ISSN 1674-1056
http://cpb.iphy.ac.cn
http://iopscience.iop.org/cpb
中国物理 B
Chinese Physics
中国物理
Chinese Physics B

Accomplishment and challenge of materials database toward big data

Yibin Xu(徐一斌)

Citation: Chin. Phys. B . 2018, 27(11): 118901. **doi:** 10.1088/1674-1056/27/11/118901
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Generation of optimal persistent formations for heterogeneous multi-agent systems with a leader constraint

Guo-Qiang Wang(王国强), He Luo(罗贺), Xiao-Xuan Hu(胡笑旋)
Chin. Phys. B . 2018, 27(2): 028901. **doi:** 10.1088/1674-1056/27/2/028901

Statistical physics of hard combinatorial optimization：Vertex cover problem

Zhao Jin-Hua, Zhou Hai-Jun
Chin. Phys. B . 2014, 23(7): 078901. **doi:** 10.1088/1674-1056/23/7/078901

Generation of minimally persistent circle formation for a multi-agent system

Luo Xiao-Yuan, Shao Shi-Kai, Zhang Yu-Yan, Li Shao-Bao, Guan Xin-Ping, Liu Zhi-Xin
Chin. Phys. B . 2014, 23(2): 028901. **doi:** 10.1088/1674-1056/23/2/028901

Output regulation for linear multi-agent systems with unmeasurable nodes

Liang Hong-Jing, Zhang Hua-Guang, Wang Zhan-Shan, Wang Jun-Yi
Chin. Phys. B . 2014, 23(1): 018902. **doi:** 10.1088/1674-1056/23/1/018902

SPICE modeling of flux-controlled unipolar memristive devices

Fang Xu-Dong, Tang Yu-Hua, Wu Jun-Jie, Zhu Xuan, Zhou Jing, Huang Da
Chin. Phys. B . 2013, 22(7): 078901. **doi:** 10.1088/1674-1056/22/7/078901

TOPICAL REVIEW — Physics research in materials genome

Accomplishment and challenge of materials database toward big data*

Yibin Xu(徐一斌)†

National Institute for Materials Science, 1-2-1 *Sengen, Tsukuba, Ibaraki* 3050047, *Japan*

(Received 4 June 2018; revised manuscript received 16 September 2018; published online 25 October 2018)

The history and current status of materials data activities from handbook to database are reviewed, with introduction to some important products. Through an example of prediction of interfacial thermal resistance based on data and data science methods, we show the advantages and potential of material informatics to study material issues which are too complicated or time consuming for conventional theoretical and experimental methods. Materials big data is the fundamental of material informatics. The challenges and strategy to construct materials big data are discussed, and some solutions are proposed as the results of our experiences to construct National Institute for Materials Science (NIMS) materials databases.

Keywords: material database, big data, material informatics, machine learning, interfacial thermal resistance, material identification

PACS: 89.20.Ff, 01.65.+g, 65.80.–g **DOI:** 10.1088/1674-1056/27/11/118901

1. Introduction

We witness the progress and victory of artificial intelligence (AI) technology in this decade in various fields of our life, from playing chess to automated driving. AI is expanding into most of human activities immensely and overwhelmingly, no exception for science and technology, a representative of crystallization of human wisdom. Today, more and more material scientists start to believe and dream that someday, we will just order a required material function to a robot, and then the robot will design and produce the material for us. However, to generate such a system, tremendous materials experiences and knowledge are necessary for making the robot "clever" and "reliable". Then we face to a realistic question: do we have enough accumulation of data to train the robot? If not, what should we do? For the first question, most people will agree that the answer is negative. Then the second question becomes the key point to realize the dream. This paper reviews the history and current status of materials data and database, and discusses the future strategy to construct materials big data.

2. Review of world materials data and databases

The evolution of data management and dissemination technology of materials data can be divided into three stages: handbook, database, and big data, and the main purpose and application of the materials data systems at each stage change from reference data to material selection and to material development (Fig. 1).

