
# The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome

Evgeny Blokhin and Pierre Villars

## Contents



**Abstract**

One of the oldest initiatives in materials informatics, the PAULING FILE project, is described. It includes the comprehensive database for inorganic crystalline

E. Blokhin (✉)
Materials Platform for Data Science, Tallinn, Estonia

Tilde Materials Informatics, Berlin, Germany
e-mail: eb@tilde.pro

P. Villars
Material Phases Data System, Vitznau, Switzerland
e-mail: villars.mpds@bluewin.ch



W. Andreoni, S. Yip (eds.), *Handbook of Materials Modeling*,
https://doi.org/10.1007/978-3-319-42913-7_62-1

compounds, their atomic structures, intrinsic physical properties and phase diagrams. On top of that, the powerful online retrieval software is introduced, called MPDS, the Materials Platform for Data Science. The practical recipes of storage, exchange and analysis of the large amounts of materials data are given. The focus is made on the modern information technologies and software engineering. As a result, from the large heterogeneous data, holistic conclusions about the entire set of known materials are drawn. They can be regarded as a guideline for the systematic large-scale predictions.

## 1 Introduction

The PAULING FILE project is the materials database with nearly 25 years history, grouping crystallographic data, phase diagrams, and physical properties of inorganic crystalline substances under the same frame. Its focus is put on the experimental observations. Each individual crystal structure, phase diagram, or physical property originates from a particular peer-reviewed publication. The world's scientific literature in materials science, engineering, physics, and inorganic chemistry is covered from 1891 to the present date. The Materials Platform for Data Science is an online edition of the PAULING FILE project, created in 2016. It presents all the PAULING FILE data in two online interfaces: graphical user interface and application programming interface. The former is intended for the materials scientists; the latter is intended for the software engineers and data scientists. An intersection of the research interests of these groups falls into the scope of the novel discipline of materials informatics.

## 2 Materials Big Data and the PAULING FILE Project

### 2.1 Modern Challenges Ahead of the Materials Community

Counterintuitively, empirical traditions are widespread in the fundamental science and particularly in materials science. The pioneer in chemoinformatics, Peter Murray-Rust (Cambridge, UK), called the materials science as one of the most conservative precise disciplines, the least transparent and open to collaboration and crowd-sourcing, compared to the other sciences, as biochemistry, astronomy, mathematics, and computer and environmental sciences (Murray-Rust 2013) – not counting paywalls. Interesting to note, the field of materials informatics is less established and younger, compared to chemo- and bioinformatics (Gasteiger and Engel 2003, here *informatics* assumes an *information exchange*). Among the novel distributed computing projects, where anyone may volunteer computing time from their personal computers to a specific cause, only a very little part is concerned with materials science. The following not exhaustive explanation can be given. Materials science and engineering are maximally tightly bound to the industries, e.g., aerospace, automotive, electronics, military industry, etc. Considerable part of

materials research and development is privately funded. A know-how in materials science costs on average higher, than in other precise disciplines. These costs can be compared to high energy physics or astronomy. However an influence of such know-hows on the technological progress is also higher, and the lag between investment and innovation is relatively moderate (about 15 years or less, Obama 2011). Further, due to its widespread and extreme complexity, materials science is also very segmented ("feudal"). Consequently, the level of competition is very high, both in academia and industry (government grants, staff positions, etc.). In terms of complexity, even mastering the basic formalism of the modern solid-state physics presents a nontrivial task. The multidisciplinary specialists are generally rare in materials science. These reasons may give an idea, why the materials science community is inherently very conservative.

Moreover, until the recent years the culture of sharing the basic research outcome (e.g., raw measurement or simulation logs, in contrast to the articles) was totally absent and sometimes even tabooed in materials science. Concerning the publications, still there is no way to obtain scientific information in machine-readable format from journals for further reuse and repurposing. However with the technical progress the modern academics are sinking in the "ocean of data." Today there is a common complaint that researchers are publishing too much and too fast. To estimate this growth, it is wrong to cite the growth of academic databases, as no database captures everything. The bibliometric analysts from the Max Planck Society and Swiss Federal Institute of Technology (Bornmann and Mutz 2015) estimated that the global scientific output roughly doubles every 9 years. Moreover, with the increase in popularity of the modern data-intensive approaches, commonly denoted as "materials design" or "materials genome," there is an evidence that the amount of raw big data in materials science generated by experiments or simulations will continue to grow exponentially. Speaking about the *big data*, the computer science community has appropriately defined this phenomenon as the data amounts governed by four metrics: volume, velocity, variety, and veracity. In materials science *volume* refers to the big sizes, exceeding personal computer facilities, and *velocity* refers to harnessing real-time data acquisition (e.g., from dynamics experiments). *Variety* is concerned with the fact that the data takes all forms in materials science, ranging from discrete numerical values to qualitative descriptions of materials behavior and imaging data. *Veracity* acknowledges the practical reality about uncertainties and a lot of "missing" data (Rajan 2015). Nowadays is the epoch of big data in materials science. Yet in this epoch of big data, materials science is still not doing enough to encourage and enable the sharing, analysis, and interpretation of the vast swatches of data that researchers are collecting. The traditional means of exchange of scientific information in materials science community are deeply imperfect. In such conditions a natural diversity appears, with the principles of the natural selection guiding the shape of the cutting-edge research. And in the last several years with the high penetration of the Internet and modern information technologies, the situation has started to change slowly. Here the ability to digest information, drawing the correct conclusions, is crucial. This is where the data science tools (e.g., machine learning) need to be linked to the

foundations of materials science: theory, modeling, and experiments. The aim is to make the *laboratories* for generating the new information and not just *repositories* for retrieving the known or expected information (Ghiringhelli et al. 2017).

## 2.2 PAULING FILE Background

There are a number of initiatives in the world trying to overcome the above-mentioned challenges in materials science (Pizzi et al. 2016). The main idea is either to systematically collect the materials data or generate and process them in a high-throughput manner. One of the notable initiatives is the PAULING FILE project, launched in 1995. Its main focus is put on the critical evaluation of the published experimental materials data. Historically, it was a joint venture of the Japan Science and Technology Corporation, Material Phases Data System company in Switzerland, and the University of Tokyo, RACE. Now it is managed solely by the Swiss company. Three steps were planned from the beginning. The first goal was to create and maintain a comprehensive database for inorganic crystalline substances, covering crystallographic data, diffraction patterns, intrinsic physical properties, and phase diagrams. The data should be checked with the extreme care. The term "inorganic substances" was defined as compounds containing no C-H bonds. In parallel to the database creation, the second goal was to develop an appropriate retrieval software to make the data accessible in a single-user interface. In longer term, as the third goal, the new tools for materials design should be created, to search the database for correlations automatically. This is known as intelligent design of the new inorganic materials with predefined intrinsic physical properties. The pilot version PAULING FILE Binaries Edition was released as a desktop software in 2002. Now the selected parts of the PAULING FILE data are included in several printed, offline and online products. Today the PAULING FILE project is quite well-known. There are already thousands of publications referring it. Its foundations, database design, and data-centric observations are published (e.g., Villars 2004; Villars et al. 2008; Xu et al. 2011; Kong et al. 2012; Villars and Iwata 2013 etc.). The recent implementation of the PAULING FILE retrieval software and the materials design tools is an online product called Materials Platform for Data Science.

