronic documents, 2018-05-14.

[16] Madian Khabsa and C. Lee Giles. The number of scholarly documents on the public web. *PLoS ONE*, 9, 2014.

[17] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–2, 2003.

[18] Kaname Kizu, Katsuhiko Tsuchiya, Yozo Kashiwa, Haruyuki Murakami, and Kôichi Yoshida. Construction of the jacketing facility and first production results of superconductor for jt-60sa. *IEEE Transactions on Applied Superconductivity*, 20:538–542, 2010.

[19] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018.

[20] Tomohiko Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, Ryo Ogawa, I. Hosako, and A. Maeda. Deep learning model for finding new superconductors. *ArXiv*, abs/1812.01995, 2018.

[21] Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):203, October 2019.

[22] H. Kotegawa, T. Kawazoe, H. Tou, K. Murata, H. Ogino, K. Kishio, and J. Shimoyama. Contrasting pressure effects in sr2vfeaso3 and sr2scfepo3. *arXiv: Superconductivity*, 2009.

[23] Martin Krallinger, José M. G. Izarzugaza, Carlos Rodríguez Penagos, and Alfonso Valencia. Extraction of human kinase mutations from literature, databases and genotyping studies. *BMC Bioinformatics*, 10:S1 – S1, 2009.

[24] Martin Krallinger, Florian Leitner, and Alfonso Valencia. *Analysis of Biological Processes and Diseases Using Text Mining Approaches*, pages 341–382. Humana Press, Totowa, NJ, 2010.

[25] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Dong-Hong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, P. Senthil Nathan, Slavko Zitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin M. Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, K. E. Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usie, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzábal, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2 – S2, 2015.

[26] Alexander Krasnitz. Cancer bioinformatics. In *Methods in Molecular Biology*, 2019.

[27] Klaus Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433, 01 2006.

[28] Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical

named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association : JAMIA*, 21(3):406–413, 2014.

[29] Huaping Liu, Feng Wang, Fuchun Sun, and Bin Fang. Surface material retrieval using weakly paired cross-modal learning. *IEEE Transactions on Automation Science and Engineering*, 16:781–791, 2019.

[30] Philippe Mangin and Rémi Kahn. *Superconductivity: An introduction.* Springer, 2016.

[31] Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. DKPro agreement: An open-source Java library for measuring inter-rater agreement. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 105–109, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[32] Aurélie Névéol, Rezarta Islamaj Dogan, and Zhiyong Lu. Semi-automatic semantic annotation of pubmed queries: A study on quality, efficiency, satisfaction. *Journal of biomedical informatics*, 44 2:310–8, 2011.

[33] Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. Incorporating genetag-style annotation to genia corpus. In *BioNLP@HLT-NAACL*, 2009.

[34] Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317, 2020.

[35] Enrique Orduña-Malea, Juan Manuel Ayllon, Alberto Martín-Martín, and Emilio Delgado López-Cózar. Methods for estimating the size of google scholar. *Scientometrics*, 104:931–949, 2015.

[36] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. "* O'Reilly Media, Inc.", 2012.

[37] Valentin Stanev, Corey Oses, A. Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and I. Takeuchi. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4, 09 2017.

[38] Lorraine K. Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6:S3 – S3, 2005.

[39] Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. SC-CoMIcs: A superconductivity corpus for materials informatics. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6753–6760, Marseille, France, May 2020. European Language Resources Association.

[40] Shigeki Yonezawa, Y. Muraoka, and Z. Hiroi. New ß-pyrochlore oxide superconductor csos2o6. *Journal of the Physical Society of Japan*, 73:1655–1656, 2004.

[41] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16(1):93, August 2016.

Acknowledgements

We would like to thank Tanifuji Mikiko for her continuous support, as well as the enthusiasm and the openness with which she lead the Data PlatForm Data Center (DPFC, `https://www.nims.go.jp/eng/research/materials-data-pf/index.html`) at NIMS. Our warmest thanks to Patrice Lopez, the author of Grobid (`https://github.com/kermitt2/grobid`) and other TDM open-source projects.

Author contributions statement

L.F. designed and developed the work (data preparation, annotation tools, IAA experiments, automatic annotations). M.I. and Y.T. supervised the project. L.R. defined the standardised dataset TEI format. L.F. S.D. A.S. A.U. M.G.E.E. P.B.C. Y.M. S.I., and K.T. performed the dataset annotation and validation. M.G.E.E. P.B.C. Y.M. S.I. K.T., and Y.T. validated the corpus. L.F. wrote the manuscript with assistance in editing from S.D., M.I., K.T., P.B.C., Y.M., and M.G.E.E.. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables

Iteration #	**IAA**	**IAA by label**	
1	0.45	`<material>`	0.45
		`<tc>`	0.56
		`<tcValue>`	0.50
		`<doping>`	0.21
2	0.65	`<material>`	0.75
		`<tc>`	0.85
		`<tcValue>`	0.85
		`<doping>`	0.39
3	0.89	`<material>`	0.89
		`<tc>`	0.91
		`<tcValue>`	0.88
		`<doping>`	0.94

Table 1: Summary of the IAA for each annotation iteration.