2.1. Materials data handbook

The collection of materials data started from the 1880s. At that time, the data sets were published as handbooks. Beilstein collected data of properties, spectra, and preparation of organic compounds from literatures, and published the first edition of his handbook in 1881.[1] Up to 1989, the 4th edition of Beilstein handbook was published in 503 volumes (over 440000 pages). Since 2009, the content has been maintained and distributed by Elsevier information systems in Frankfurt under the product name "Reaxys".[2]

Similar work has been done by Landolt and Börnstein[3] since 1883. The current edition of Landolt–Börnstein handbook includes 195000 pages. More than 72000 element systems, 150000 chemical substances, and 530000 substance–property pairs are included. SpringerMaterials[4] is a web platform, which provides data service based on Landolt–Börnstein handbook.

The Journal of Physical and Chemical Reference Data[5] is published by American Institute of Physics (AIP) Publishing since 1972. One source of the journal contributions is the national standard reference data system (NSRDS), which was established in 1963 to coordinate the production and dissemination of critically evaluated reference data in physical sciences. National Institute of Standards and Technology (NIST) coordinates several data evaluation centers, located in universities, industrial, and government laboratories, as well as within NIST.

In Japan, National Research Institute for Metals (NRIM) published the first NRIM creep data sheet in 1966 and fatigue

*Project supported by "Materials Research by Information Integration" Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST).

†Corresponding author. E-mail: xu.yibin@nims.go.jp

© 2018 Chinese Physical Society and IOP Publishing Ltd

data sheet in 1976. This work has been continued after the reorganization of NRIM into National Institute for Materials Science (NIMS) in 2001. And NIMS published the first NIMS corrosion data sheet in 2002. Up to now, 57 volumes of creep data sheet, 123 volumes of fatigue data sheet, and 4 volumes of corrosion data sheet have been published.[6–8]

Besides these large-scaled comprehensive chemical and physical data collections, many handbooks on specific material properties have also been published. For example, ASM handbook[9] on mechanical properties, and Thermophysical Properties Research Center (TPRC) handbook[10] on thermosphysical properties.

The data handbooks or journals are featured by high-quality data evaluated and selected by experts with high reputation in these fields. The editors did not only put those data together, but also classified and indexed them. For example, Beilstein developed the Beilstein system to classify compounds according to their constitutional features.[1]