Now a minimum of required definitions must be given. A *database* is a collection of interrelated stored data that serves the needs of the multiple users. In the so-called relational paradigm, these stored data are organized in the *tables*. The motivations for using databases rather than files include greater availability to a diverse set of users, integration of data for easier access to and updating of complex transactions, and less redundancy of data. A *database management system* (DBMS) is a generalized software for manipulating the databases. A DBMS supports a logical view (*database schema*), physical view (access methods, data clustering), data definition and manipulation language, and utilities, such as transaction management and concurrency control, data integrity, crash recovery, and security (Teorey et al. 2005). For example, Oracle is used as an internal DBMS for PAULING FILE. The PAULING FILE database has the following structure. The standard unit of data is

called an *entry*. All the entries are subdivided into three kinds: crystalline structures, physical properties, and phase diagrams. They are called S-, P-, and C-entries, correspondingly. Entries have persistent identifiers, analogous to digital object identifiers (DOIs), e.g., S377634, P600028, and C100027. Another dimension of the PAULING FILE data is the *distinct phases*. The three kinds of entries are interlinked via the distinct materials phases they belong. A tremendous work has been done by PAULING FILE editors in the past 25 years to manually distinguish more than 150,000 inorganic materials phases, appearing in the world scientific literature. Each phase has a unique combination of (*a*) chemical formula and (*b*) modification. These are defined using the structure prototype, Pearson symbol, and space group. Each phase has an integer identifier called *phase id*. In the next three sections, each of the entry types (S-, P-, and C) together with their interlinkage will be covered in details.

## 2.3 Physical Properties

The P-entries of PAULING FILE include the experimental and to a limited extent simulated data for a broad range of intrinsic physical properties of inorganic compounds in the solid, crystalline state. The considered physical properties belong to one of the following seven general domains: (*a*) electronic and electrical properties, (*b*) optical properties, (*c*) magnetic properties, (*d*) mechanical properties, (*e*) phase transitions, (*f*) superconductivity, and (*g*) thermal and thermodynamic properties. The taxonomy consists of three levels: the mentioned general domains, sub-domains, and the particular physical properties. For instance, the domain "electronic and electrical properties" contains the sub-domain "electron energy band structure," which in turn contains the "Fermi energy" property, etc. Currently there are about 100 sub-domains and nearly 2000 particular physical properties. The taxonomy was compiled by Fritz Hulliger (Swiss Federal Institute of Technology in Zurich, Switzerland), Roman Gladyshevskii (Ivan Franko National University of Lviv, Ukraine), and Karin Cenzual (University of Geneva, Switzerland). To a certain degree, it reflects the development of the solid-state physics during the last century.

The physical properties are stored in four different ways: numerical values, figure descriptions, property classes (such as ferromagnet, piezoelectric, etc.), and indications of existence of the particular data in a source publication, e.g., different spectra. The symbols for the most common physical properties have been standardized, mainly based on the *CRC Handbook of Chemistry and Physics* (Lide 1997–1998). The numerical values are stored in the published units and converted to the standard SI units. For certain properties at the atomic level, other units such as eV or $\mu$B are used. Properties expressed with respect to a defined quantity of substance (per kg, per mole) are converted to per atom-gram. Each numerical property value is accompanied by information about the experimental conditions for the particular measurement. A great flexibility is provided via the links to the reference tables. Thanks to that, the new properties may be selected, and their symbols, units, and ranges of magnitude can be controlled.

All the data are taken from the primary literature. Each P-entry corresponds to a particular data source and can contain several numerical values, figure descriptions, and keywords. For an investigation of a compound through a temperature- or pressure-induced structural phase transition, there will be two P-entries, for instance, one for the room-temperature modification and one for the low-temperature modification. By default, ferroelectric transitions are assumed to be accompanied by structural changes and will justify the creation of two P-entries, whereas magnetic, electric, or superconducting transitions are not. Data for the phases with a certain homogeneity range are grouped under a representative chemical formula. The actual composition for a particular measurement, when differing from the composition representing the P-entry, is specified among the parameters. As for the crystal structure part, there will be three database entries for a continuous solid solution between two ternary compounds: one for each ternary boundary compositions and a third one grouping samples containing four chemical elements. Some simulated data from the ab initio calculations are also included, in particular energy band structures, but focus is on experimentally measured data and values directly derived from measurements.

In addition to the physical properties (in the form of numerical values, figure descriptions, or keywords), and compulsory items such as the chemical formula, large amounts of information concerning the sample preparation and experimental conditions are stored. The following database fields may be present in a physical properties P-entry: (*a*) compound, such as chemical system, published chemical formula (investigated samples), representative standardized chemical formula, modification; (*b*) bibliographic data, such as reference, authors, language, title etc.; (*c*) preparation, such as starting materials and method of synthesis; (*d*) sample description, such as form, chemical analysis, stability with respect to temperature, pressure, and composition, elastic behavior, density, color, chemical reactivity; (*e*) crystallographic data, such as structure prototype, space group, and cell parameters.

The PAULING FILE data are checked for consistency using the original software package ESDD (evaluation, standardization, derived data), containing more than 100 different modules (Cenzual et al. 2000). The checking is carried out progressively level by level, also on the individual database fields. These include formatting of numerical values, units and symbols for physical properties, Hermann-Mauguin symbols, Pearson symbols, consistency of journal code, year and volume, pages for literature references, formatting of chemical formulae, usual order of magnitude, spelling, etc. Consistency checks within the individual datasets include atom coordinates, Wyckoff letters, site multiplicity, comparison of chemical elements in chemical system, chemical formula, and comparison of computed and published values. Further quantities for checking are cell volume, density, absorption coefficient, interplanar spacings, Pearson symbol, space group, cell parameters, refined composition, chemical formula, units, symbols for physical properties, Bravais lattice, diffraction conditions, site symmetry, anisotropic displacement parameters, and so on. Special checking of the crystallographic data includes comparison of the interatomic distances with the sum of the atomic radii, comparison of interatomic distances within chemical units, checks on charge

balance, search for missed symmetry elements, and comparison with the type-defining entry (cell parameter ratios, atom coordinates). Consistency checks within the entire database include comparison of densities, comparison of cell parameter ratios for isotypic compounds, check for compulsory data, check of database links, and so forth. Wherever possible, misprints detected in the original publications are corrected. Since editing mistakes can never be completely avoided, all modifications of the originally published data and interpretations of ambiguous data are stored in remarks. The ESDD software further computes the following parameters: at.% of the different elements, molar mass, refined composition and formula, computed density, interplanar spacings from functions of Bragg angle, equivalent isotropic displacement parameters, linear absorption coefficient, Miller indices referring to the published space group setting. It converts compositions expressed in wt.% to at.% and values expressed in various published units to standard units, including units per mole or wt.% to units per gram-atom, respecting the number of significant digits. As seen, an extreme care is taken to provide maximal quality of the stored data.

## 2.4 Crystalline Structures

Currently, PAULING FILE contains more than 350,000 crystalline structure S-entries. The minimal requirement for an S-entry in the PAULING FILE is a complete set of published cell parameters, assigned to a compound of well-defined composition. Whenever the published data are available, the crystallographic data also include atom coordinates, displacement parameters, and experimental diffraction lines and are accompanied by information concerning preparation, experimental conditions, characteristics of the sample, phase transitions, and dependence of the cell parameters on temperature, pressure, and composition. In order to give an approximate idea of the actual structure, a complete set of atom coordinates and site occupancies is proposed for S-entries where a prototype could be assigned. The crystallographic data are stored as published but also have been standardized according to the method proposed by Parthé and Gelato, using the program STRUCTURE TIDY (Gelato and Parthé 1987). When relevant, they are further adjusted so that the data for isotypic S-entries can be directly compared. Derived data include atomic environments of the individual atomic sites, based on the maximum gap method (Brunner and Schwarzenbach 1971; Daams et al. 1992), and the Niggli-reduced cell. The S-entries are checked for inconsistencies within the S-entry and by comparing different S-entries, using the program package ESDD mentioned in the previous section. For 5% of the S-entries, one or more misprints in the published crystallographic data are detected and corrected. Warnings concerning remaining short interatomic distances, deviations from the nominal composition, etc. are added in remarks. SI units are used everywhere, and the crystallographic terms follow the recommendations by the International Union of Crystallography.

Similarly to the physical properties, all the data are extracted from the primary literature. When available, supplementary materials deposited as CIF files or in

the other formats are used as data source. Approximately 10% of the processed documents exist in an original version (e.g., Russian) and a translated version (English); duplicates are avoided, and both references are stored. Crystallographic data, simulated by the ab initio calculations or optimized by the other methods, are only considered being confirmed by experimental observations. Distinct S-entries are created for all the complete refinements reported in a particular paper. For cell parameters without published atom coordinates, an S-entry is prepared for each chemical system and crystal structure. For example, for a continuous solid solution between two ternary compounds, there will be three S-entries: one for each ternary boundary composition and one for the quaternary system. The latter may contain a remark describing the composition dependence of the cell parameters. For the choice of the retrievable cell parameters, preference is given to values determined under ambient conditions.