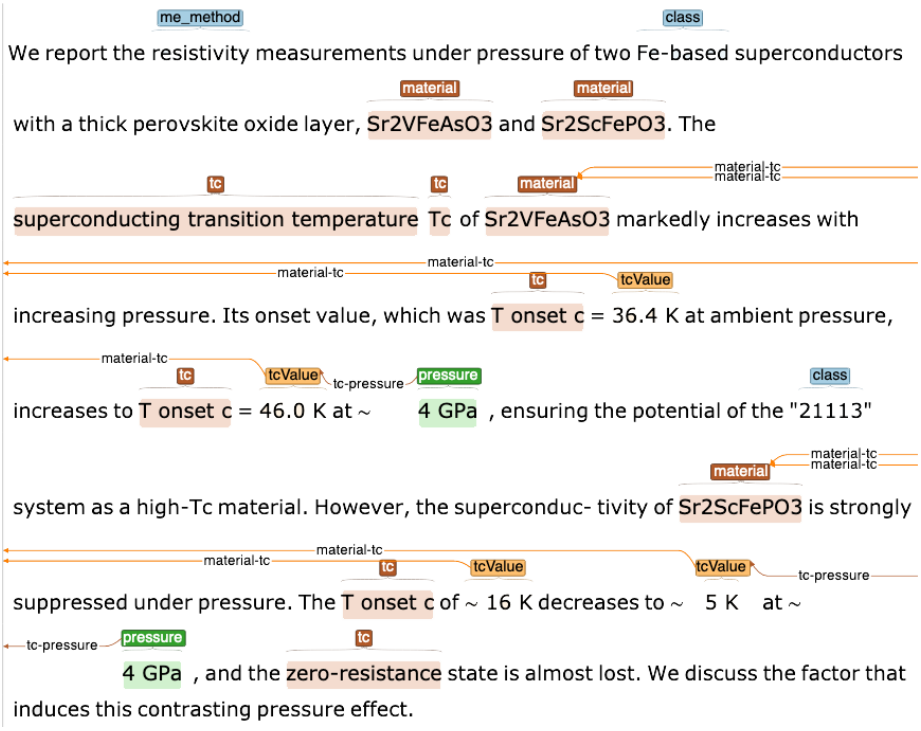


Figure 1: Example in the annotated corpus. The excerpt was taken from [22].

Text	Label 1	#	Label 2	#
LiFeAs	<material>	89	<class>	1
Bi-2212	<material>	34	<class>	1
cobalt oxide	<material>	89	<class>	1
RE-123	<material>	34	<class>	1

Table 2: Inconsistencies resulting from human mistakes.

Text	Label 1	#	Label 2	#
superconducting transition	<material>	1	<tc>	61
NCCO	<material>	14	<tc>	1
superconducting transition temperatures	<material>	1	<tc>	11
occurrence of superconductivity	<material>	1	<tc>	1

Table 3: Inconsistencies resulting from the overlapping of <material> and <class> labels.

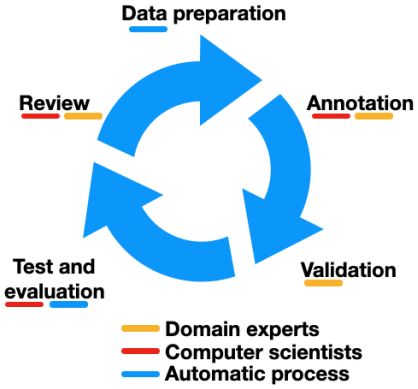


Figure 2: Annotation workflow. Different colours illustrate the involvement of each group at each step of the workflow.

