Chemistry of Metals and Alloys

Chem. Met. Alloys 11 (2018) 43-76
Ivan Franko National University of Lviv
www.chemetal-journal.org

PAULING FILE - towards a holistic view

Pierre VILLARS[1]*, Karin CENZUAL[2], Roman GLADYSHEVSKII[3], Shuichi IWATA[4]

[1] *Material Phases Data System (MPDS), Unterschwanden 6, Vitznau, CH-6354 Luzern, Switzerland*
[2] *Geneva University, Laboratory of Crystallography, quai E. Ansermet 24, CH-1211 Genève, Switzerland*
[3] *Ivan Franko National University of Lviv, Department of Inorganic Chemistry, Kyryla i Mefodiya St. 6, UA-79005 Lviv, Ukraine*
[4] *The Graduate School of Project Design, 3-13-16, Minami-aoyama, Minato-ku, Tokyo, 107-8411, Japan*
* *Corresponding author. E-mail: villars.mpds@bluewin.ch*

Received July 31, 2017; accepted December 26, 2018; available on-line March 29, 2019

The PAULING FILE is a relational database for materials scientists, grouping crystallographic data, phase diagrams, and physical properties of inorganic crystalline substances under the same frame. Focus is on experimental observations and the data are processed from the original publications, covering world literature from 1900 to present date. Each individual crystal structure, phase diagram, or physical properties database entry contains data from a particular publication, but the linkage of the different data sets is achieved *via* the Distinct *Phases* concept, considering the chemical system and the crystal structure, or domain of existence, of different *phases*. At the present stage of development (2016), the PAULING FILE contains over 310'000 crystal structure entries (including atom coordinates and displacement parameters, when relevant) for some 140'000 different *phases*, more than 44'000 phase diagrams (with updated *phase* assignment) for near 9'800 chemical systems, and 120'000 physical properties entries (420'000 numerical values and 130'000 figure descriptions) for some 50'000 *phases*. To reach this result, more than 160'000 scientific publications have been processed. The aim of the PAULING file is to give easy access to large amounts of different kinds of critically analyzed experimental data, and by this propose a general overview on crystalline inorganic substances, offering possibilities to reveal yet undiscovered patterns among data and facilitate a sensible and efficient search for new materials with tailored properties. In combination with different data mining and knowledge discovery techniques, the PAULING FILE provides examples of holistic views on inorganic crystalline substances, confirming that *"the whole is greater than the sum of its parts"*.

Databases / Inorganic materials / Phase diagrams / Crystal structures / Physical properties / Holistic view

1. Introduction

To get a holistic view on materials is the ultimate goal of not only materials scientists, but also materials users. Deep insight into materials has been enhanced through continuous interaction between observation, experiment, data compilation, theoretical modeling, and calculation, as illustrated through history by prominent scientists such as Dmitri Mendeleev and Linus Pauling. Now, when moving into the "Data Era", such interactions can be carried out on a large scale and relatively easily by building up procedures using digitized logics and data. More than 20 years ago, two of us (P.V., I.S.) undertook the task to do so, following the spirit of Linus Pauling, facing many challenges and learning from mistakes in developing materials data systems since the 1960s. This was the beginning of the PAULING FILE project [1] in the 1990s.

There are two general approaches/concepts to obtain a holistic view on materials. The first one is the bottom-up approach (BUA), based on materials data. This data-driven approach is the key guideline of the PAULING FILE project. The second one is the top-down approach (TDA), for which guidelines and logics are taken from outside, from models (mathematics, physics, chemistry, and biology) and surrounding environments (nature and artifacts). TDA can be developed into a combination of powerful sets of scientific disciplines and/or market needs, based on logics, which requires networking of multi-facet knowledge models, bridging gaps and tuning mismatches.

Logics described by explicitly digitized scientific knowledge (*i.e.* not tacit knowledge) have been used in simulation programs with embedded algorithms, rules in Artificial Intelligence (AI) systems and inspiring interfaces. In different fields,

there have been pioneering projects, among which:
• structure-property correlations for organic materials (E.J. Corey, 1960s);
• phase diagrams based on thermodynamics (NIST & ASM, since 1970s);
• structure maps based on crystallography and/or quantum mechanics (D.G. Pettifor, 1980s);
• deformation/fracture mechanism maps based on defect theory (M.F. Ashby, 1970s).

To integrate digitized models has become attractive and popular since the 1980s thanks to the vertiginous development of high-performance computing, consequently named "third science". Thus many projects have been proposed, so-called multi-scale modeling, or multi-physics simulation connecting first principles calculations, Molecular Dynamics (MD), rate equations, the Finite Elements Method (FEM), and other methods that were developed in the 1980s and 1990s.

The data-driven approach to obtain a holistic view requires a digitized system, a prototype of which was developed as a Computer-Aided Design (CAD) system of databases, simulation and AI in the middle of the 1970s. In the middle of the 1980s, almost all ideas, such as meta-data, meta-knowledge, Distributed Data Systems (DDS), intelligent systems of data classification and data mining, knowledge management, learning and inference logics, had been discussed in the dawn of the new networked information environment and the increasing availability of high-performance computers.

Holistic views are obtained through converging interplay between bottom-up and top-down approaches. This was successfully proved "manually" in 1985 [2], and can in a strategic way be transcribed into computational algorisms. The PAULING FILE project proposed the possibility of such a converging interplay *via* a digital platform. Concepts and prototypes for digital systems supporting similar interplay had also been proposed by E.J. Corey in 1969, and further developed in 1975, but the essential difference of the Pauling File project is to discover knowledge directly from data, rather than to reuse a set of predefined knowledge. New knowledge is emerging from data in principle, even if we reuse a set of predefined knowledge chunks for convenience.

Triggered by a workshop held in Como in 1993 [3], important contributions towards a holistic view on materials were conceived in the form of networked intelligent systems of key data and models with powerful PCs and high-performance computers, and several projects to build material data systems, targeting practical needs from materials users in industry, were kicked off in the middle of the 1990s. The Virtual Experiment for Materials Design (VEMD) project [4] and the PAULING FILE project, focusing on TDA and BUA, respectively, were typical projects launched at that moment. However, as shown in the following sections, it takes time to implement a digital system of practical impact. The editorial team of the PAULING FILE have devoted their time for several decades to develop the PAULING FILE to practical competitiveness. Without a comprehensive compilation of highly quality-controlled data, followed by strategic mixing of inductive and deductive inferences, holistic views have not yet "emerged" digitally.

As for TDA, it is still in an incubation period, especially from the viewpoint of materials users. Implications of PAULING FILE experience to get a set of holistic views were not learned properly for VEMD cases, where focus was on offering holistic views to materials users as a piling up of comparative studies. In order to reach the first milestone, the following directions are at present under development by allocating logics for each fact, improving coherence of the collected facts and logics, balancing combined uncertainties through comparative studies of complexities:
• Internet of Things (IoT) (huge monitoring data and data mining, *etc.*) of engineering products (recent General Electric engines as an implementation of J.H. Westbrook's ideas covering from scientific basics to commercial services);
• open and closed approaches to data, necessary to overcome the weakness of BUA and TDA by AI and collective knowledge in the cloud environment;
• market-in and market-out synergetic collaborations between materials producers and materials users.

In order to become innovative and adapt to the market in materials data, as well as to the recent cloud environment, lessons should be learned and/or unlearned from the PAULING FILE and VEMD projects. A few remarks in this sense are summarized at the end of this chapter, as a reference for the future.

1.1 Creation and development of the PAULING FILE

The shortcomings of the empirical (BUA) approach provided the basic motivation for the initiation of the PAULING FILE project, which was launched in 1995 as a joint venture of the Japan Science and Technology Corporation (JST), Material Phases Data System (MPDS), and The University of Tokyo, RACE. The PAULING FILE project [1,5] planned three steps: 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 extreme care, and the term "inorganic substances" was defined as compounds containing no C-H bonds. In parallel to the database creation, appropriate retrieval software should be developed to make the different groups of data mentioned above accessible *via* a single user

interface. In longer term, new tools for materials design should be created, which would more or less automatically search the database for correlations for intelligent design of new inorganic materials with pre-defined intrinsic physical properties. To test the concept, the prototype PAULING FILE – Binaries Edition was published on-line and off-line in 2002 [6,7]. Since about the same time, the PAULING FILE is under the leadership of MPDS alone. Selected PAULING FILE data are included in several printed, off-line and on-line products, most of them updated on a yearly basis, and three multinary on-line editions of the PAULING FILE will be available in 2017 [6,8,9].

2. PAULING FILE - Crystal Structures

The minimal requirement for a crystal structure database entry in the PAULING FILE is a complete set of published cell parameters, assigned to a compound of well-defined composition. Whenever published data are available, the crystallographic data also include atom coordinates, (an)isotropic displacement parameters and experimental diffraction lines, and are accompanied by information concerning preparation, experimental conditions, characteristics of the sample, phase transitions, 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 database entries where a prototype could be assigned (by the authors or by the editors), but atom coordinates were not refined. The crystallographic data are stored as published, but have also been standardized according to the method proposed by Parthé and Gelato [10,11], using the program STRUCTURE TIDY [12], and, when relevant, further adjusted so that the data for isotypic entries can be directly compared [13]. Derived data include Atomic Environments of individual atom sites, based on the maximum gap method [14-16], and the reduced Niggli cell. The database entries are checked for inconsistencies within the database entry (*e.g.* chemical elements, charge balance, interatomic distances, space group, symmetry constraints), and by comparing different database entries (*e.g.* cell parameters and atom coordinates of isotypic compounds) with a program package including more than 30 modules [17]. For 5% of the database 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 [18,19].

2.1 Data selection

The data are extracted from primary literature. Thesis works are not considered and conference abstracts are processed only in exceptional cases. When available, supplementary material deposited as cif files or in other formats is used as data source. Approximately 10% of the processed documents exist in an original and a translated version; duplicates are carefully avoided and both references are stored. Crystallographic data simulated by *ab-initio* calculations or optimized by Differential Pair Distribution Function (d-PDF) or other methods, are only considered when confirmed by experimental observations. Distinct database entries are created for all complete refinements reported in a particular paper. For cell parameters without published atom coordinates, a database entry is prepared for each chemical system and crystal structure (a distinct *phase*, see definition below). For example, for a continuous solid solution between two ternary compounds, there will be three database entries: one for each ternary boundary composition and one for the quaternary system, the latter possibly containing a remark describing the composition dependence of the cell parameters. For the choice of retrievable cell parameters, preference is given to values determined under ambient conditions.