All the S-entries are subdivided into different categories, according to the level of investigation: complete structure determined, coordinates of non-H atoms determined, cell parameters determined and prototype with fixed coordinates assigned, cell parameters determined and prototype assigned, only cell parameters determined, etc. In addition to the crystallographic data, large amounts of information concerning the sample preparation and experimental investigation are also included in the PAULING FILE. Basic data are stored as published (for rapid comparison with the original paper) and standardized (for efficient data checking and retrieval and for a homogeneous presentation). The following database fields may be present in a crystal structure S-entry: (*a*) classification, such as chemical system, published and standardized chemical formula, modification, colloquial name, structure prototype, Pearson symbol, space group number, Wyckoff sequence, mass per formula unit, computed density, level of structural investigation etc.; (*b*) bibliographic data, such as reference, authors, language, title; (*c*) published and standardized crystallographic data, including detailed information on the atom coordinates, etc. and transformation from published to standardized data; (*d*) Niggli-reduced cell, including transformation from published to Niggli-reduced cell; (*e*) isotropic, anisotropic displacement parameters; (*f*) published diffraction lines, Bragg angle or equivalent parameter, interplanar spacing, intensity, Miller indices, radiation, and remarks; (*g*) preparation, such as starting materials (purity, form), method of synthesis, etc.; (*h*) mineral name, and locality; (*i*) compound description, such as chemical analysis, stability, color, sample form, chemical reactivity, measured density, etc.; (*j*) determination of cell parameters and structure determination, such as sample, experimental method, radiation, and conditions; (*k*) figure descriptions, such as number in the original publication, title, parameters, and ranges; and, finally, (*l*) editor's or general remarks.

As said, each S-entry gets the structure prototype assigned. The structure prototype is a well-known concept in inorganic chemistry, where a large number of compounds often crystallize with very similar atom arrangements. The compilation Strukturbericht (Ewald and Hermann 1931) started already in the beginning for the twentieth century to classify crystal structures into prototypes, named by codes such as *A2*, *B2*, or *G1*. Though these notations are still in use, structure

prototypes are nowadays generally referred to by the name of the compound for which this particular atom arrangement was first identified. The PAULING FILE uses a longer notation, which includes also the Pearson symbol (a lowercase letter for the crystal system, an uppercase letter for the Bravais lattice, sum of multiplicities of all, fully or partially occupied atom sites) and the number of the space group from the *International Tables for Crystallography* (Hahn 1983). All datasets with the published atomic coordinates are classified into the structure prototypes. Each structure prototype is defined based on the first experimentally determined compound with the respective geometrical arrangement of atoms within the unit cell. More than 36,000 different structure prototypes have been identified and assigned to the S-entries. When not published, the editor also assigns the space group setting of the published cell parameters.

There exist an infinite number of ways to select the crystallographic data (cell parameters, space group setting, representative atomic coordinates) that define a crystal structure. The number remains high even when the basic rules recommended by the *International Tables for Crystallography* (Hahn 1983) are respected, due to the allowed symmetry operations such as permutations, origin shifts, etc. It follows that even identical or very similar atom arrangements may not be recognized as such. The classification of crystal structures into structure prototypes is largely facilitated by the use of standardized crystallographic data. The crystallographic data in the PAULING FILE are stored as published but also standardized. This second representation of the same data is such that compounds crystallizing with the same prototype (isotypic compounds) can be directly compared. It is prepared in a three-step procedure. First, the published data are checked for the presence of overlooked symmetry elements and, if found, converted into a space group of higher symmetry. Second, the resulting data are standardized with the program STRUCTURE TIDY. Third, the resulting data are compared with the standardized data of the type-defining database entry (Villars et al. 2018).

The atomic environments, also called coordination polyhedra, are defined for each S-entry using the method of Brunner and Schwarzenbach (Brunner 1971; Daams et al. 1992). One hundred different polyhedral types have been identified in the PAULING FILE data. In most structures, the coordination numbers vary from 1 to 22. It should be noted that this purely geometrical approach was developed for the intermetallic compounds and does not distinguish bonding types. As a consequence, the selected atomic environment may include both cations and anions. However, the method is simple to apply and useful in the majority of cases. Also, this approach offers an additional possibility to check the crystal structure data for geometrical correctness. The atomic environments can be used as the second independent structure classification. For instance, one can easily find geometrically similar prototypes. Notably, the PAULING FILE database supports geometrical restraint criteria for retrieval. That is, one may request information for the crystal structures containing, e.g., tetrahedra and octahedra.

## 2.5 Phase Diagrams and Distinct Phases

The phase diagram section of the PAULING FILE contains temperature-composition phase diagrams for binary systems, as well as the horizontal and vertical sections and liquidus or solidus projections for ternary, quaternary, and other high-order systems. Both experimentally determined and calculated diagrams are processed. Primary literature is considered in the first priority. Also the diagrams from a few well-known compilations, such as the compendium of binary phase diagrams (Massalski et al. 1990) and the series of books on ternary phase diagrams (Petzow and Effenberg 1988–1995), have been included. All the diagrams have been converted to at.% and °C and redrawn in a single scale, so that the different reports for the same chemical systems can easily be compared. Single-phase fields are colored in blue and three-phase fields in yellow. The phases identified on the diagrams are named according to the PAULING FILE conventions, but also the original names are stored. Each phase diagram is linked to a C-entry, which usually contains the following database fields: (*a*) classification, such as chemical system and type of the diagram; (*b*) investigation, such as experimental or calculation technique, APDIC standardization; (*c*) bibliographic data, such as reference, authors, affiliation, language, and title; (*d*) original diagram details, such as figure number in the publication, borders, scales, sizes, etc.; (*e*) redrawn diagram details, such as concentration range, temperature, and conversion of concentration; and (*f*) list of the phases present on the diagram, standardized phase name, name used in the original publication, structure prototype assigned by the editor, structural information given in the original publication, and link to a representative crystal structure S-entry. For binary systems also the temperature and reaction type for the upper and (or) lower limit of existence of the phase are stored.

The physical property, crystal structure, and phase diagram entries are related via the distinct phases concept. At the database level, all three different types of data (P-, S-, and C-entries) are linked to the distinct phases table. To prepare this table, each chemical system has been evaluated. For example, the three major distinct phases reported for $TiO_2$ crystal (rutile, anatase, and brookite) are separated with respect to the temperature or pressure. Then the reported physical properties or crystalline structures of $TiO_2$ are associated only with the corresponding distinct phase. Finally, each C-entry (phase diagram) is formed by a particular set of the known distinct phases. Thus the unique interlinkage of data is achieved. A certain number of characteristics, attributed to the phases (compound classes, mineral names, etc.), are stored in the distinct phases table. Each distinct phase obtains a unique name containing a representative chemical formula, when necessary followed by a specification such as “ht,” “rt,” “cub,” etc. There are the following special cases. *First*, the phases that crystallize in the same structure prototype, but are separated by a two-phase region in phase diagrams, are distinguished. The same is true for the temperature- or pressure-induced isostructural phase transitions where a discontinuity in the cell parameters is reported. *Second*, the structures with different degrees of ordering have in some cases been considered separately, in

others not, depending on the possibility to assign unambiguously one or the other modification. Structure refinements considering, for instance, split atom positions are often grouped under the parent prototype. *Third*, the structure proposals, stated to be incorrect in the later literature, have been grouped under a single phase in agreement with the more recent reports. That is, e.g., an S-entry reporting a hexagonal cell may in such a case be grouped under an orthorhombic phase. *Fourth*, the definition of a structure prototype used here suggests that a continuous solid solution may smoothly shift from one prototype to another. Refinements considering one or the other prototype are then grouped together. *Fifth*, the physical properties reported ignoring the crystal structure, and in principle referring to ambient conditions, are assigned to the "rt" modification, or, if the temperature dependence is not known, to the most commonly observed modification. *Sixth*, by default the paraelectric-ferroelectric phase transitions are assumed to be accompanied by a structural transition, and different phases are considered above and below the transition temperature. On the contrary, magnetic ordering is assumed not to modify the nuclear structure to a significant extent, therefore not leading to the new phases. Still there exist the chemical systems that are little explored, so that the reports in the literature are contradictory. The phase assignment becomes here very difficult, and the list of distinct phases will sometimes contain more phases than there exist in reality. It should be noted that there is a certain amount of subjectivity when assigning a phase identifier. Nevertheless, this approach represents a substantial advantage.