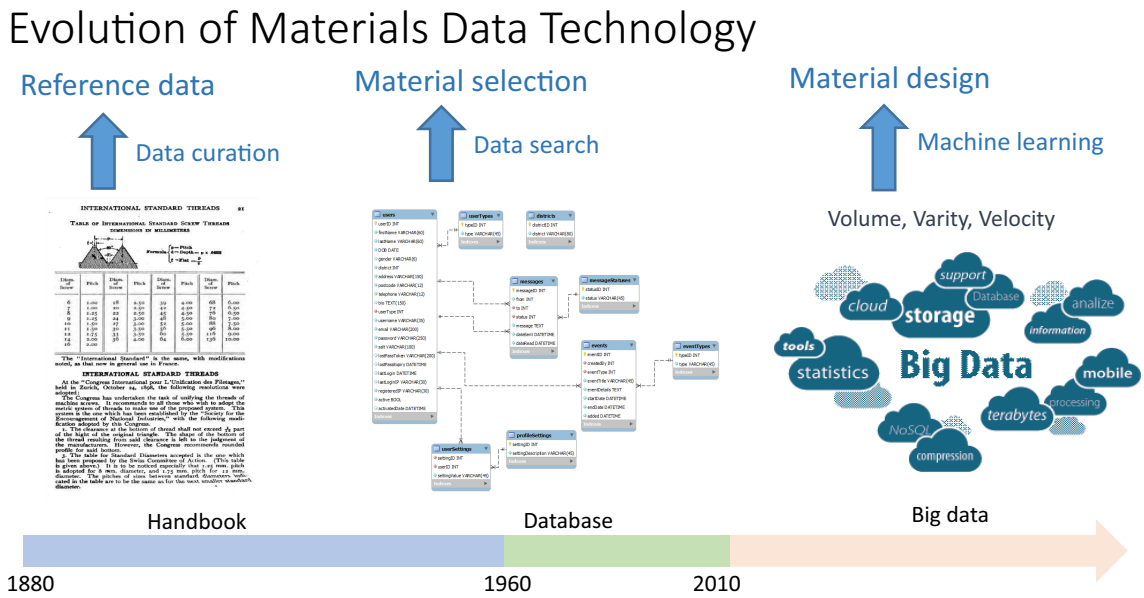


Fig. 1. (color online) Evolution of materials data technology.

2.2. Materials database

The computerized database started in the 1960s. In 1970, Codd[11] published an important paper to propose the use of a relational database model, and his ideas changed the way people thought about databases. In his model, the schema or logical organization of the database is disconnected from physical information storage, and this became the standard principle for database systems. Since then, various relational database system products have been developed, such as MS SQL Server, DB2, Allbase, Oracle, etc.

The database enables us to store, update, and search a large amount of data, quickly, accurately, and securely. It also makes it very easy to transform and reorganize the data, and to generate a new data set for new purposes.

Since the 1990 s, some materials databases have been developed, e.g., Pauling File project,[12] which was launched in 1995 as a collaboration between Japan Science and Technology Agency (JST) and Material Phases Data System (MPDS), a Swiss company. Since January 2016, the data has been copyrighted by NIMS and MPDS. The first goal of Pauling File project is to create and maintain a comprehensive materials database for non-organic (no C–H bonds) solid state materials, covering phase diagrams, crystallographic data, diffraction patterns, and physical properties. Up to now, the Pauling File data have been contained in 12 products. AtomWork[13] and AtomWork Adv.[14] are two database systems developed by NIMS. AtomWork is a free database opened to the Internet as a part of NIMS materials database system MatNavi,[15] which contains the Pauling File data before 2001. AtomWork Adv. is a fee-based database released in 2018, which contains the full Pauling File data and will be updated annually. The present version of AtomWork Adv. contains 42406 phase diagrams, 303885 crystal structures, 550507 x-ray diffraction patterns, and 365517 materials properties, extracted from 141490 publications.

In these years, first principle computation has become an efficient method to generate materials data such as electronic structures and properties. Many databases have been established in this way. For example, Novel Materials Discovery (NOMAD) repository[16] is developed as a joint project of Fritz-Haber-Institut (FHI), Humboldt-Universität zu Berlin (HUB), and Max Planck Computer and Data Facility (MPCDF). Materials Project[17] is an integrated framework of data and analysis tools developed by University of California at Berkeley and Lawrence Berkeley National Laboratory.

3. Materials data science and informatics

Materials science started with experiments. By concluding and sublimating the experimental data and experiences, we establish knowledge and theories. Then as application of the theories, computation has become an important method to predict the property and find new materials. Finally, the results of theories and computations are verified by experiments. Data science is expected to accelerate and improve this process, because today's computer can store and process such large volumes of data that it allows us to set up algorithms based correlations among material phenomena without understanding the physical and chemical process underground. It is especially useful for studying materials phenomena which are too complex to be modeled by physical or chemical methods. For this purpose, we need data captured from various phenomena with respect to material composition, processing, structure, property, and performance, and with different characteristic spatial and temporal scales.

3.1. "Materials Research by Information Integration Initiative" project

In 2015, a Japanese national project "Materials Research by Information Integration Initiative (MI^2I)" has been launched. The purpose of this project is to accelerate material exploration and development by combining materials data and data science method. As the fundamental of this project, a data platform MI^2I-DPF[18] has been constructed. MI^2I-DPF is an integration of materials data, data application tools, and computational environment. MI^2I-DPF is contemporary opened to the project members and the members of the industrial consortium. The users can access and download nearly a million data entries from 6 main databases of MatNavi through application programming interface (API), and use the tools to analysis or do simulation with the data on the platform. An example of the research results of this project using this system is as follows.