2.2 Categories of crystal structure entries

As stated above, the minimal requirement for a database entry in the crystal structure part of the PAULING FILE is a complete set of published cell parameters. The database entries are subdivided into different categories, according to the level of investigation, of which the most common are:

- complete structure determined;
- coordinates of non-H atoms determined;
- cell parameters determined and prototype with fixed coordinates assigned;
- cell parameters determined and prototype assigned;
- cell parameters determined.

Atom coordinates are included in the PAULING FILE for the first four categories. Less frequent categories are: average structure, commensurate approximant, part of atom coordinates determined, cell parameters determined and parent structure assigned (for filled-up derivatives such as carbides, hydrides), subcell determined.

The brief summary defining the level of investigation may be followed by information about additional studies, such as:

- absolute structure determined;
- composition dependence studied;
- electron density studied;
- magnetic structure studied;
- pressure dependence studied;
- refinement in superspace;
- temperature dependence studied.

2.3 Database fields

In addition to the crystallographic data, large amounts of information concerning the sample preparation and experimental investigation are 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 database entry:

• *Classification:* chemical system; chemical formula (as published, standardized); modification; colloquial name; structure prototype; Pearson symbol; space group number; Wyckoff sequence; mass per formula unit; computed density; level of structural investigation; additional studies

• *Bibliographic data:* data source; authors (affiliation); language; title

• *Published crystallographic data:* space group; cell parameters; number of formula units per cell; atom coordinates (site label; element(s); site multiplicity, Wyckoff letter; site symmetry; *x*; *y*; *z*; partial site occupancy)

• *Standardized crystallographic data:* space group; cell parameters; number of formula units per cell; atom coordinates (site label; element(s); site multiplicity, Wyckoff letter; site symmetry; *x*; *y*; *z*; partial site occupancy); transformation from published to standardized data

• *Niggli-reduced cell:* cell parameters; transformation from published to Niggli-reduced cell

• *Displacement parameters:* isotropic; anisotropic; computed equivalent isotropic

• *Published diffraction lines:* Bragg angle or equivalent parameter; interplanar spacing; intensity; Miller indices; radiation; remarks

• *Preparation:* starting materials (purity, form); method of synthesis (crucible, atmosphere, solvent); annealing or crystal growth

• *Mineral:* mineral name; locality

• *Compound description:* chemical analysis (method, composition from analysis); stability with respect to temperature, pressure, composition; color; optical characteristics; sample form (crystal habit, grain size); chemical reactivity; measured density

• *Determination of cell parameters:* sample; experimental method; radiation; temperature; pressure, theta range, software used

• *Structure determination:* sample; experimental method; diffractometer/reactor; radiation; temperature; pressure; scan mode; theta range; number of reflections; linear absorption coefficient, absorption correction; starting model; refinement; number of refined parameters; numbers of reflections; condition for observed reflections; R factors; software used

• *Remarks:* general remarks; errata; editor remarks (modifications of published data, warnings); remarks on/from related references; dependence of cell parameters on temperature, pressure, composition

• *Figure descriptions:* figure number in the original publication; title; parameters; ranges

The data extracted and stored for a ternary aluminide are shown in Table 1.

2.4 Structure prototypes

The *structure type* is a well-known concept in inorganic chemistry, where a large number of compounds often crystallize with very similar atom arrangements. The compilation Strukturbericht [20] started already in the beginning for the 20th century to classify crystal structures into types, named by codes such as A1, B1 or A15. Though these notations are still in use, structure types are nowadays generally referred to by the name of the compound for which this particular atom arrangement was first identified, *i.e.* for the types enumerated above: Cu, NaCl, Cr_3Si. The PAULING FILE uses a longer notation, which includes also the Pearson symbol (a lower-case letter for the crystal system, an upper-case letter for the Bravais lattice, sum of multiplicities of all, fully or partially occupied atom sites) [21] and the number of the space group in the International Tables for Crystallography [18]: Cu,cF4,225; NaCl,cF8,225; Cr_3Si,cP8,223.

All data sets with published atom coordinates are in the PAULING FILE classified into structure prototypes, following the criteria defined in TYPIX [22]. According to this definition, isotypic compounds must crystallize in the same space group, have similar cell parameter ratios, and the same Wyckoff positions should be occupied in the standardized description (see below), with the same or similar values of the atom coordinates. If all these criteria are fulfilled, the atomic environments should be similar. Different ordering variants (substitution derivatives) are distinguished but, in the general case, no distinction is made between structures with fully and partly occupied atom sites. Because of the difficulty to locate protonic hydrogen atoms by X-ray diffraction, the positions of H atoms in structures containing more than two chemical elements (with the exception of hydrides) are ignored in the classification.

Each structure prototype is defined on a database entry in the crystal structure part of the PAULING FILE. These database entries are grouped in the so-called Structure Type Pool (STP), and may later be replaced. More than 36’000 different prototypes have up to date been identified and added to the Structure Type Pool.

When possible, a structure type has been assigned also to data sets without atom coordinates. The structure type is often stated in the original publication, in other cases it is assigned directly by the editors. The assigned prototype may in some cases be an approximation of the real structure, ignoring for instance a certain disorder. When not published, the editor assigns also the space group setting to which the published cell parameters refer.

Table 1 Example of data stored for a PAULING FILE crystal structure entry.

Summary
Standardized formula $YNiAl_4$; Alphabetic formula Al_4NiY; Published formula $YNiAl_4$; Formula from refinement Al_4NiY
Structure prototype $YNiAl_4$,oS24,63; Space group *Cmcm* (63); Wyckoff sequence 63,fc^{3}a
Computed density 4.07 Mg m^{-3} Molar mass 255.5
Level of investigation complete structure determined

Bibliographic data
Reference Sov. Phys. Crystallogr. (1972) 17, 453-455; Kristallografiya (1972) 17, 521-524;
Language Russian/English; Title Crystal structure of the compounds $YNiAl_4$ and $YNiAl_2$

Author	***Department***	**Organization**	**City**	**Country**
Rykhal' R.M.	Department of Inorganic Chemistry	Lviv Ivan Franko National University	Lviv	Ukraine
Zarechnyuk O.S.	Department of Inorganic Chemistry	Lviv Ivan Franko National University	Lviv	Ukraine
Yarmolyuk Y.P.	Department of Inorganic Chemistry	Lviv Ivan Franko National University	Lviv	Ukraine

Published crystallographic data
Space group *Cmcm* (63)
Cell parameters *a* = 0.408, *b* = 1.544, *c* = 0.662 nm, *α* = 90, *β* = 90, *γ* = 90°,
V = 0.417 nm^3, *a/b* = 0.264, *b/c* = 2.332, *c/a* = 1.623, Z = 4
Atom coordinates

Site	**Elements**	**Wyckoff position**	**Site symmetry**	***x***	***y***	***z***	**Partial occupancy**
Y	Y	4*c*	*m2m*	0	0.121	1/4	
Ni	Ni	4*c*	*m2m*	0	0.771	1/4	
Al1	Al	8*f*	*m..*	0	0.314	0.054	
Al2	Al	4*c*	*m2m*	0	0.943	1/4	
Al3	Al	4*b*	2/*m..*	0	1/2	0	

Standardized crystallographic data
Space group *Cmcm* (63)
Cell parameters *a* = 0.408, *b* = 1.544, *c* = 0.662 nm, *α* = 90, *β* = 90, *γ* = 90°,
V = 0.4170 nm^3, *a/b* = 0.264, *b/c* = 2.332, *c/a* = 1.623, Z = 4
Atom coordinates

Site	**Elements**	**Wyckoff position**	**Site symmetry**	***x***	***y***	***z***	**Partial occupancy**
Al1	Al	8*f*	*m..*	0	0.186	0.054	
Y	Y	4*c*	*m2m*	0	0.379	1/4	
Al2	Al	4*c*	*m2m*	0	0.557	1/4	
Ni	Ni	4*c*	*m2m*	0	0.729	1/4	
Al3	Al	4*a*	2/*m..*	0	0	0	

Transformation origin shift 0 1/2 1/2

Niggli-reduced cell
a = 0.408, *b* = 0.662, *c* = 0.7985 nm, *α* = 90, *β* = 104.802, *γ* = 90°,
V = 0.2085 nm^3, *a/b* = 0.616, *b/c* = 0.829, *c/a* = 1.957

Atomic Environments

Site	**Coordination number**	**Atomic Environment Type**	**Composition**
Al1	12	cuboctahedron	$Ni_3Al_6Y_3$
Y	19	distorted pseudo Frank-Kasper (19)	$Al_{13}Ni_4Y_2$
Al2	12	cuboctahedron	NiY_3Al_8
Ni	9	tricapped trigonal prism	Al_7Y_2
Al3	12	cuboctahedron	Al_8Y_4

Preparation

Starting material	**Purity**	**Form**
Y	99.9 wt.%	
Ni	electrolytic, 99.98 wt.%	
Al	99.98 wt.%	

Synthesis arc-melted; Atmosphere purified argon; Composition of sample $Al_{70}Ni_{15}Y_{15}$

Description of the sample
Measured density 4.27 Mg m^{-3}

Determination of the cell parameters
Sample single crystal; Method rotation photographs; Radiation X-rays, Cu Kα

Structure determination
Sample single crystal; Method Weissenberg photographs; Radiation X-rays, Cu Kα; Data collection 0 k l;
Model crystal chemical considerations; Refinement least-squares refinement, 69 reflections; R factors R = 0.150

Processing information
Document 102868; S-entry 1407077; Processing 12-MAY-03; Checking 23-APR-04; Last update 12-MAY-03

2.5 Standardized crystallographic data

There exist an infinite number of ways to select the crystallographic data (cell parameters, space group setting, representative atom coordinate triplets) that define a crystal structure. The number remains high even when the basic rules recommended by the International Tables for Crystallography [18] are respected, due to space-group allowed operations such as permutations, origin shifts, *etc*. It follows that even identical or very similar atom arrangements may not be recognized as such (see Fig. 1). The classification of crystal structures into structure prototypes is largely facilitated by the use of standardized crystallographic data (several examples are given in [23]).