## 2.6 Applications

Thanks to the large amount of information stored in dozens of tables and hundreds of distinct database fields, the PAULING FILE offers almost unlimited possibilities for retrieval. It can of course be used for all kinds of trivial search, based on the chemical system, or literature data, but also much, much more. The conversion to standard units facilitates the search for properties within a particular numerical range, and the assignment of distinct phases plays an essential role, making it possible to combine searches on data stored in the three parts of the database: crystal structures, phase diagrams, and physical properties. The hundreds of interconnected database fields can be used to create different products. The PAULING FILE data are included in various printed products, as well as offline and online software, such as desktop catalogs, simulation environments, materials investigation toolkits, etc. Some of these products contain only structure data, others phase diagrams and crystallographic data, and others the three groups of data. Following the preference of the producers, some products contain only the published cell parameters, others only the standardized cell parameters, and yet others both published and standardized crystallographic data. Some of the products are limited to the PAULING FILE data, whereas others also contain data from other sources.

The Materials Platform for Data Science (MPDS) is a recent online infrastructure, presenting all the three parts of the PAULING FILE data. It contains

nearly 70,000 phase diagrams, over 350,000 crystalline structures, and nearly 700,000 physical property entries. About 300,000 scientific publications in materials science, chemistry, and physics serve as source for these data. About 80% of the data can be requested remotely in a mass manner (via the so-called MPDS application programming interface) in a developer-friendly format, ready for any external modern data-intensive applications. Next sections give an overview of this PAULING FILE implementation, its technical details and usage scenarios, as well as ongoing work in this field in general.

## 3 Materials Genome and Materials Platform for Data Science

### 3.1 Materials Genome Background

The concept of materials genome was taken from bioinformatics, referring to the Human Genome Project, publicly funded initiative started in 1990 and successfully ended in 2003 (Schmutz et al. 2004). In June 2011 the US President announced the multiagency Materials Genome Initiative to help US businesses and universities to discover, develop, and deploy new materials twice as fast, at a fraction of the cost. In 2012, Materials Genome Initiative commitments include $12 million of research at the Department of Energy and $17 million in materials research at the Department of Defense. To the end of 2014 several dozens of universities were participating. The oldest and, probably, the most recognized participant is the Materials Project, an undertaking of the groups of Gerbrand Ceder and Kristin Persson (Lawrence Berkeley National Laboratory). The Materials Project team had identified hundreds of new compounds, several of which now function as lithium battery electrode materials. The software toolkit for materials design, development of the Materials Project, simplifying routine computational tasks, is actively used by about 100 scientists in different organizations around the world, and approximately every second user contributes in the open-source code. Additionally, there exists an open database, prepared using this toolkit (approx. 100,000 compounds). Importantly, the Materials Project team on a half consists of the experimental scientists, who deal with lab synthesis. This leads to an incredibly strong collaboration of theory, modeling, software development, and experiment.

The first European counterpart of the Materials Project was the NoMaD Project, started in 2013 from the collaboration of Fritz Haber Institute (FHI) of the Max Planck Society, Berlin, and Humboldt University of Berlin (HU) with the aim to create an international ab initio materials science data repository. In the end of 2014, the first version of NoMaD user interface was publicly launched, and in 2015 NoMaD Project was successfully funded by European Union's infrastructure call for Centers of Excellence (CoE) in computational sciences. As of 2017, the NoMaD data repository contains more than three million ab initio simulation files (more than 10 Tb disk space on estimate), contributed by the community and taken from the other repositories. In total, more than ten data formats for all the major well-known quantum simulation packages are supported. Based on these data volumes, the

online materials science encyclopedia and the software analytics toolkit are publicly released. A possible disadvantage of NoMaD is the focus on the community's data centralization, which is currently not very well accepted in the materials science community.

Speaking about the publicly funded projects of materials genome, an ongoing commercial activity should also be mentioned. The notable product with the long history (since 1998) is MedeA scientific software environment (Christensen et al. 2017), which presents one of the most advanced and sophisticated simulation desktop toolkits. The MedeA employs computational workflows and allows to manage high-throughput database-driven simulations. Also there are two prominent materials informatics startups in the San Francisco Bay Area, USA: Exabyte and Citrine. The focus of the *Exabyte* software platform is the high-throughput ab initio simulations of the materials, performed in the cloud, i.e., at the commodity hardware cluster infrastructure, rented at one of the public vendors. As a simulation engine, the well-known VASP package is employed. In two recent years the team was able to perform a comparative analysis of about a thousand different materials by utilizing extensive on-demand scalability of the developed cloud platform (Bazhirov et al. 2017). The total costs of the runs ended up not exceeding a few thousand US dollars. The *Citrine*'s online data platform is called Citrination (O'Mara et al. 2016). It was launched in 2015 and now houses over several millions of data points. So far the platform has received various contributions from about 2000 different institutions worldwide, including universities, government laboratories, and companies. The disposed data are completely free and opened. Citrine itself contributes to its platform, searching and disposing the datasets from the opened online sources. The platform provides a free mass access interface for all its data. Citrine also claims to develop the artificial intelligence-based tools that enable new insights from the collected materials data.

It is seen that all the abovementioned initiatives have one main feature in common, namely, they build their own software infrastructures to process information efficiently and tackle the challenge of materials big data. They also develop the novel data-intensive analysis methods. The development of such tooling is mostly conducted within the *open-source* paradigm. This means, the program code of the complex tools is provided for free, and anyone can adopt it for own aims. This has a strong rationale. After a certain complexity threshold, the software product becomes practically unusable, because only a very limited group of professionals are able to deal with it. Often such people are not motivated by money, but they can be attracted by chances and challenges of the possible technological breakthrough. Thus, open-sourcing the parts of the code is great advertising, which allows to attract such talents that could hardly financially be attracted. If the code is popular enough for the outside contributions, a force multiplier is created that helps to get more work done faster and cheaper. More contributors mean more use cases being explored and, finally, the more robust software. Importantly, the user community must be grown around the open-source tooling. Normally such community is fairly amorphous and requires guidance and patronage. The expenses for the development are well covered by the talented contributions from outside, reputation and acknowledgment,

which, in turn, can be converted to the other means of profit. It is also planned to open-source the certain parts of the software of Materials Platform for Data Science. And a part of the PAULING FILE data is now already opened online under the Creative Commons Attribution 4.0 International license. These are (*a*) all entries found by keywords "cell parameters – temperature diagrams" and "cell parameters – pressure diagrams," (*b*) all data for compounds containing both Ag and K, and (*c*) all data for binary compounds of oxygen. In total, (*a–c*) present quite a rich dataset, suitable for educational data-mining purposes.

## 3.2 Development of the PAULING FILE Materials Infrastructure

As of today, materials informatics is a collection of recipes taken from computer science and adopted for the modern materials science. The main difficulty is purely technical from an academic point of view – how to handle materials big data efficiently. Here some recipes are discussed. All they were considered while development and maintenance of the PAULING FILE online retrieval software – MPDS (Materials Platform for Data Science).

One of the possible ways to tackle complexity is the unified modeling language (UML), general-purpose, developmental, modeling language for software engineering (see e.g., Miles and Hamilton 2008). From a bird's eye view, UML is a convention of drawing concepts in a human-understandable manner. Although it cannot be interpreted by a computer, UML provides formal description of the problematic field, which is then much easier to encode in a computer programming language. Another UML advantage is its standardization and high popularity in the field of software engineering (suited for collaborative work). The concepts in UML must be related with the defined relationship types.