3.2. Study on interfacial thermal resistance by data science approach

Various material interfaces exist in devices and materials such as composites, alloys, and ceramics. Especially for nanostructured materials, the properties of interfaces play important roles to determine the properties of the materials. However, interface is one of the most difficult issues in material science, because it is affected by numerous equilibrium and nonequilibrium factors.

Interfacial thermal resistance (ITR) is caused by scattering or reflection of phonons and electrons at the interface. Two physical models,[19] acoustic mismatch model (AMM) and diffusion mismatch model (DMM), have been set up to calculate ITR of an interface based on the physical properties, concretely, density, acoustic velocity, and unit cell volume of the materials at two sides. However, comparing the calculation results to the experimental data, we can see that the accuracy of the calculation is not satisfied (Fig. 2).[20] The reasons of the deviation between theory and experiment can be considered as: (i) the property data used in the calculation has large uncertainty, because properties such as acoustic velocity obtained by different methods or reported by different papers usually have large deviation; (ii) the descriptors used in AMM and DMM models are not sufficient. Except the above physical properties, ITR may be affected by many other chemical and materials factors. Some experimental studies have confirmed that the chemical bonding and interfacial structure are factors with significant influences on ITR.[21,22]

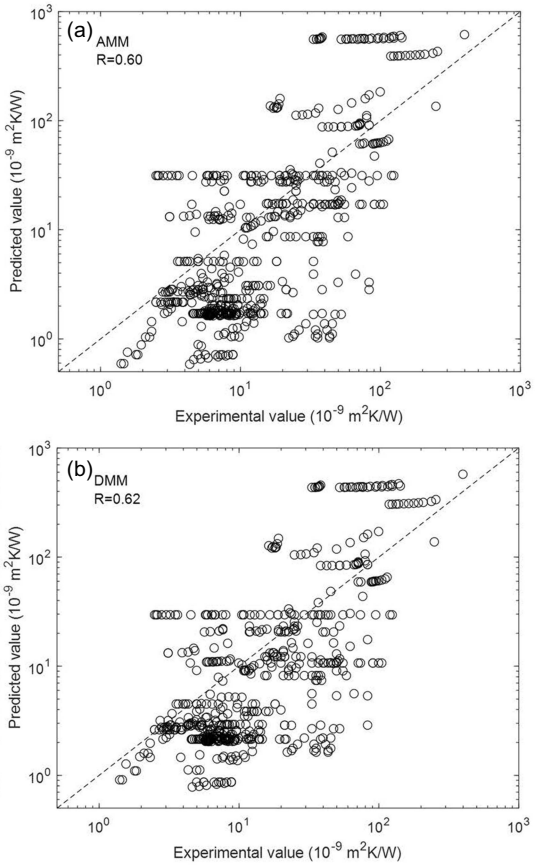


Fig. 2. Correlation between the experimental values and the values predicted by the (a) AMM and (b) DMM.[20]

We try to establish a data model[20] to predict ITR with higher accuracy from two approaches: (I) avoiding to use physical properties with high uncertainty; (II) introducing descriptors other than physical properties. To find an appropriate set of property descriptors, we check the correlations among 11 heat transfer related properties, and select 4 properties which are relatively independent and have less uncertainty. The optimized property descriptor set includes specific heat, melting point, density, and unit cell volume. We also introduce a new descriptor film thickness, since ITR measurements are usually done on samples with a film deposited on a

substrate, and film thickness is one of the most basic parameters of film deposition. Although there is no clear theoretical explanation of the influence of film thickness on ITR, the data show an obvious correlation between them. With the new descriptor set and machine learning methods, we obtain much better prediction of ITR than AMM and DMM (Fig. 3).[20]