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 3-step procedure:

(1) The published data are checked for the presence of overlooked symmetry elements [24] and, if relevant, converted into a space group of higher symmetry.

(2) The resulting data are standardized with the program STRUCTURE TIDY [12].

(3) The resulting data are compared with the standardized data of the type-defining database entry, and, if relevant, additional space-group permitted operations are performed [13].

(1) Checking of symmetry

A crystal structure can always be refined and described in a subgroup of the actual space group. To an extreme, any structure can be described in the triclinic space group *P*1, having no other symmetry elements than identity and translation. To know the correct space group is important not only for the recognition of isotypic structures, but also in connection with intrinsic physical properties. Particular properties are effectively restrained to certain symmetries, *e.g.* ferroelectricity can only be observed for polar space groups, whereas pyroelectricity is excluded for crystal structures possessing an inversion center. Therefore, the crystallographic data in the PAULING FILE are checked for the presence of overlooked symmetry elements [24]. Whenever it is possible to describe the structure in a space group of higher symmetry, or with a smaller unit cell, without any approximations, this is done. Fig. 2 shows how the structure of WAl_5, reported in space group $P6_3$ (173), can be described in space group $P6_322$ (182), after applying an origin shift of 0 0 3/4 to the published data [25].

(2) Standardization

At the next step, the crystallographic data are standardized following the method proposed by Parthé and Gelato [10,11], using the program STRUCTURE TIDY [12]. The standardization procedure applies criteria to select the space group setting, the cell parameters, the origin of the coordinate system, the representative atom coordinates, and the order of the atom sites. The main criteria are summarized below.

The coordinate system must be right-handed and refer to a space group setting defined in the International Tables for Crystallography [18], with the following additional constraints:

RbO

Published crystal structure

Spacegroup	Immm (71)
Cell parameters	a = 0.4201(5) nm, b = 0.7075(5) nm, c = 0.5983(5) nm
Cell length ratio(s)	a/b = 0.594, b/c = 1.183, c/a = 1.424
Cell volume [nm³]	0.17783

Atom coordinates

Label	Site identifier	Wyckoff notation	Point set symmetry	x	y	z
Rb	Rb	4g	m2m	0	0.25	0
O	O	4i	mm2	0	0	0.374

Standardized crystal structure

Crystal system	orthorhombic
Spacegroup	Immm (71)
Wyckoff sequence	71,ig
Pearson symbol	oI8
Cell parameters	a = 0.4201 nm, b = 0.5983 nm, c = 0.7075 nm
Cell length ratio(s)	a/b = 0.702, b/c = 0.846, c/a = 1.684
Cell volume [nm³]	0.17783
Number of formula units	4

Atom coordinates

Label	Site identifier	Wyckoff notation	Point set symmetry	x	y	z
Rb	Rb	4i	mm2	0	0	0.25
O	O	4g	m2m	0	0.374	0

Transformation	new axes a,-c,b

CsS

Published crystal structure

Spacegroup	Immm (71)
Structure type	Rb_2O_2
Cell parameters	a = 0.6992(2) nm, b = 0.9615(2) nm, c = 0.5232(2) nm
Cell length ratio(s)	a/b = 0.727, b/c = 1.838, c/a = 0.748
Cell volume [nm³]	0.351737

Atom coordinates

Label	Site identifier	Wyckoff notation	Point set symmetry	x	y	z
Cs	Cs	4h	m2m	0	0.2187(2)	1/2
S	S	4e	2mm	0.1505(11)	0	0

Standardized crystal structure

Crystal system	orthorhombic
Spacegroup	Immm (71)
Wyckoff sequence	71,ig
Pearson symbol	oI8
Cell parameters	a = 0.5232 nm, b = 0.6992 nm, c = 0.9615 nm
Cell length ratio(s)	a/b = 0.748, b/c = 0.727, c/a = 1.838
Cell volume [nm³]	0.351737
Number of formula units	4

Atom coordinates

Label	Site identifier	Wyckoff notation	Point set symmetry	x	y	z
Cs	Cs	4i	mm2	0	0	0.2813
S	S	4g	m2m	0	0.3495	0

Transformation	new axes c,a,b; origin shift 0 1/2 0

Fig. 1 Data sets for RbO and CsS, as published and after standardization, revealing their isotypism. Data as shown in the PAULING FILE – Binaries Edition [7].

Published crystal structure

Spacegroup	$P6_3$ (173)
Cell parameters	a = 0.49020(3) nm, c = 0.88570(5) nm
Cell length ratio(s)	c/a = 1.807
Cell volume [nm³]	0.18432

Atom coordinates

Label	Site identifier	Wyckoff notation	Point set symmetry	x	y	z
W	W	2b	3..	1/3	2/3	0.5
Al1	Al	2b	3..	1/3	2/3	0.0
Al2	Al	2a	3..	0	0	0.0
Al3	Al	6c	1	1/3	1/3	0.25

Standardized crystal structure

Crystal system	hexagonal
Spacegroup	$P6_322$ (182)
Wyckoff sequence	182,gdcb
Pearson symbol	hP12
Cell parameters	a = 0.49020 nm, c = 0.88570 nm
Cell length ratio(s)	c/a = 1.807
Cell volume [nm³]	0.18432
Number of formula units	2

Atom coordinates

Label	Site identifier	Wyckoff notation	Point set symmetry	x	y	z
Al3	Al	6g	.2.	0.33333	0	0
W	W	2d	3.2	1/3	2/3	3/4
Al1	Al	2c	3.2	1/3	2/3	1/4
Al2	Al	2b	3.2	0	0	1/4

Transformation	origin shift 0 0 3/4

Fig. 2 The structure of WAl_5, reported in space group $P6_3$ (173), can be described in space group $P6_322$ (182), after applying an origin shift of 0 0 3/4 to the published data. Data set from the PAULING FILE – Binaries Edition [7].

• triclinic space groups: Niggli-reduced cell;
• monoclinic space groups: b-axis unique, "best" cell;
• orthorhombic space groups: $a \leq b \leq c$, when not fixed by the space group setting;
• trigonal space groups with *R*-lattice: triple hexagonal cell;
• space groups with two origin choices: origin choice 2 (origin at inversion center);
• enantiomorphic space groups: smallest index of the relevant screw axis.

For the 148 non-polar space groups there exist between 1 and 24 possibilities to rotate, invert or shift the coordinate system, respecting the conditions listed above. For each possibility the standardization program prepares a complete description where the representative triplet of each atom site must obey a series of eliminative conditions:
• first triplet in the International Tables for Crystallography [18];
• $0 \leq x, y, z < 1$;
• minimum value of $(x^2+y^2+z^2)$;
• minimum value of x, then y, then z.

For polar space groups similar data sets are prepared where one atom site after the other, belonging to the "lowest Wyckoff set" (set of Wyckoff sites containing the first letters in the alphabet) represented in the structure, fixes the origin on the polar axis. One of the data sets, prepared as described above, is selected based on the following eliminative conditions:
• minimum value of $\sum(x_j^2+y_j^2+z_j^2)^{1/2}$ summing over all atom sites;
• minimum value of $\sum x_j$ summing over all atom sites, then $\sum y_j$, then $\sum z_j$;
• minimum value of $x_n^2+y_n^2+z_n^2$ for the n^{th} atom site.
Finally, the atom sites are reordered according to the following eliminative criteria:
• inverse alphabetic order of Wyckoff letters;
• increasing x, then y, then z.

In order to obtain similar standardized data sets for refinements with and without hydrogen positions, the positions of H (D, T) atoms in structures containing more than two chemical elements (with the exception of hydrides), are not taken into consideration for the choice of the standardized data set. The coordinates of the hydrogen atoms, when determined, are transformed according to the same operations as the remaining coordinates, and the atom sites are listed at the end of the standardized data set. Protonic hydrogen atoms are also ignored in parameters used for structural classification, such as the Pearson symbol or the Wyckoff sequence.

(3) Comparison with the type-defining data set

In the general case the standardization procedure produces directly comparable data sets for isotypic compounds. This is, however, not always true since particular situations may occur, *e.g.*:
• Two refinable cell parameters have similar values. Which one is the larger one may differ for isotypic

compounds and the constraint $a \leq b \leq c$ will lead to different standardized descriptions.

• The condition imposing that all the angles of the Niggli-reduced cell must be either ≤ 90° or ≥ 90° may cause flipping of triclinic unit cells, when the value of one of the angles switches from slightly larger than 90° to slightly smaller than 90°.

• The constraint that refinable atom coordinates must be ≥ 0 is responsible for a certain number of diverging standardizations observed for isotypic structures with refinable atom coordinates close to 0.

• The order of the atom sites may differ for isotypic compounds where several atom sites in the same Wyckoff position with similar refinable *x*-coordinates (*y*-, *z*-) are present.

To remedy these problems, each standardized data set is compared with the standardized database entry that defines the prototype in the PAULING FILE. The program COMPARE [13] generates the different space-group permitted crystallographic representations. Each representation is compared with the standardized description of the type-defining entry, based on the value of the sum of the "minimum distances" between corresponding atom sites, expressed in "fractional coordinates", multiplied by the site multiplicity: B(est) S(etting) C(riterion) = $\sum m_i [(\Delta x_i)^2+(\Delta y_i)^2+(\Delta z_i)^2]^{1/2}$, summing over all atom sites. The standardized data set is replaced by the data set having the smallest BSC value, and the isotypism is then checked by detecting atom coordinates differing by more than 0.1 from those of the type-defining entry.

For data sets with no published coordinates, the cell parameters are standardized following the criteria defined for the unit cell and space group setting. For data sets with unknown space group, the cell parameters are standardized assuming the space group of lowest symmetry in agreement with the Pearson symbol, *e.g. P*222 for no more information than orthorhombic (o**) or orthorhombic primitive (oP*). For triclinic structures, the cell is adjusted by comparing with the cell of the type-defining database entry.