The PAULING FILE concepts as implemented in the MPDS platform are represented in Fig. 1, with the very short UML legend at the bottom. Notably, UML provides a clear formal way to understand, how a certain problem domain is organized. Namely, it is seen at Fig. 1 that (*a*) any data-mining tool (e.g., visualization) is based on the MPDS PAULING FILE data; (*b*) data are subdivided into three parts: crystalline structure part (S), physical properties part (P), and phase diagrams part (C); (*c*) each part is represented by entries; and (*d*) each entry is concerned with the relevant phase and scientific publication. Also, there are users with the different data access permissions. Thus, the UML presents the important guideline for the further development.

The MPDS is an online software, working according to a client-server architecture. There are many advantages of the online products over the offline products: absence of installation, cross platform operation, no special requirements to the client PC, transparent updates, enhanced security and reliability, and more. With the ubiquitous penetration of the Internet and the wide availability of the server resources, the online model becomes clearly preferable. The details are presented in Fig. 2.

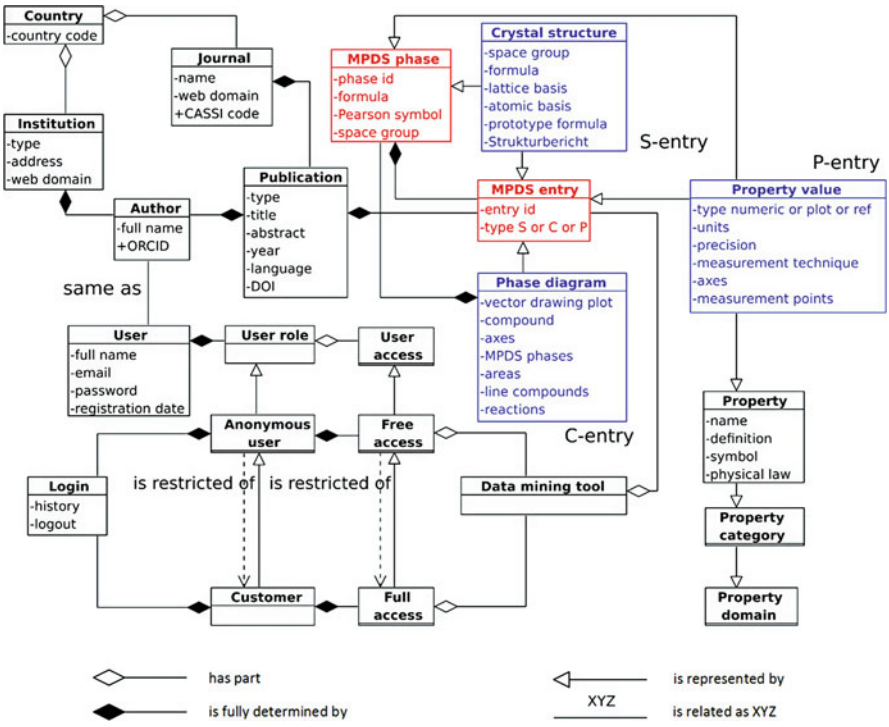


**Fig. 1** UML diagram for PAULING FILE concepts, as implemented in the MPDS

Generally, the World Wide Web is based on the idea of interconnection. Indeed, in a modern environment no isolated software per se makes sense, and each application communicates with the others. In order to provide fast and efficient interaction experience and not to develop a new access interface for each case, the application programming interfaces (APIs) are commonly used. The API regulates communications between any kind of the software (be it a complex of data-mining programs, simulation platform, or any other big data consumer). The main idea is that the functionalities are collected in a single place and exposed (encapsulated) via the online API. The APIs in the online medium normally adhere to the principle of representational state transfer or REST (Fielding 2000). The REST presents guiding constraints for client-server software architecture and could be called as meta-API. It is also employed for the MPDS API (see Fig. 2), which presents all the PAULING FILE data in a developer-friendly, machine-readable way, using the opened formats, such as CIF, JSON, and MIF (see below). Importantly, the API is not only how the clients communicate with the server. In a wider sense, online API is a software architecture, declaring the way of all the communications. This way the audience is not bound to the existing human-oriented graphical user interfaces (irrespective of their convenience) and able to use the service provider maximally efficient for their aims.

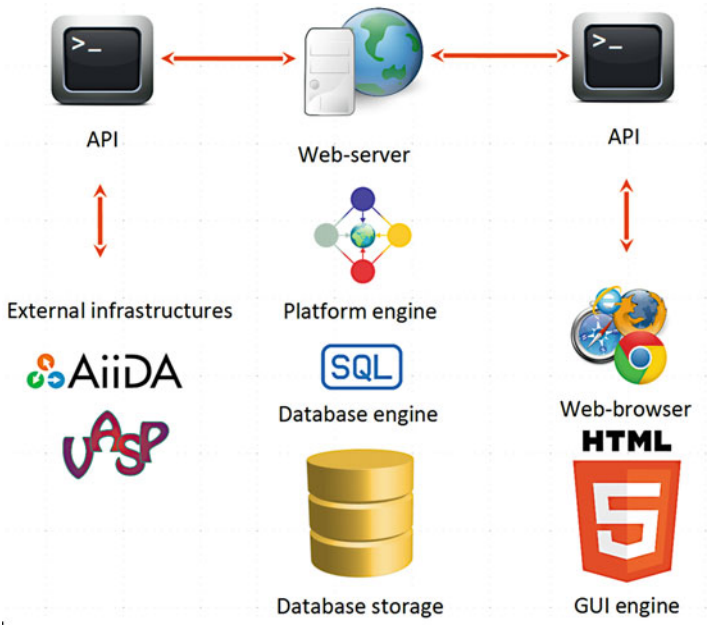


**Fig. 2** Client-server architecture of the MPDS

Nevertheless, much attention at the MPDS is given to its graphical user interface. The search input field allows to type different combinations of basic search terms seamlessly at once, so that they are correctly recognized and the matching entries are shown. The algorithm, responsible for treatment of such seamless searches, works as follows. First, the stop-words ("the," "and," "about," etc.) are filtered. Then, the words are checked to belong to the one-term categories: chemical formulae, chemical elements, and crystalline lattices. These categories cannot contain more than a single word. The found words are interpreted and excluded. Then, the remaining words are checked to belong to the materials classes and physical property categories. The matches of the longest combinations of words are checked in the beginning, and then the shorter combinations are checked. Unrecognized words are marked as "ignored."

From the very beginning of MPDS development performance of all the subsystems was a cornerstone. This is especially important for the centralized online software, when the server is a single point of failure. Generally, for any online project it is always recommended to focus on the operation speed. It should be noted that the software designed for 5 simultaneous users differs considerably from that designed for 50 or 500 simultaneous users. There are two approaches to scaling the performance: vertical and horizontal (Michael et al. 2007). *Vertical scaling* means the more powerful server is deployed with the increased loads, but the software architecture remains the same. This solution is a quick and efficient,

although limited. First, the more powerful, the more expensive is the server, and the dependence is not linear. Second, there is a limit of computational capacity per a single server. And the database performance is normally saturated much lower this limit. So for a database after a certain tolerance vertical scaling makes no sense. The solution in this situation is the *horizontal scaling*, when the software architecture is changed in such a way to distribute the increased loads evenly among the server cluster, consisting of inexpensive commodity hardware. The idea is that the new replicas with the MPDS software are added to the server cluster by simple quick copying. A new replica joins the server cluster and starts accepting the incoming HTTP requests. Furthermore, if a replica fails by whatever reason, it can be automatically replaced by a new one. This way the server cluster stays highly available, and even the huge loads cannot influence the performance, since more cheap replicas could be added. However while the server price remains under control, the development is complex and requires time.