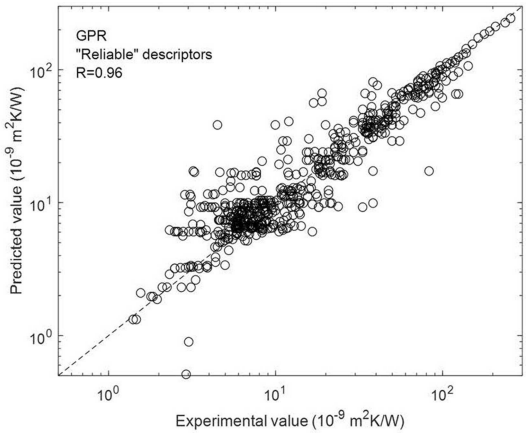


Fig. 3. Correlation between the experimental values and the prediction by Gaussian process regression method.[20]

This example shows the potential of data science in material study. Because it neglects the complicated physical and chemical process between material composition, material structure and material property, it is especially suitable for studies on subjects which cannot be modeled by physical and chemical models.

Data shortage is the biggest difficulty when we apply data science methods to material science. Although the importance of materials database has been well noticed today, data accumulation takes long terms. Nevertheless, for materials, the correlations among composition, structure, and properties allow one descriptor to be substituted by another one or more descriptors, which provides us a possibility to select descriptors with relatively sufficient and high quality data.

4. From materials database to big data

The voluminosity and complexity of big data make the traditional data capturing and processing methods difficult to deal with. To construct materials big data, we face challenges of data capture, data storage, data analysis, search, sharing, visualization, information privacy, and data source. Some of these challenges are expected to be solved by progresses of information technology, however, some of them are problems of material science. In this section, we focus our discussion on some issues specified for materials data.

4.1. Material identification

The first problem when we try to combine data from different data sources is to identify the similarity between two materials, because in most cases, the data are obtained on different samples. Scientifically speaking, material can be defined by its chemical composition and structure. However, the characterization of a material's structure is not an easy task, because it covers a large range of scales from atomic to macroscopic scale. Practically, many people use the process conditions to identify a material, however, the conditions are equipment dependent. MatML[23] is a specification from NIST designed for the interchange of materials information. It uses chemical composition and processing conditions to describe a material. Most databases use only chemical composition or chemical formula to identify a material; in these case, materials with different structures are not distinguishable. Based on our experiences with the data of single crystal, ceramics, alloys, and polymers, we have developed a material identification system according to the fundamentals of materials science. In this system, materials are identified at four different levels: chemical system, compound, substance, and material. Figure 4 shows the four levels and the identifiers of each level.

Chemical system is the fundamental of all materials. It indicates the element or elements of which the materials is composed. Compound is the second level, which identifies a material at molecular level. For most inorganic materials, the compound can be defined by its chemical formula. However, for organic or polymer materials, molecular structure must be specified. The third level is substance. At this level, the state (gas, liquid, or solid) of a compound should be defined. For solid state, the crystalline state and crystal structure should also be given. In most cases, a substance corresponds to a phase in a phase diagram. The forth level is material. To define a material, many types of information are needed, for example, the material form, size, micro-structure, process condition, etc. It is difficult to define a general descriptor set applicable to all types of material. Since samples prepared by different labs

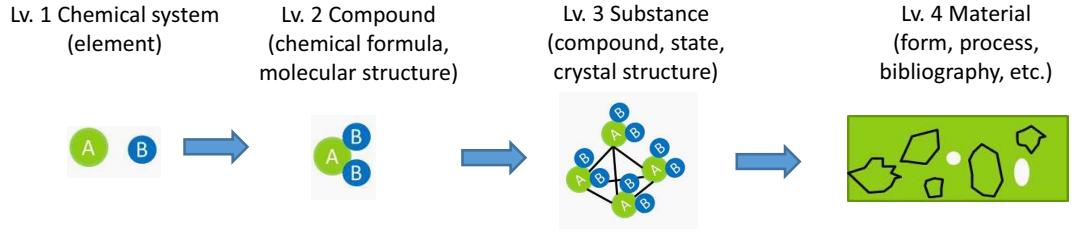


Fig. 4. (color online) Four-level material identification system.