2.6 Assigned atom coordinates

In order to give an approximate idea of the actual structure, a complete set of atom coordinates and site occupancies is proposed for database entries where a structure prototype could be assigned (by the authors and/or by the editors), but atom coordinates were not determined. Two different cases occur:

(1) A structure type where all atom coordinates are fixed by symmetry is assigned.

The editor, based on the chemical formula, will in this case assign also a probable atom distribution. For off-stoichiometric compositions, different situations are proposed as a first approximation, depending on the structure type. In the general case, fully occupied atom sites with mixed occupation are assumed, whereas for structure types such as NaCl, ZnS, CaF_2, NiAs, or Ni_2In, vacancies are assumed on one atom site.

(2) A structure type with refinable atom coordinates is assigned.

The atom coordinates of the type-defining entry are proposed as a first approximation. The atom distribution is inserted by a program that compares the chemical formula of the type-defining entry with a chemical formula modified by the editor so that the substitution element by element is emphasized [17]. The structure types having both entries with and without refined atom coordinates in the PAULING FILE have been analyzed and information concerning their behavior with respect to off-stoichiometry is stored if vacancies or mixed occupation are expected to occur selectively on particular atom sites. The positions of protonic H positions are not included among the assigned coordinates, but sites occupied by *e.g.* O, OH or OH_2 are distinguished.

No attempt has been made to propose a data set closer to the real structure, *e.g.* by copying refinements for the same compound, since assigned atom coordinates and site occupancies can anyhow not replace a structure refinement.

2.7 Atomic Environment Types (AETs)

For the approach used here [15,16], the Atomic Environment (AE), also called coordination polyhedron, is defined using the method of Brunner and Schwarzenbach [14]. According to this method the interatomic distances between an atom and its neighbors are plotted in a next-neighbor histogram, as shown on the left hand-side of Fig. 3 for the Ti atom in $BaTiO_3$ rt. In most cases a clear maximum gap is revealed and the atoms situated at distances to the left of the maximum gap are considered to belong to the AE of the central atom. This rule is called the "maximum gap rule" and the coordination polyhedron, the Atomic Environment Type (AET), is constructed with the atoms to the left of the maximum gap. The polyhedron around the Ti atom in Fig. 3 is a (distorted) octahedron.

In those cases where the maximum gap rule leads to AETs with not only the selected central atoms but also additional atoms enclosed in the polyhedron, or to AETs with atoms located on one or more of the faces or edges of the coordination polyhedron, the so-called "maximum-convex-volume rule" is applied. This rule is defined as the maximum volume around the central atom delimited by convex faces, with all the atoms of the AE lying at the intersection of at least three faces. This rule is also used in those cases where no clear maximum gap is detected.

All the structure entries with refined or fixed atom coordinates in the PAULING FILE are analyzed, applying the rules given above. 100 different AETs have been identified, of which the 50 most frequent ones are listed in Table 2. Each Atomic Environment Type is identified by a code and the name of the

coordination polyhedron; the count in the second column gives the number of times this AET is present in Pearson's Crystal Data [26], release 2016/17. In most structures the coordination numbers (CN) vary from CN = 1 to CN = 22.

It may be noted that this purely geometrical approach, which was developed for intermetallic compounds, does not distinguish types of bonding. As a consequence, the selected Atomic Environment may include both cations and anions, both atoms forming covalent bonds with the central atom and counter-ions, or large atoms at contact distances and small atoms with little interaction. The procedure further considers all atom sites as being fully occupied and, consequently, a tetrahedron (*e.g.* a sulfate ion) in statistical disorder between two orientations will be classified as a cube. However, the method is simple to apply and of great use in the majority of the cases.

Table 2 The 50 most frequent Atomic Environment Types (AETs) with their counts (number of point sets) in PCD 2016/17.

	Count	AET-code	Coordination polyhedron
1	295'885	1#a	single atom
2	234'712	2#a	non-collinear
3	177'943	6-a	octahedron
4	168'982	4-a	tetrahedron
5	107'137	3#a	non-coplanar triangle
6	40'263	12-b	cuboctahedron
7	29'728	9-a	tricapped trigonal prism
8	26'893	2#b	collinear
9	24'678	12-a	icosahedron
10	17'863	3#b	coplanar triangle
11	16'693	14-b	rhombic dodecahedron
12	16'536	8-a	square prism (cube)
13	16'075	8-b	square antiprism
14	15'285	5-a	square pyramid
15	14'824	5-c	trigonal bipyramid
16	10'840	7-g	monocapped trigonal prism
17	10'151	14-a	14-vertex Frank-Kasper
18	9'424	10-a	fourcapped trigonal prism
19	8'921	6-b	trigonal prism
20	8'806	16-a	16-vertex Frank-Kasper
21	8'504	4#c	coplanar square
22	7'730	7-h	pentagonal bipyramid
23	7'499	20-a	pseudo Frank-Kasper (20)
24	7'155	13-a	pseudo Frank-Kasper (13)
25	6'484	4#d	non-coplanar square
26	6'067	12-d	anticuboctahedron
27	5'074	8-d	distorted square antiprism-a
28	4'405	8-g	double anti-trigonal prism
29	4'329	4#b	tetrahedron, central atom outside
30	4'160	15-a	15-vertex Frank-Kasper
31	4'088	10-b	bicapped square prism
32	4'021	17-d	7-capped pentagonal prism
33	3'989	11-a	pentacapped trigonal prism
34	3'860	6-d	pentagonal pyramid
35	3'360	8-i	side-bicapped trigonal prism
36	3'339	11-b	pseudo Frank-Kasper (11)
37	3'203	8-c	hexagonal bipyramid
38	3'042	10-c	bicapped square antiprism
39	2'941	18-a	eight-equatorial-capped pentagonal prism
40	2'816	22-a	polarity, eight-equatorial-capped hexagonal prism
41	2'363	10-e	distorted equatorial fourcapped trigonal prism
42	2'023	5#d	square pyramid, central atom outside
43	1'998	8-j	distorted square antiprism-b
44	1'987	12-f	hexagonal prism
45	1'343	14-d	bicapped hexagonal prism
46	1'243	20-h	twelve-pentagonal-faced polyhedron
47	1'117	7-a	monocapped octahedron
48	961	18-d	sixcapped hexagonal prism
49	945	6-h	distorted trigonal prism
50	908	12-c	bicapped pentagonal prism

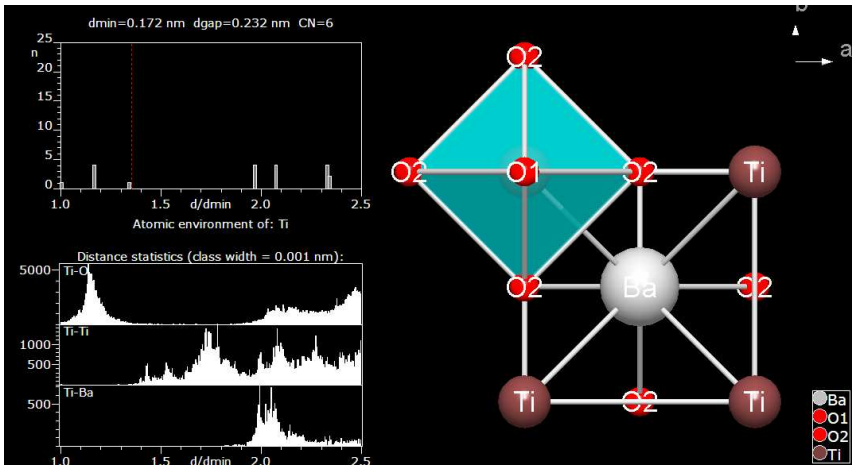


Fig. 3 Next-neighbor histogram (NNH) (top left) and the corresponding coordination polyhedron (AET) for an entry for $BaTiO_3$ rt in Pearson's Crystal Data [26].

The Atomic-Environment approach offers an additional possibility to check the crystal structure data for geometrical correctness. Coordination polyhedra also constitute a tool for the classification of crystal structures into geometrically similar types [27], with the definition used for AET used here called "coordination types" [16].

2.8 Cell parameters from plots

Since 2009, values are extracted from plots of cell parameters (or functions of these) *vs.* temperature, pressure, or composition [28] and stored in the database. Three cases are distinguished: experimental points, fit to experimental points, linear dependence. The extracted values are converted to SI units and used to produce new figures, illustrating thermal expansion, phase transitions, or compression under pressure (see examples in Fig. 4). Values extracted from the same publication for the same *phase* and temperature/pressure/composition, are identified and linked, converted to basic cell parameters *a, b, c, α, β, γ* and standardized. After checking, these can be used for retrieval and it is possible to assign approximate atom coordinates.

3. PAULING FILE – Phase Diagrams

The phase diagram section of the PAULING FILE contains temperature-composition phase diagrams for binary systems, as well as horizontal and vertical sections and liquidus/solidus projections for ternary and multinary systems. Both experimentally determined and calculated diagrams are processed. Primary literature is considered in first priority, but diagrams from a few well-known compilations, among which the compendium of binary phase diagrams edited by Massalski *et al.* [29] and the series of books on ternary phase diagrams edited by Petzow and Effenberg [30], have been included.

All the diagrams have been converted to at.% and °C and redrawn in a standardized version, so that different reports for the same chemical system 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 PAULING FILE conventions, but also the original names are stored in the database. Examples of phase diagrams redrawn for the PAULING FILE are shown in Fig. 5.

Each phase diagram is linked to a database entry, which contains the following database fields:

• *Classification:* chemical system; type of diagram
• *Investigation:* experimental/calculated; calculation technique; APDIC/non-APDIC; remark
• *Bibliographic data:* data source; authors (affiliation); language; title
• *Original diagram:* figure number in the original publication; borders; scales; original size
• *Redrawn diagram:* concentration range; temperature (range); conversion of concentration
• *List of 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; link to a representative PAULING FILE crystal structure entry. For binary systems also the temperature and reaction type for the upper and/or lower limit of existence of the *phase* are stored.