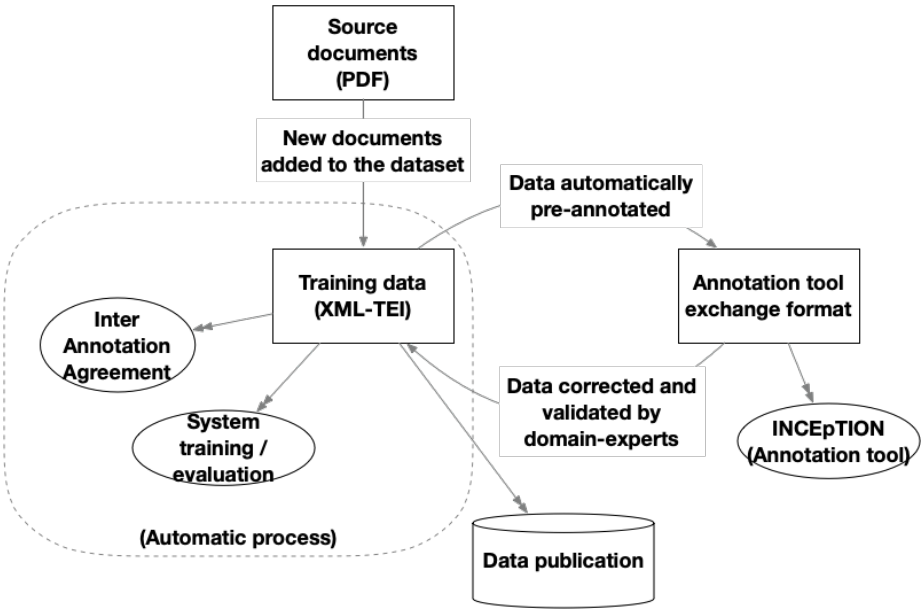
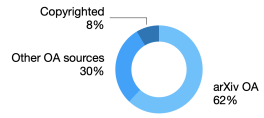


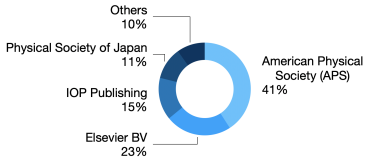
Figure 3: Summary of the data transformation flows.

Documents	**Files** 142	**Paragraphs** 2800	**Sentences** 18344	**Tokens** 1118432
Entities	**Entities** 16040	**Unique entities** 7151		**Labels** 6
Links	**Links** 1399	**Links$_{ip}$** 1286		**Links$_{ep}$** 113

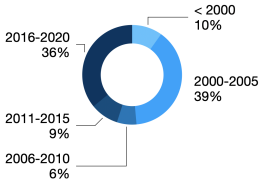
Table 4: Statistical overview of the dataset. Links$_{ip}$ indicates the number of links within the same paragraph (intra-paragraph). Links$_{ep}$ indicate the number of links from different paragraphs (extra-paragraphs).



(a) Papers distribution by Licence: Open Access (CC-BY) vs copyrighted.



(b) Distribution by publisher.



(c) Distribution by year of publication.

Figure 4: Distribution of paper in the dataset by (a) license, (b) publisher, and (c) year of publication.

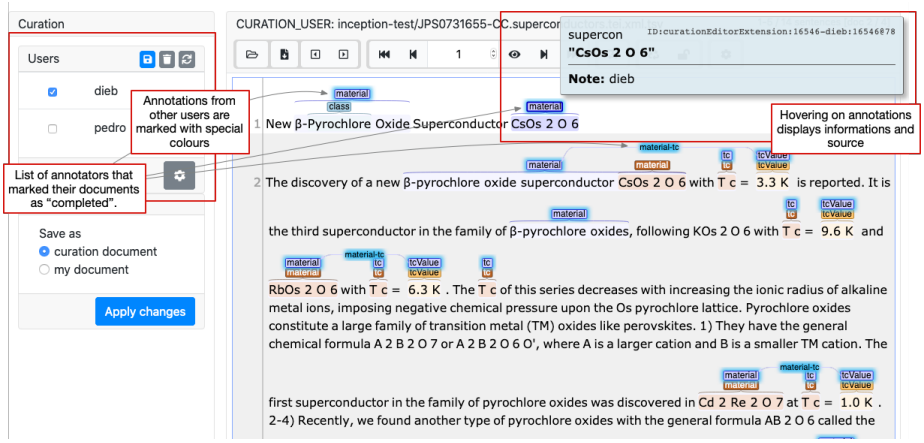


Figure 5: INCEpTION curation interface. The example is taken from [40].

Label	Average
<material>	0.956
<me_method>	0.887
<pressure>	0.723
<class>	0.925
<tcValue>	0.863
<tc>	0.831
Micro average	0.911

Table 5: Average IAA between the annotated and validated documents

Label	Domain experts	Non-domain experts	Novices
<material>	0.969	0.950	0.924
<me_method>	0.890	0.862	0.901
<pressure>	0.836	0.741	0.746
<class>	0.990	0.836	0.899
<tcValue>	0.895	0.734	0.841
<tc>	0.874	0.776	0.830
All labels	0.940	0.882	0.896
# paragraphs	1066	1648	325

Table 6: Calculated IAA for annotations produced by domain experts, non-domain experts, and novices compared to the validated version. Annotations from domain experts are cross validated.