or under different processing conditions are usually different in micro-structure and properties, we treat different samples as different materials, except that the data providers indicate that they are identical at the material level. The four-level system allows us to distinguish each individual sample, meanwhile, to cluster materials according to their common physical and chemical features. Even two samples are different at the material level, if they belong to the same substance, compound, or chemical system, we can still find the relationship between them.

4.2. Link from single phase material to complex material systems

The above material identification system provides a possibility to compare and link materials data at the level of material, substance, compound, and element. However, many practically used materials are composed by multiple substances or materials. In order to make such complex materials comparable, we have designed a hierarchical structure for material description, which can be used to describe materials from atoms to composites. As an example, the hierarchical structure of the SiC/Ti alloy composite is shown in Fig. 5.

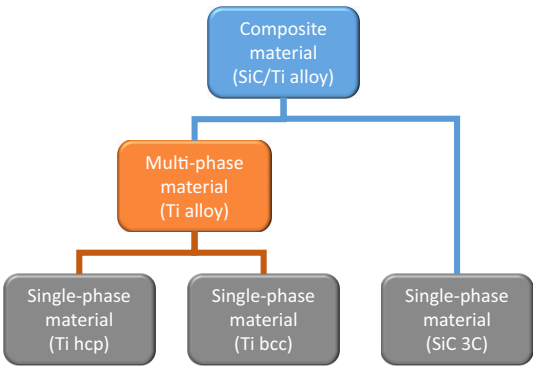


Fig. 5. (color online) An example of hierarchical structure to describe the SiC/Ti composite.

4.3. Material identification record management system

Since material identification is a common issue when integrating data from different data sources, a system which manages the identification record of materials in different resources (Fig. 6) is efficient for this purpose. To construct such a system, a standardized format of material identification record is necessary. We have designed an XML format based on the four-level material identification method. In this format, we prepare 9 types of materials: atom, molecular, cluster, single phase material, multi-phase material, composite, interface, organic, and polymer. A similar hierarchical structure of Fig. 5 is used to describe a complex material. For example, a multi-phase material is composed of multiple single phase materials, and a composite is composed by multiple materials. As an example, the XML schema for single phase material is shown in Fig. 7.

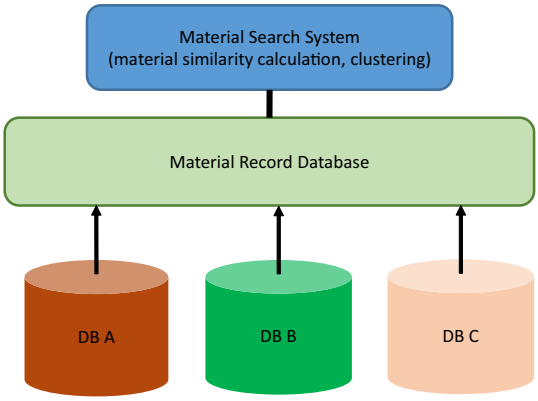


Fig. 6. (color online) Concept of material identification information management system.