4. PAULING FILE – Physical Properties

The physical properties section of the PAULING FILE stores experimental and (to a limited extent) simulated data for a broad range of intrinsic physical properties of inorganic compounds in the solid,

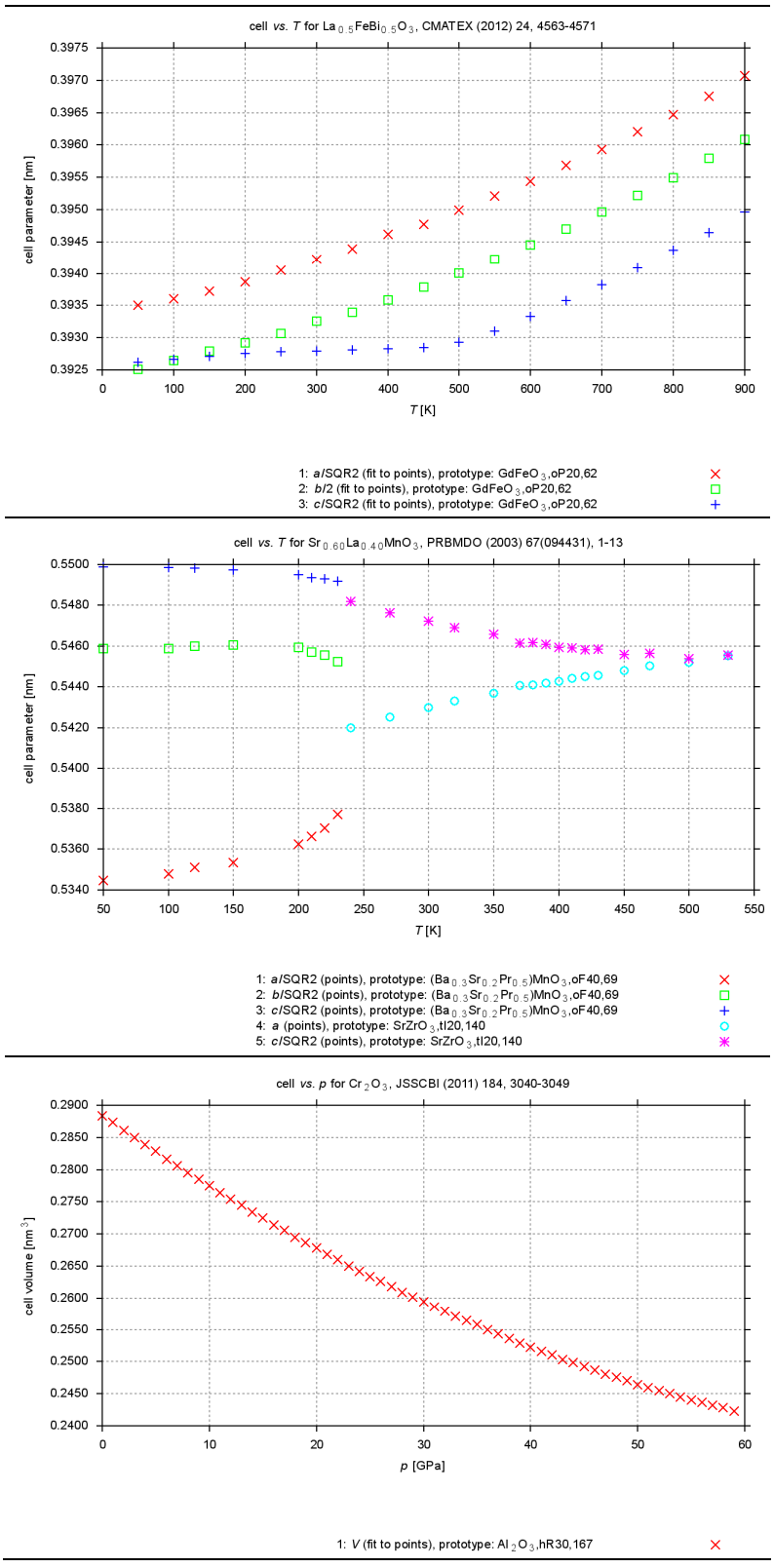


Fig. 4 Examples of cell parameter plots from Pearson's Crystal Data [26], release 2016/17: (a) thermal expansion for $La_{0.5}FeBi_{0.5}O_3$, (b) parameter change through the phase transition for $Sr_{0.6}La_{0.4}MnO_3$, (c) pressure dependence of the cell volume for Cr_2O_3.

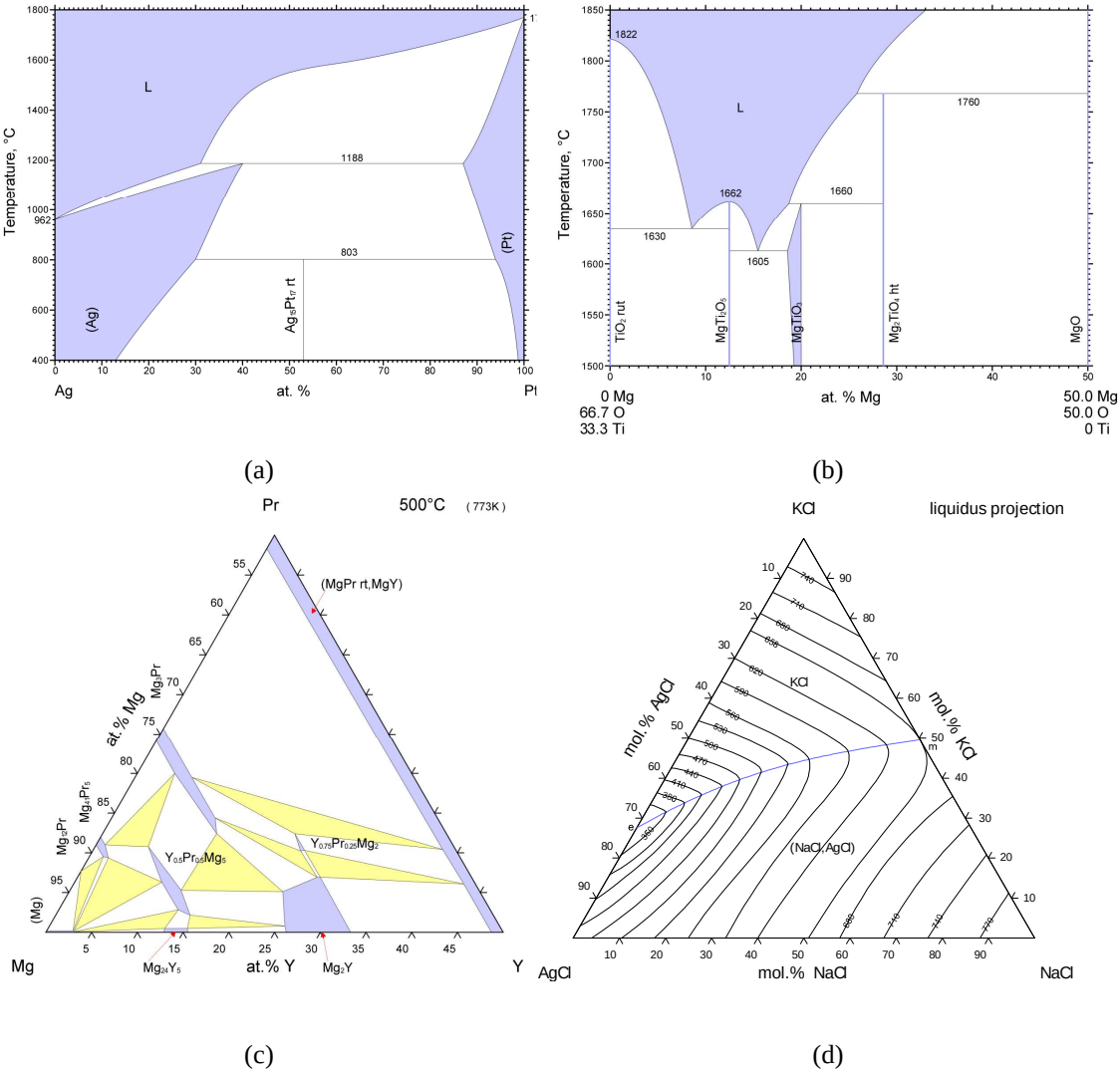


Fig. 5 Examples of phase diagrams as redrawn for the PAULING FILE: (a) phase diagram of a binary system, (b) vertical section and (c) isothermal section of the phase diagram of a ternary system, (d) liquidus projection of the phase diagram of a quaternary system.

crystalline state. Processing is literature-oriented and each database entry groups selected data extracted for a particular *phase* in a particular publication. Focus is on the characterization of inorganic substances (single-phase samples), rather than on the optimization of materials. When published, the entries also contain information about synthesis and sample preparation, as well as information that helps to establish the links to phase diagram and crystal structure entries, such as colloquial names, crystallographic data, limits of stability of the *phase* with respect to temperature, pressure, or composition.

The physical properties are stored in four different ways:

- numerical values,
- figure descriptions (*Y vs. X*),
- property classes such as superconductor, ferroelectric, *etc*.,
- keywords indicating the existence of particular data, *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 [31]. Numerical values are stored in as-published units and also converted to standard units. Most standard units are SI units, however, for certain properties at the atomic level, more suitable units such as eV or μ_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 that particular measurement. Great flexibility is provided through the links to reference tables, thanks to which new properties may be selected and their symbols, units, and ranges of magnitude can be controlled.

4.1 Data selection

Data are taken from primary literature. Each database entry corresponds to a particular combination data source – inorganic crystalline *phase*, but 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 database 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 database entries, whereas magnetic, electric or superconducting transitions are not.

Data for *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 database 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 *ab initio* calculations are also included, in particular energy band structures, but focus is on experimentally measured data and values directly derived from measurements.