## 3.3 Storage and Exchange of Materials Data

Similar to any other scientific data, materials data are commonly stored in the files and databases. The files usually present raw data (simulation or measurement logs, compressed archives, etc.) Reusable processed data are normally stored in the databases. It should be noted although that such subdivision is very conditional, moreover a file could be a database and vice versa (e.g., HDF5 container, SQLite, etc.). A database model is a type of data model that defines the logical structure of a database and fundamentally determines in which manner data can be stored, organized, and manipulated.

The MPDS employs several database models, in particular, relational and semantic graph model. Additionally, the document-oriented model is used as an intermediate step while preparing data. In the document-oriented model each PAULING FILE entry is treated as a *document*. Such documents may have varying number of properties and are all stored in a single giant database table. All the document properties are searchable. The model is implemented using the PostgreSQL DBMS. Being very simple, this model allows unique flexibility (single table, arbitrary data queries). The cost of such flexibility is very low access speed. Being inappropriate for production, it suits very well for development. Being robust and mature, the relational model is a core of the MPDS production system. It is also supported by the PostgreSQL DBMS, in addition to the intermediate document-oriented model. The standard data manipulation language within the relational model is called structured query language (SQL) and based upon the relational algebra. Eventually, the single table from the document-oriented model is taken, refined, and split into many simpler tables related to each other. This process is called normalization (Teorey et al. 2005). This way much greater access speed is achieved. The refined table structure is however tightly bound to the chosen usage scenario and does not provide extra flexibility.

The disadvantage of the relational model is that the data must obey the strict rules, defined as the database schema. Normally, these rules do not imply an existence of other terms at the database level outside an application business logic. However, the expert systems and artificial intelligence applications must act in an opened world, making inferences and determining new facts, basing on the newly collected information. Their databases must be able to include the new terms and to follow the new logic dynamically and therefore do not fit into the traditional relational paradigm. Usually, such databases employ the so-called semantic Web approach (DuCharme 2013). The data in such model are represented as a *graph*. Such graph is comprised by a set of statements in the form "subject-property-object," called triples. Using the semantic graph database model, the unprecedentedly flexible and expressive queries on top of the knowledge graph become possible. The MPDS currently uses the semantic graphs only indirectly, complementary to the relational model; however much wider adoption of this model is planned. As a back end for the semantic graph model, the Virtuoso DBMS is employed.

As mentioned, the development of the single data exchange format in materials science is an extremely complex task. Nowadays computer science suggests a very convenient paradigm, when a specific information container is accompanied with the rules definition, i.e., an automatic validation tooling. There are various examples following this idea: SQL and schemata for the relational databases, A-Box and T-Box for the semantic graphs, XML and XSD for the machine-readable data transfer and markup, JSON and schemata for the human-readable data transfer, etc. In this respect, two successful achievements in developing the common exchange format in materials science should be mentioned: Crystallographic Information File (Hall et al. 1991) and Materials and Physical Information File (Michel and Meredig 2016). Both of them are supported at the MPDS and partially at the other mentioned materials informatics infrastructures.

The Crystallographic Information File (CIF) was established in the 1990s by the International Union of Crystallography (IUCR). CIF is based on a text container called STAR (Self-Defining Text Archive and Retrieval), where the physical properties, obtained, e.g., as a result of X-ray diffraction or theoretical modeling, are labeled by the standard tags. The standard tags determine the parameters of the unit cell, symmetry, atomic positions, relevant scientific publication metadata, etc. These tags are defined in the external CIF dictionaries (cf. XSD schemata for the XML documents), so it is possible to validate a CIF file against a CIF dictionary and even to infer the new physical properties from those available. The difference is that the CIF format allows the arbitrary tags. They are ignored by CIF parser but later can become the part of standard CIF dictionaries, according to IUCR. Furthermore, CIF format supports the relational data model, so one can refer to the specific atom in the crystalline structure by its identifier. The drawback is the absence of a convenient multilevel hierarchy support, so here the STAR container concedes to XML. CIF format is used for the online crystal structure visualization at the MPDS, and no other proprietary or self-made formats are employed. Only the Web browser is needed for visualization, and no plugins, applets, or other software is required. Normally, structure rendering is done on the GPU (i.e., graphical card).

However, if the GPU is outdated or not available, rendering is done on the CPU (central processor). In this case the quality of rendering is reduced. The total size of the code served online for the rendering is only about 150 Kb, and after the code is loaded, no further Internet connection is required. The CIF visualization at the MPDS is based on the open-source technologies.

The JSON format is simpler, more flexible, and more permissive. Historically, it is much more common for the software development than a narrow-purpose CIF. The container of the Materials and Physical Information File (MIF), introduced by a Citrine startup, is built on top of JSON, taking all its advantages. JSON also provides its own schema approach, used for validation, documentation, and interaction control, i.e., a contract for the JSON data required by a given application, and how that data can be modified. JSON schemata for all types of the PAULING FILE data, including MIF specifications, were developed in 2016 and now are publicly available. The JSON format, including MIF, is the most widely adopted within the mentioned materials infrastructures.

## 3.4 Computer-Assisted Data Analysis

Undoubtedly plotting and visualizations are extremely helpful for data analysis. One of the interesting plotting features of the MPDS is the semantic graphs of terms. Generally, all the MPDS data are a giant semantic graph of the structured knowledge in materials science, accumulated by the humankind. Unfortunately, no human being may observe this graph as a whole. The online interactive visualizations attempt to show only tiny portions of a giant MPDS semantic graph, related to the particular user's input, in a very simplified form. Fortunately, modern semantic technologies (DuCharme 2013) are able to comprehend a giant MPDS graph all at once. This is the planned direction of the MPDS platform development in future. Another type of visualizations, the dynamically rendered phase diagrams (C-entries), displayed online at the MPDS, are fully digitized, programmatically drawn plots. The rendering engine works in all the modern Web browsers, requires no plugins, and is based on the open-source Web technologies. The phases at the phase diagram are associated with the parametric equations in the form $x = x(t)$, $y = y(t)$, $0 <= t <= 1$, where $x$ stands for the composition and $y$ for the temperature.

Recently the MPDS has launched the first version of its machine-learning predictions. To demonstrate some practical usage scenarios of the materials data-mining using MPDS API, a relatively unsophisticated yet powerful predictive machine-learning algorithm, the decision tree regression, was chosen. A decision tree is a statistical model, which describes the data going from the observations about some item (e.g., a crystalline structure) to the conclusions about the item's target value (e.g., a corresponding physical property). The MPDS data contain crystal structures with the corresponding physical properties, so it is feasible to train a model on this dataset. The following physical properties were chosen: isothermal bulk modulus, enthalpy of formation, heat capacity at constant pressure, and melting temperature. Multiple decision trees were built by repeatedly resampling training

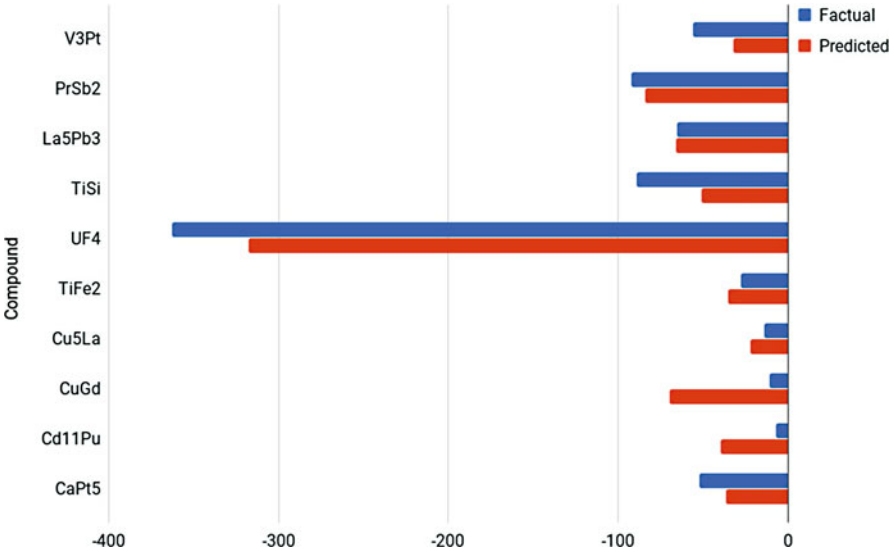


**Fig. 3** Formation enthalpy, kJ g-at.$^{-1}$: *predicted* versus *factual* values; mean absolute error is 42 kJ g-at.$^{-1}$, comparable with the results of the ab initio simulations

data with replacement and voting for the trees yielding a consensus prediction. This algorithm is known as a "*random forest*" regressor. The "random forest" is a statistical estimator that fits a number of classifying decision trees on various subsamples of the dataset and use averaging to improve the predictive accuracy (Breiman 2001). Its presently used state-of-the-art open-source implementation is very efficient and takes seconds to train a model from MPDS data on an average desktop PC (McKinney 2010).