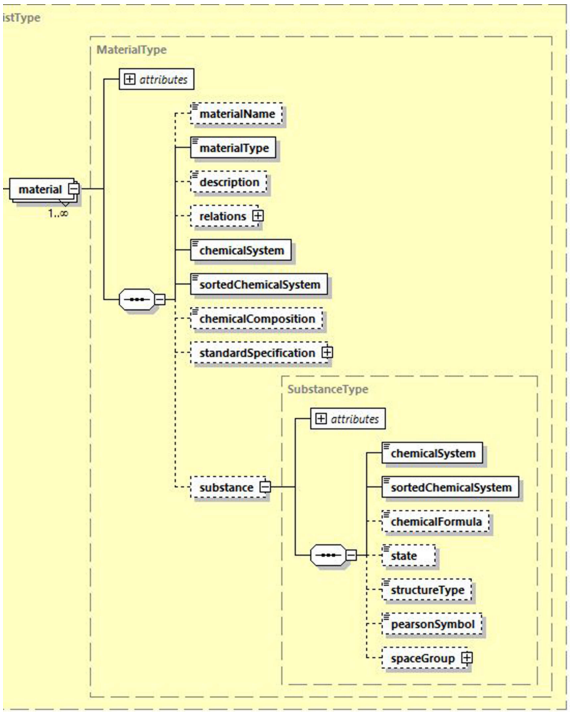


Fig. 7. (color online) XML schema of material identification record for single phase material.

5. Conclusion

Data science is becoming an important and efficient approach to deal with material issues which are too time consuming for experiments and too complicated for theoretical and computational methods. Materials data is the fundamental, as well as a bottleneck in this scientific innovation. In order to construct materials big data, we are facing to various challenges including up-to-date data collection, long-term data conservation, data sharing, and maximal utilization, etc. Fortunately, physical and chemical laws under all material phenomena can help us to reduce the number of data required. Some solutions have been proposed in this paper as the results of our experiences to construct NIMS materials databases. A successful example has been shown to predict interfacial thermal resistance by machine learning methods.

References

[1] Luckenbach R 1981 *J. Chem. Inf. Comput. Sci.* **21** 82

[2] Reaxys, Elsevier, https://www.elsevier.com/solutions/reaxys [2018-05-25]

[3] SpringerMaterials–The Landolt–Börnstein Database, SpringerMaterials, Customer Presentation, 2009

[4] SpringerMaterials, https://materials.springer.com [2018-05-25]

[5] Journal of Physical and Chemical Reference Data, https://www.nist.gov/srd/journal-physical-and-chemical-reference-data-jpcrd [2018-05-25]

[6] Creep data sheet, http://smds.nims.go.jp/creep/index_en.html [2018-05-25]

[7] Fatigue data sheet, http://smds.nims.go.jp/fatigue/index_en.html [2018-05-25]

[8] Corrosion data sheet, http://smds.nims.go.jp/corrosion/index_en.html [2018-05-25]

[9] ASM handbooks, https://www.asminternational.org/handbooks [2018-05-25]

[10] Touloukian Y S and Ho C Y 1970 *Thermophysical Properties of Matter, the TPRC Data Series* (Purdue University)

[11] Codd E F 1972 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.86.9277 [2018-05-25]

[12] The Pauling File project, http://paulingfile.com/index.php?p=scope#the PAULINGFILE project [2018-05-25]

[13] Inorganic materials database AtomWork, http://crystdb.nims.go.jp/index_en.html [2018-05-25]

[14] Inorganic materials database AtomWork Adv., https://atomwork-adv.nims.go.jp [2018-05-28]

[15] NIMS materials database MatNavi, http://mits.nims.go.jp/index_en.html [2018-05-25]

[16] NOMAD repository, http://nomad-repository.eu [2018-05-25]

[17] The Materials Project, https://materialsproject.org [2018-05-25]

[18] Materials Research by Information Integration Initiative Data Platform, https://mi2i.nims.go.jp/en [2018-05-25]

[19] Swartz E T and Pohl R O 1989 *Rev. Mod. Phys.* **61** 605

[20] Zhan T, Fang L and Xu Y 2017 *Sci. Rep.* **7** 7109

[21] Xu Y, Kato R and Goto M 2010 *J. Appl. Phys.* **108** 104317

[22] Zhan T, Xu Y, Goto M, Tanaka Y, Kato R and Sasaki M 2015 *RSC Adv.* **5** 49703

[23] MatML, https://www.matml.org/overview.htm [2018-09-16]