4.2 Database fields

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 in the PAULING FILE. The following database fields may be present in a physical properties database entry:

- *Compound:* chemical system; published chemical formula (samples investigated); representative standardized chemical formula; modification
- *Bibliographic data:* data source; authors (affiliation); language; title
- *Preparation:* starting materials (purity, description); method of synthesis (crucible, atmosphere, solvent); annealing or crystal growth
- *Sample description:* sample form; chemical analysis; stability with respect to temperature, pressure, and composition; elastic behavior; density; color; chemical reactivity
- *Crystallographic data:* structure prototype; space group; cell parameters; remark

For each physical property:

- *Numerical values:* symbol; value in published unit; value in standard unit; temperature; other experimental conditions (pressure, magnetic field, wavelength, *etc.*); direction; composition or chemical element; remark
- *Figures:* number in the original publication; parameters; ranges; remark
- *Keywords:* code for additional topic treated in the publication
- *Property class:* one or several property classes

Fig. 6 shows the part concerning the properties of a data sheet taken from the PAULING FILE – Binaries Edition [7].

Optical properties

Refractive index	n= 9.6 information taken from fig. 3(solid); peak value	rt; hν= 5.24 10^{-2} eV
Refractive index	n= 3.55 information taken from fig. 3(solid)	rt; hν= 3.72 10^{-2} eV
Permittivity	ε_∞= 2.68	
Permittivity	ε_1= 13 information taken from fig. 4	rt; hν= 3.64 10^{-2} eV
Permittivity	ε_2= 1.38 10^2 information taken from fig. 4; peak value	rt; hν= 5.32 10^{-2} eV
Permittivity	ε_s= 7.72 $\varepsilon_s = \varepsilon_\infty + \Sigma 4\pi p_i$	rt
Permittivity	ε_s= 8.06 from Lyddane-Sachs-Teller relation	rt
Optical phonon	based on crystal symmetry 1 IR-active optical mode (F_{1u}) and 1 Raman-active mode (F_{2g}) are expected at zero wave vector; triply degenerate F_{1u} mode is split into a doubly degenerate transverse optical (TO) mode and a longitudinal optical(LO) mode (see table 1)	
Data on	Raman scattering	

Figures

Figure	Title	Parameters
1	optical reflectivity - wavenumber diagram of Li_2O at rt IR reflectivity; solid curve: experimental results; circles: calculated from dispersion theory using parameters given in table 1	R[/] vs. 1/λ[30-110 mm^{-1}]
3(broke)	extinction coefficient - wavenumber diagram of Li_2O at rt calculated from dispersion theory using parameters given in table 1	k[/] vs. 1/λ[30-110 mm^{-1}]
3(solid)	refractive index - wavenumber diagram of Li_2O at rt calculated from dispersion theory using parameters given in table 1	n[/] vs. 1/λ[30-110 mm^{-1}]
4	permittivity - wavenumber diagram of Li_2O at rt real (solid curve) and imaginary (broken curve) parts of permittivity, calculated from dispersion theory using parameters given in table 1	$\varepsilon_{1,2}$[/] vs. 1/λ[30-110 mm^{-1}]

Fig. 6 Part of the data sheet of a physical properties entry for Li_2O in the PAULING FILE – Binaries Edition [7].

4.3 Physical properties considered in the PAULING FILE

The physical properties considered in the PAULING FILE belong to one of the following categories: electronic and electrical properties, ferroelectricity, magnetic properties, mechanical properties, optical properties, phase transitions, superconductivity, thermal and thermodynamic properties. Table 3 lists the main properties considered in the PAULING FILE for the first two categories. Items in square brackets are keywords, for which numerical values are in principle not extracted. Primary properties, to which particular attention is paid for the extraction of numerical values, are emphasized with bold characters. Thanks to the flexible construction of the relational database, new properties can easily be added.

5. Data quality

Only reliable data can be used for sensible data mining and great importance is given to the quality of the data in the PAULING FILE. The articles selected for processing are analyzed by scientists specialized in crystallography, phase diagrams, or solid-state physics, most of them with a doctor degree and own experience in solid-state chemistry or physics research [1]. A minimum of 50% editing activity is required, in order to achieve efficiency and homogeneity in data processing and some of the editors have already processed more than 5'000 scientific papers.

5.1 Computer-aided checking

The PAULING FILE data are checked for consistency with the help of an original software package, ESDD (Evaluation, Standardization, Derived Data), containing more than 100 different modules [17]. The checking is carried out progressively, level by level.

Checks on individual database fields:
• formatting of numerical values,
• units and symbols for physical properties,
• Hermann-Mauguin symbols, Pearson symbols,
• consistency journal code–year–volume, first–last page for literature references,
• formatting of chemical formulas,
• usual order of magnitude,
• spelling.

Consistency within individual data sets:
• consistency atom coordinates – Wyckoff letters – site multiplicity,
• comparison of chemical elements in chemical system – chemical formula – refinement –preparation,
• comparison of computed and published values: cell volume, density, absorption coefficient, interplanar spacings,
• consistency Pearson symbol – space group – cell parameters,
• consistency refined composition – chemical formula,
• consistency units – symbols for physical properties,
• consistency Bravais lattice – diffraction conditions,
• consistency site symmetry – anisotropic displacement parameters.

Special checking of crystallographic data:
• comparison of interatomic distances with the sum of atomic radii,
• comparison of interatomic distances within chemical units (carbonates, phosphates, *etc.*),
• check on charge balance for oxides and halides,
• search for overlooked symmetry elements,
• comparison with the type-defining entry (cell parameter ratios, atom coordinates).

Consistency within the database:
• comparison of densities,
• comparison of cell parameter ratios for isotypic compounds,
• check for compulsory data,
• check of database links.

Wherever possible, misprints detected in the original paper are corrected, based on arguments explained in remarks. 5% of the crystal structure entries contain errata referring to misprints in the published crystallographic data. 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/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. The modular construction facilitates incorporation of new checking procedures.

6. Distinct phases

The first part of the challenge in building up a comprehensive database consisted in compiling large amounts of data. However, to provide a global (holistic) overview of the database content and allow combined retrieval, it was also necessary to link the different database entries from the three parts of the PAULING FILE in a more efficient way than provided through the bibliographic information and the chemical system. The concept of Distinct *Phases* was introduced for this purpose.

Table 3 Electronic and electrical and ferroelectric properties considered in the PAULING FILE. Bold characters emphasize primary parameters, square brackets indicate keywords.

Electronic and electrical properties
- metal/nonmetal character
  - **temperature for metal-nonmetal transition**
    - pressure derivative
  - pressure for metal-nonmetal transition
- electron energy band structure
  - [electron energy band structure]
  - [Brillouin zone]
  - [Fermi energy]
  - [Fermi surface]
- electron density of states
  - **electron density of states at Fermi level**
  - [electron density of states diagram]
  - [electron density maps]
- energy gap
  - **energy gap**
    - pressure derivative
    - temperature derivative
    - composition derivative
  - **energy gap for direct transition**
    - pressure derivative
    - temperature derivative
  - **energy gap for indirect transition**
    - pressure derivative
    - temperature derivative
  - **thermal energy gap**
  - exciton energy
    - pressure derivative
    - temperature derivative
- activation energy
- electrical conductivity/resistivity
  - **electrical resistivity**
    - temperature derivative
    - concentration derivative
  - electrical resistivity anisotropy
  - phonon resistivity
    - temperature derivative
  - magnetic resistivity
    - temperature derivative
  - **ionic conductivity**
  - electron conductivity
  - hole conductivity
- residual resistivity
  - residual resistivity
  - **residual resistivity ratio (RRR)**
- spin-disorder resistivity
  - [spin-disorder resistivity data]
- spin-fluctuation resistivity
  - [spin-fluctuation resistivity data]
- piezoresistivity
  - piezoresistivity
    - pressure derivative
    - temperature derivative
  - magnetic contribution to piezoresistivity
- magnetoresistivity
  - magnetoresistivity
  - temperature derivative
- Hall coefficients
  - **Hall coefficient**
    - pressure derivative of Hall coefficient
    - temperature derivative of Hall coefficient
  - **ordinary Hall coefficient**
  - **extraordinary Hall coefficient**
- effective mass
  - **effective mass of electrons in conduction band**
  - effective mass of electrons anisotropy
  - **effective mass of holes in valence band**
    - pressure derivative
  - effective mass of electrons/holes ratio
  - effective mass of polarons
- charge carrier concentration
  - **electron concentration**
  - **hole concentration**
  - electron/hole concentration ratio
  - charge carrier concentration
  - donor concentration
  - acceptor concentration
  - donor/acceptor concentration ratio
- charge carrier mobility
  - **electron mobility**
    - pressure derivative
  - **hole mobility**
    - pressure derivative
  - electron/hole mobility ratio
  - Hall mobility
    - pressure derivative
  - ion mobility
- charge-density wave
  - [charge-density wave energy gap]
- charge transfer
  - **effective charge**
  - **mean valence**
- quadrupole splitting
- [electric-field gradient]

Ferroelectricity
- ferroelectric transitions
  - **ferroelectric Curie temperature**
    - pressure derivative
  - **antiferroelectric Néel temperature**
    - pressure derivative
  - temperature for transition between different ferroelectric states
- permittivity (dielectric constant)
  - **permittivity**
    - pressure derivative
    - temperature derivative
  - real part of permittivity
  - imaginary part of permittivity
  - permittivity change at phase transition
  - **static permittivity**
    - pressure derivative
    - temperature derivative
  - **high-frequency permittivity**
    - pressure derivative
    - temperature derivative
- dielectric loss tangent
- electric polarization
  - electric polarization
  - **spontaneous electric polarization**
    - pressure derivative
  - electric dipole moment
- paraelectric state
  - **paraelectric Curie temperature**
    - pressure derivative
  - paraelectric Curie coefficient
- ferroelectric hysteresis
  - coercive electrical field
  - remanent polarization
- ferroelectric phase diagram
  - [electrical field – composition diagram]
  - [electrical field – temperature diagram]
- piezoelectricity
  - piezoelectric coefficients
- pyroelectricity
  - pyroelectric coefficients

The linkage of the three different groups of data is achieved *via* a Distinct *Phases* table, to which each individual crystal structure, phase diagram, and physical properties entry is linked *via* a coded *phase* identifier (chemical system and an arbitrary number). To prepare this table, each chemical system has been evaluated and the distinct *phases* identified based on information available in the PAULING FILE. As an example, the eight phases reported in the Al–Ta system are listed in Table 4.