The evaluation process was repeated at least 30 times to achieve a statistical reliability. The results of a randomly chosen evaluation process are shown in Fig. 3. The prediction quality is acceptable and on average may even compete with the ab initio simulation results. The difference is that the simulation normally requires hours or days of computation time, whereas the machine-learning model yields the results in milliseconds on the same hardware. Another difference is that the ab initio simulations in practice require careful fine-tuning of the method, whereas the chosen method of machine learning is a black box, where no initial setup is needed. The disadvantage of the machine-learning model is that no physical meaning of predictions is implied. The underlying complex physical phenomena, as well as the lack of training data, may lead to the poor prediction quality. The size of the training dataset should not be necessarily huge, but there is some minimal threshold. For example, predictions using the smaller dataset of the open MPDS data demonstrate worse quality. Here it is important to note that nowadays more accurate machine-learning techniques exist, such as deep learning neural networks, and the quality of predictions may be further increased.

In the search for the new technological opportunities in the materials data mining, the end point of the route is the artificial intelligence techniques. This is an understandable fact, since the more sophisticated technology, the more similar its intellectual output to the human's one. Historically, there are two approaches to the construction of the intelligent agents (Jones 2008): statistic ("bottom-up" or connectionist) and deterministic ("top-down" or symbolic). The statistic approach is known as "machine learning," i.e., the machine is expected to discover the world on its own, as the humans do. The deterministic approach is known as "inference engine," i.e., the facts are prepared in advance and considered logically. This approach will be shortly reviewed below. The so-called *ontologies* are employed. The language for ontology expression is called OWL (ontology Web language) and based on the logical calculus, used in mathematics, philosophy, linguistics, and computer science (Baader et al. 2007). Applied to materials science, the ontologies are almost unknown, although this term appears in the publications last two decades from time to time. Conceptually, the ontologies and databases look similar. A database consists of the schema and data (i.e., arrangement of tables plus their content). An ontology consists of the axioms and facts (called T-Box and A-Box). There is however an important difference compared to databases. A database schema acts as constraints on structure of data, whereas ontology axioms act as implications or inference rules. The logical expressiveness of the ontology axioms can be much higher (richer) than any database schema. Currently the schema of the MPDS platform database is quite carefully tuned to meet the designed usage scenarios. On the other hand, this is also a potential limitation. Being highly optimized to satisfy online searches, the MPDS platform database does not however perform very well for nonstandard complex queries and hierarchical data manipulations. In this sense, the ontologies fit the MPDS data model very well. Nonetheless, currently the most important drawback of the ontologies compared to the databases is that the performance cannot be unfortunately even theoretically guaranteed, as the OWL logical calculus is extremely complex mathematically. Apparently, all the successful production deployments of the ontologies somehow overpass this limitation (Cuel and Young 2015). Notably, the ontologies per se do not present much practical sense. In fact, they are just logical constructions attached to the data. Only the whole set of accessory technologies and implementations (called the *semantic web stack*) is able potentially to shed more light to the laws of nature, enclosed in the materials data.

## 3.5 Data-Centric Observations

Following Dmitri Mendeleev and Lothar Meyer, who observed periodical patterns in the properties of the chemical elements in the 1860s, one may try to find similar patterns in the whole range of the known materials and make a step toward hypothetical periodic table of materials. Based on the observations across the PAULING FILE database, the 12 empirical principles were formulated (Villars 1994; Villars et al. 2008; Villars and Iwata 2013). These principles can be called the

cornerstones of nature, and all they can be explained using the modern electronic structure theory. They define (*a*) the compound formation, (*b*) the ordering of chemical elements within a structure prototype, and (*c*) the linkage between the position of a chemical element in the periodic table and its occupied sites in the structure prototype.

*First*, compound formation: the atomic size, electrochemical, valence-electron, and cohesion energy factors are governing compound formation. For example, one can observe that about 30% of all chemical element combinations form no compounds within the binary, ternary, and quaternary systems.

*Second,* correlation of the number of chemical elements and atomic environment type: the maximal diversity of the atomic environments is achieved within the binary and ternary inorganic compounds. The quaternary, quinary, and other higher-order compounds strongly prefer the atomic environment types with the low coordination number. One may also call this observation as a surprising reduction of geometrical diversity with the increasing element count.

*Third*, active concentration ranges: there is the systematic occurrence of daltonide inorganic phases for binary, ternary, and quaternary inorganic systems within the certain rather limited (active) concentration ranges.

*Fourth*, stoichiometric ratio condition: there are highly preferred stoichiometric ratios for the vast majority of daltonide compounds.

*Fifth*, compound simplicity: the vast majority of the inorganic compounds have on average only ten atoms per unit cell, thus showing only three or fewer atomic environment types within the crystal structures. One may note here that the nature indeed prefers simplicity.

*Sixth*, compound symmetry: 10% of the space groups cover nearly 70% of the inorganic compounds. The most frequent 11 space groups are 12, 62, 63, 139, 166, 191, 194, 216, 221, 225, and 227. As seen, the high symmetry is preferred.

*Seventh*, atomic environments arrangement: 18 out of about 100 possible atomic environment types are highly preferred and were found for 90% of all the PAULING FILE compounds. In particular, the most frequent types of polyhedra are tetrahedron, octahedron, cube, tri-capped trigonal prism, four-capped trigonal prism, icosahedron, cubooctahedron, bi-capped pentagonal pyramid, and anti-cubooctahedron.

*Eighth*, chemical element ordering tendency: only about 30 structure prototypes have more than 1000 representatives, and the 1000 most populous prototypes and their representatives cover the majority of the crystalline structures of the PAULING FILE.

*Ninth*, correlation of the structure prototype and the periodic system: the vast majority of the crystalline structures show a very strict regularity between the position of the chemical element in the periodic system and its Wyckoff position occupation. This is confirmed on an example of the 1000 most populous prototypes.

*Tenth*, linking of structure and stability: the atomic size, electrochemical, valence-electron, and atomic number factors determine the crystalline structures of the intermetallic compounds. Again, this is applicable to binary, ternary, and quaternary systems. One may reveal clear patterns for, e.g., former versus non-

former systems, iso-stoichiometric structure stability maps, and complete solid solubility between binary compounds within the same prototype.

*Eleventh*, generalized atomic environment type stability: using the periodic number (from Lothar Meyer's periodic table), one may subdivide different atomic environment types into distinct stability domains. It was found that the chemical elements with the periodic number more than 54 control the atomic environment types, independently of whether they act as the central or coordinating atoms. Thus, there exists a clear separation between the possible and impossible atomic environment types. Interestingly, the diversity of atomic environment types is very much reduced for quaternaries, as compared to binaries and ternaries.

*Twelfth*, complete solid solution stability: the atomic size, electrochemical, and valence-electron factors control solid solubility. For example, in a ternary system, where two of the binary boundary systems have the same structure prototype, one may predict whether a complete or limited solid solution is formed. Similarly, for a given chemical element, either a limited or extended solid solubility can be predicted.

As seen, all intrinsic physical properties of a single-phase inorganic solid are strongly linked to its crystal structure, emphasizing the importance of the crystal structure classification. And the 12 principles outlined above can only be discovered by the examination of a large amount of critically evaluated experimentally determined data. They ultimately lead to the restraints, which are a requirement for the development of a practicable and trustworthy computational materials design approach.