A *phase* is in the PAULING FILE defined by the chemical system, the crystal structure (when known), and/or the domain of existence with respect to temperature, pressure, or composition. Each distinct *phase* has been given a unique name containing a representative chemical formula, when necessary followed by a specification such as “ht”, “rt”, “3R”, “hex”, *etc.* The crystal structure is defined referring to the structure prototype, if known. For not yet (fully) investigated structures, partial structural information is given, if available, *e.g.* the complete Pearson symbol may be replaced by *t*** (tetragonal) or *cI** (cubic body centered). Information about colloquial names and stability with respect to temperature, pressure, or composition, collected in the three parts of the database, is used to assign a *phase* identifier to physical properties and phase diagram entries with no structural data. Special cases:

• *Phases* that crystallize with the same structure type, but are separated by a two-phase region in phase diagrams, are distinguished. The same is true for temperature- or pressure-induced isostructural phase transitions where a discontinuity in the cell parameters is reported.

• 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 to the database entries. Structure refinements considering, for instance, split atom positions are often grouped under the parent type.

• Structure proposals stated to be incorrect in later literature have been grouped under a *phase* identifier in agreement with more recent reports. A crystal structure entry reporting a hexagonal cell may in such a case, for instance, be grouped under an orthorhombic *phase*.

• The definition of a structure type applied here makes that a continuous solid solution may smoothly shift from one type to another. A typical case is the progressive transition of a *phase* A_xB from a NiAs-type to a Ni_2In-type structure by filling first one *A* site, then a second one. Refinements considering one or the other type have been grouped together.

• 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.

• By default a paralectric-ferroelectric phase transition is 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.

There exist of course still parts of chemical systems that are little explored and reports in the literature are sometimes contradictory. The *phase* assignment becomes here difficult and the list of distinct *phases* will sometimes contain more *phases* than there exist in reality. It follows that there is a certain amount of subjectivity when assigning a *phase* identifier; we believe, however, that this approach represents a substantial advantage for the user.

6.1 Chemical formulas and phase names

The chemical formulas have been standardized so that the chemical elements are always written in the same order, an order that roughly corresponds to the order of the groups in the periodic system. Chemical units, such as water molecules or sulfate ions, are distinguished and written within square brackets. Deuterium and tritium are considered as distinct chemical elements.

Table 4 Distinct *phases* in the Al–Ta system.

System	at.% Ta	*Phase*	Prototype	Space group
Al-Ta	25	$TaAl_3$	$TiAl_3$,tI8,139	*I*4/*mmm*
Al-Ta	36.11	$Ta_{39}Al_{69}$ ht	$Ta_{39}Al_{69}$,cF444,216	*F*-43*m*
Al-Ta	40	Ta_2Al_3 rt	*,aP*,*	-
Al-Ta	41.67	$TaAl_{1.4}$ rt	*,hP*,*	-
Al-Ta	51.16	$Ta_{22}Al_{21}$	$Ta_{22}Al_{21}$,mP86,14	$P12_1/c1$
Al-Ta	58.62	$Ta_{17}Al_{12}$	$Mg_{17}Al_{12}$,cI58,217	*I*-43*m*
Al-Ta	62.5	Ta_5Al_3	Mn_5Si_3,hP16,193	$P6_3/mcm$
Al-Ta	67	$Ta_{0.67}Al_{0.33}$	$(Cr_{0.49}Fe_{0.51})$,tP30,136	$P4_2/mnm$

In the crystal structure part of the PAULING FILE, whenever a structure type has been assigned to the published data, the chemical formula is written so that the number of formula units per cell is the same as for the type-defining compound. A *phase* containing 50 at.% *A* and 50 at.% *B*, for example, will be called $A_{0.50}B_{0.50}$ if the structure type is Cu,cF4,225 (Z = 4), but *AB* if it is CuAu,tP2,123 (Z = 1) and A_2B_2 if it is Cu_3Au,cP4,221 (Z = 1). A two-phase sample of the same composition would be written $A_{50}B_{50}$. Such conventions imply a certain hypothesis on the atom distribution in the case of off-stoichiometric formulas. In particular it is necessary to choose between a formula assuming a structure with vacancies and one with mixed occupation, *e.g.* between $A_{0.9}B$ and $A_{0.95}B_{1.05}$. Adding to this the uncertainty on the chemical composition itself, especially when the authors did not recognize the crystal structure, this must be taken as a formal way of writing and no claims are made on its correctness.

Each *phase* is assigned a name, which, in the general case, is a representative chemical formula, written as described above. Whenever several *phases* are known for the same chemical composition, a short code specifying the modification is added. Preference is given to terms such as “rt“ (room-temperature), “ht“ (high-temperature), “lt“ (low-temperature), or “hp“ (high-pressure), possibly followed by a digit when a series of temperature- or pressure-induced phase transitions are known. If only one modification, stable at room temperature, is known, the field modification is left blank. The specification “ht“ is in principle added for *phases* that are only stable above room temperature (298.15 K), and by analogy, the specification “lt“ for *phases* that are only stable below room temperature. In cases where no or contradictory information about phase stability is found in the literature, a specification such as “cub“ (cubic), “rhom“ (rhombohedral), “orth“ (orthorhombic), *etc.*, may be preferred. Ramsdell notations are used for polytypic compounds such as CdI_2. Mineral names can also be used as specifications, and are then abbreviated to the first three letters.

Special notations are used in the phase diagram part, where a chemical element in parentheses indicates a terminal solid solution based on this element. For complete solid solutions two or more chemical elements, or chemical formulas (if relevant, with specifications) are written within parentheses, separated by commas, *e.g.* (LiBr,AgBr) or (Ag_2La,Ag_2Ce rt).

6.2 Phase classifications

A certain number of characteristics, attributed to the *phases*, are stored in the table Distinct *Phases*.

• *Compound classes*: The classification into compound classes is to a first extent based on the existence of complex anions such as sulfate, nitrate, carbonate, fulleride, *etc*. Simple compound classes, such as intermetallics (both elements situated on the left hand-side of the Zintl line of the periodic system), oxides, sulfides, *etc.*, are also distinguished, as well as hydrates.

• *Structure classes*: Certain structure prototypes have been grouped into families, initially based on crystal-chemical tables in TYPIX [22]. The family *close-packed structures*, for instance, groups structures built up of close-packed layers in any kind of stacking, without interstitial atoms. The structure classes *perovskites*, *AlB_2 family*, *close-packed structures*, *b.c.c. atom arrangement*, *rocksalt family*, and *high-T_c cuprates* have at present the largest numbers of representatives. The nomenclature of zeolites, using 3-letter codes to characterize different frameworks, is taken from the Database of Zeolite Structures [32]. It may be noted that since the classification is applied to the prototypes, *phases* that have not been assigned a structure prototype will also not be assigned a structure class.

• *Property classes*: Property classes such as antiferromagnet, ferroelectric, metal, semiconductor, ionic conductor, superconductor, *etc.*) are distinguished based on data available in the physical properties part of the PAULING FILE. It follows that a *phase*, which from the chemical formula is expected to have metallic character, will not be assigned this class if properties leading to this conclusion have not (yet) been processed for that particular *phase*. On the contrary, *phases* with a significant range of existence in composition, temperature or pressure, may exhibit very different properties, depending on the doping level, temperature or pressure, and all of the property classes assigned to the *phase* (*e.g.* antiferromagnet, ferromagnet, and spin glass for the same *phase*) may not apply to the representative chemical formula, or only to a particular temperature or pressure range.

• *Mineral names:* The names reported in the original publications have been checked by consulting Strunz Mineralogical Tables [33] and the list of minerals approved by the International Mineralogical Association [34], and updated consequently. The mineral names are stored in the table Distinct *Phases*, so that all database entries for this *phase* will be linked to this information. When a continuous solid solution has been confirmed experimentally, several mineral names have sometimes been assigned to the same *phase*, *e.g.* enstatite/ferrosilite or annite/phlogopite 1M.

• *Color:* Color has tentatively been assigned also at the phase level [35], but is in some cases strongly composition-dependent, or due to small amounts of impurities.

7. Towards a megadatabase

After almost 25 years of existence, the PAULING FILE has reached a respectable size in the fields of crystal structures and phase diagrams of inorganic

substances. Focus is here on the yearly update, and old, not yet processed publications represent a few percent. On the contrary, in spite of the relatively high number of database entries, the coverage of physical properties is still at a low level, considering the huge amount of data published in this field. In 2016, the PAULING FILE contains over 310’000 structural data sets (including atom coordinates and displacement parameters, when relevant) for some 140’000 different *phases*, more than 44’000 phase diagrams (with updated *phase* assignment) for near 9’800 chemical systems, and 120’000 physical properties entries (about 420’000 numerical values and 130’000 figure descriptions) for some 50’000 *phases*.

To reach this result, over 140’000 scientific publications have been processed, from more than 1’500 different journals. Some 250 scientific journals are browsed from cover to cover for the yearly updates. Fig. 7 shows the distribution of the database entries according to the top journals in each part of the database, where in some cases related titles have been grouped (*e.g.* Journal of Less-Common Metals is included under its successor Journal of Alloys and Compounds). As a principle, only primary literature is considered in the PAULING FILE, but also a few handbooks have been processed for the phase diagram part. It may be noticed that the number of database entries processed from “others” is particularly high for crystal structures.