## 3.6 Applications

A trustworthy linkage between the published experimental inorganic solids and the high-throughput DFT calculations opens the new perspectives of the database-driven, data-intensive research and discovery of the new materials. At the moment two such high-throughput computational initiatives in China and Switzerland are starting to employ the MPDS API with the PAULING FILE data.

The majority of nowadays' practical materials science problems (clean energy sources, energy storage, superconductivity, etc.) are concerned with the big amounts of the complex knowledge to be assimilated, if not by the gifted humans then by the artificial intelligence agents. With the progress of the neuroscience and medicine, shedding the light on the nature of the human brain, new powerful computer science techniques emerge and later find their application in materials science. However, development of the complex tools in a closed (proprietary) environment is extremely difficult and even inefficient. This is why an academic tradition of openness and free exchange of the ideas realizes in an open-source strategy, showcasing the economically advantageous altruism.

## 4 Summary

The PAULING FILE, a unique and probably the oldest effort on organizing the materials data, was reviewed. Starting from the foundations of materials science, such as the taxonomy of physical properties, concepts of the structure prototypes, phases, phase diagrams, etc., it spans to the modern data-intensive science, employing the novel data storage and analysis techniques and providing high-quality materials insights, as implemented in the MPDS online platform. This is important while the efficient and quick knowledge exchange in materials science is obstructed, ahead of today's big data challenge. The other efforts and their contributions to the materials informatics are also noted. An importance of fostering the emerging ecosystem of today's materials informatics is emphasized. It is based on the standards of JSON, MIF, and CIF, supporting the databases, virtual laboratories, connected with the online REST APIs, which tend to become the software infrastructure backbones. A conceptual UML modeling of the PAULING FILE at the highly abstraction level was showcased. Three types of the PAULING FILE scientific data are highly interlinked, therefore, inseparable within the MPDS online platform. The rich data structure assumes combination of storage approaches, in order to benefit of their advantages and mitigate deficiencies at the same time. Exploring the large amounts of different materials data, a holistic view on inorganic substances may be presented. Coupled with the high-throughput ab initio simulations, it can provide a key to the discovery of materials genome, playing a role of periodic table for entire set of materials. There are certain hopes for the artificial intelligence techniques, automatically generating the new materials discovery ideas, and, citing Linus Pauling, the best way to have a good idea is to have lots of ideas.

**Acknowledgments** The authors acknowledge funding support from NIH Grant U01HL114476.

## References

Baader F, Horrocks I, Sattler U (2007) Description logics, Chapter 3. In: Handbook of knowledge representation. Elsevier, Amsterdam

Bazhirov T, Mohammadi M, Ding K, Barabash S (2017) Large-scale high-throughput computer-aided discovery of advanced materials using cloud computing. Bull Am Phys Soc 62. https://adsabs.harvard.edu/abs/2017APS..MAR.C1007B

Bornmann L, Mutz R (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol 66(11):2215–2222

Breiman (2001) Random forests. Mach Learn 45:5

Brunner GO, Schwarzenbach D (1971) Zur Abgrenzung der Koordinationssphäre und Ermittlung der Koordinationszahl in Kristallstrukturen. Z Kristallogr 133:127–133

Cenzual K, Berndt M, Brandenburg K, Luong V, Flack E, Villars P (2000) ESDD software package, copyright: Japan Science and Technology Corporation, updates by O. Shcherban. Structure-Properties Company, Lviv

Christensen M et al (2017) Software platforms for electronic/atomistic/mesoscopic modeling: status and perspectives. Integr Mater Manuf Innov 6(1):92

Cuel R, Young R (eds) (2015) Formal ontologies meet industry. In: 7th international workshop proceedings, Springer
Daams JLC, van Vucht IHN, Villars P (1992) Atomic-environment classification of the cubic "intermetallic" structure types. J Alloys Compd 182:1–33
DuCharme B (2013) Learning SPARQL, 2nd edn. O'Reilly Media, Sebastopol, CA
Ewald PP, Hermann C (eds) (1931) Strukturbericht. Akad. Verlagsgesellschaft M.B.H, Leipzig
Fielding R (2000) Architectural styles and the design of network-based software architectures. Doctoral dissertation, University of California, Irvine
Gasteiger J, Engel T (2003) Chemoinformatics: a textbook. Wiley, Weinheim
Gelato L, Parthé E (1987) STRUCTURE TIDY- a computer program to standardize crystal structure data. J Appl Crystallogr 20:139–143
Ghiringhelli LM, Vybiral J, Ahmetcik E, Ouyan R, Levchenko SV, Draxl C, Scheffler M (2017) Learning physical descriptors for materials science by compressed sensing. New J Phys 19:023017
Hahn T (ed) (1983) International tables for crystallography, vol A. In: D. Reidel (ed) Dordrecht, Springer
Hall SR, Allen FH, Brown ID (1991) The Crystallographic Information File (CIF): a new standard archive file for crystallography. Acta Crystallogr A47(6):655–685
Jones MT (2008) Artificial intelligence: a systems approach. Jones & Bartlett Learning, Sudbury
Kong C, Villars P, Iwata S, Rajan K (2012) Mapping the materials gene for binary intermetallic compounds–a visualization schema for crystallographic databases. Comput Sci Discov 5:1
Lide D, editor-in-chief (1997–1998), CRC handbook of chemistry and physics, Boca Raton, CRC Press.
Massalski TB, Okamoto H, Subramanian PR, Kacprzak L (eds) (1990) Binary alloy phase diagrams, 2nd edn. ASM International, Materials Park
McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the 9-th python in science conference, p 51
Michael M, Moreira J, Shiloach D, Wisniewski R (2007) Scale-up x scale-out: a case study using Nutch/Lucene. In: 2007 IEEE international parallel and distributed processing symposium, p 1
Michel K, Meredig B (2016) Beyond bulk single crystals: a data format for all materials structure–property–processing relationships. MRS Bull 41(8):617–623
Miles R, Hamilton R (2008) Learning UML 2.0: a pragmatic introduction to UML. O'Reilly Media
Murray-Rust P (2013) Personal communications and online blog. https://blogs.ch.cam.ac.uk/pmr
O'Mara J, Meredig B, Michel K (2016) Materials data infrastructure: a case study of the Citrination platform to examine data import, storage, and access. J Miner, Met Mater Soc 68:2031
Obama B (2011) Materials genome initiative of the US Government. https://obamawhitehouse.archives.gov/mgi
Petzow G, Effenberg G (1988–1995) Ternary alloys: a comprehensive compendium of evaluated constitutional data and phase diagrams, 15 vols. Wiley-VCH, Weinheim
Pizzi G, Cepellotti A, Sabatini R, Marzari N, Kozinsky B (2016) AiiDA: automated interactive infrastructure and database for computational science. Comput Mater Sci 111:218–230
Rajan K (2015) Materials informatics. Mater Today 15:470
Schmutz J, Wheeler J et al (2004) Quality assessment of the human genome sequence. Nature 429:365
Teorey T, Lightstone S, Nadeau T, Jagadish H (2005) Database modeling & design, 4th edn. Elsevier, Amsterdam
Villars P (1994) In: Westbrook JH, Fleischer RL (eds) Intermetallic compounds, principles and practice, vol 1. Wiley, New York, pp 227–275
Villars P, Cenzual K, Daams J, Chen Y, Iwata S (2004) Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB. J Alloys Compd 367(1–2):167–175. https://doi.org/10.1016/j.jallcom.2003.08.060
Villars P, Daams J, Shikata Y, Chen Y, Iwata S (2008) Data-driven generalized atomic environment prediction for binary and multinary inorganic compounds using the periodic number. Chem Met Alloys 1:210–226

Villars P, Iwata S (2013) PAULING FILE verifies/reveals 12 principles in materials science supporting four cornerstones given by nature. Chem Met Alloys 6:81–108
Villars P, Cenzual K, Gladyshevskii R, Iwata S (2018) PAULING FILE – towards a holistic view. In: Materials informatics. Wiley
Xu Y, Yamazaki M, Villars P (2011) Inorganic materials database for exploring the nature of material. Jpn J Appl Phys 50:11S