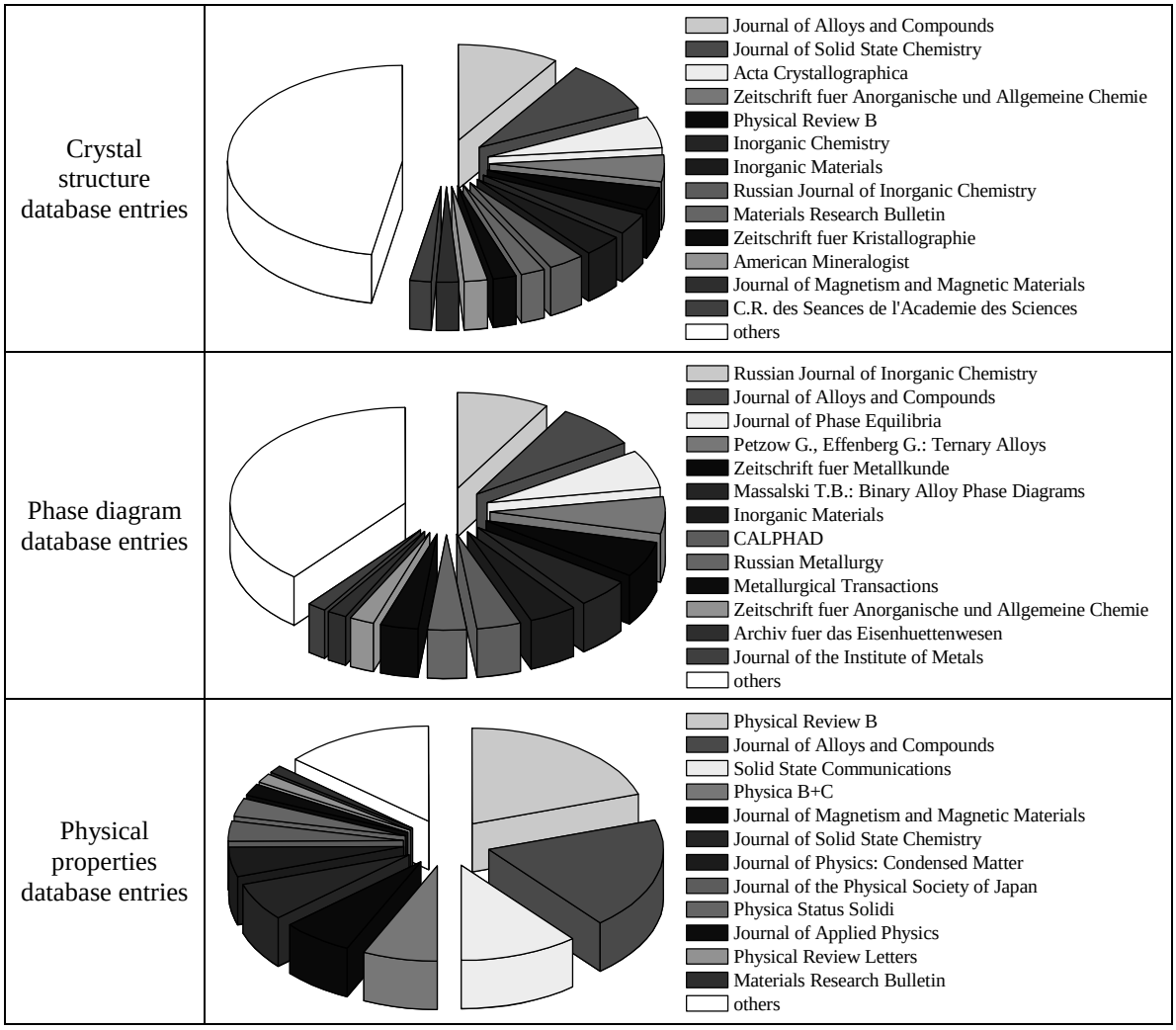


Fig. 7 Distribution of the database entries in the PAULING FILE (June 2016) according to the data source. The journals are listed in decreasing order of the number of database entries, in clockwise order from the top on the diagrams.

Fig. 8 shows the distribution of the database entries per publication year. The regular shape of the diagrams for crystal structure and phase diagram entries confirms the good coverage of the world literature for these two sections of the PAULING FILE. The diagram of crystal structure entries (Fig. 8(a)) also shows the proportion of database entries with refined (or fixed) coordinates, which, thanks to the development of the experimental methods and software for structure refinement, has increased significantly over the last 20 years. Fig. 8(b) confirms that the number of experimental investigations of phase diagrams per year is decreasing, whereas the number of thermodynamic assessments is increasing.

The third overview, shown in Fig. 9, proves that, contradicting common ideas based on earlier works by the same authors (*e.g.* [36]), the PAULING FILE is not limited to intermetallics. On the contrary, except for the phase diagram part, oxides dominate. Fig. 10 shows the number of numerical values, figure descriptions, and keywords, processed in June 2016, distributed over the eight property categories considered in the PAULING FILE (electronic and electrical properties, ferroelectricity, magnetic properties, mechanical properties, optical properties, phase transitions, superconductivity, thermal and thermodynamic properties). The most common physical properties extracted from the publications are magnetic susceptibility, electrical resistivity, heat capacity, and different transition temperatures.

Tables 5 and 6 give some numbers from the main product for the crystallographic data, Pearson's Crystal Data [26], release 2016/17. The first table shows the distribution according to the number of chemical elements, and the second one the distribution according to the level of structural investigation. It can be seen from the latter that the entries have been classified into 36'080 different structure prototypes. Each year some 15'000 new entries are added to Pearson's Crystal Data, most of them based on recent literature.

8. Applications

Thanks to the large amount of information stored in 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. It is, for example, possible to search for inorganic substances having low density (< 3 Mg m^{-3}) and a high melting point (> 2'700 K). Several distinct *phases* fulfill these requirements, among them AlP and BN cub with adamantane structures, tetragonal (?) BeO ht, and CaS, which crystallizes with the structure type NaCl,cF8,225. Other examples could be quaternary ferro(ferri)magnets with ordering temperatures above

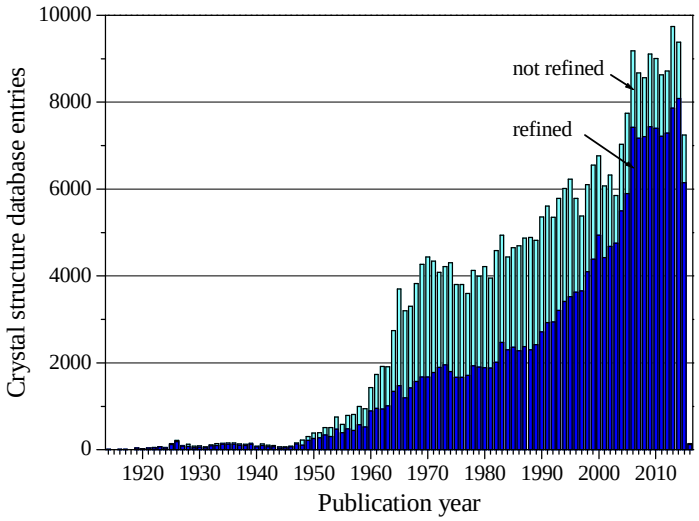
Table 5 Numbers of distinct chemical systems, *phases*, and entries in PCD-2016/17, subdivided into 1, 2, 3, and more than 3 chemical elements, and the total numbers.

Number of chemical elements	Number of chemical systems	Number of *phases*	Number of entries
1	97	434	3'017
2	2'570	18'496	52'839
3	18'096	61'358	112'309
> 3	40'142	85'063	120'656
any	**60'905**	**165'351**	**288'847**

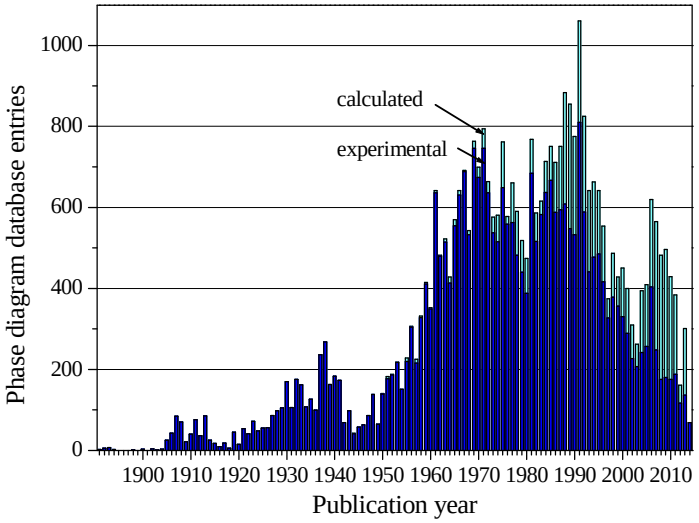
Table 6 Number of entries in PCD-2016/17 according to the level of structural investigation.

Level of structural investigation[1]	Number of entries
all atom coordinates refined or fixed, data set defining a prototype	36'080
all atom coordinates refined or fixed, not type-defining	148'298
part of atom coordinates determined	640
cell determined, prototype and atom coordinates assigned by the editor	85'455
cell determined for filled-up derivative, parent type assigned by the editor	1'297
cell determined	17'077
Total	**288'847**

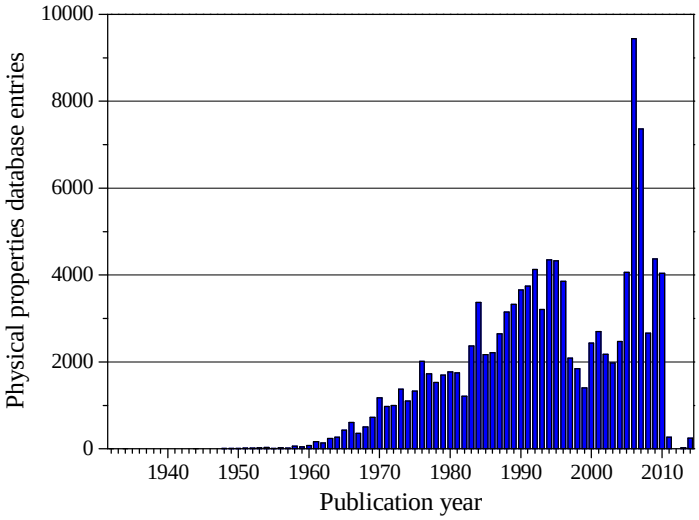
[1] positions of protonic hydrogen atoms are ignored in the classification



(a)



(b)



(c)

Fig. 8 Distribution of the database entries in the PAULING FILE (June 2016) according to the publication year: (a) crystal structure, (b) phase diagram, (c) physical properties database entries. Phase diagram data from handbooks are not included.

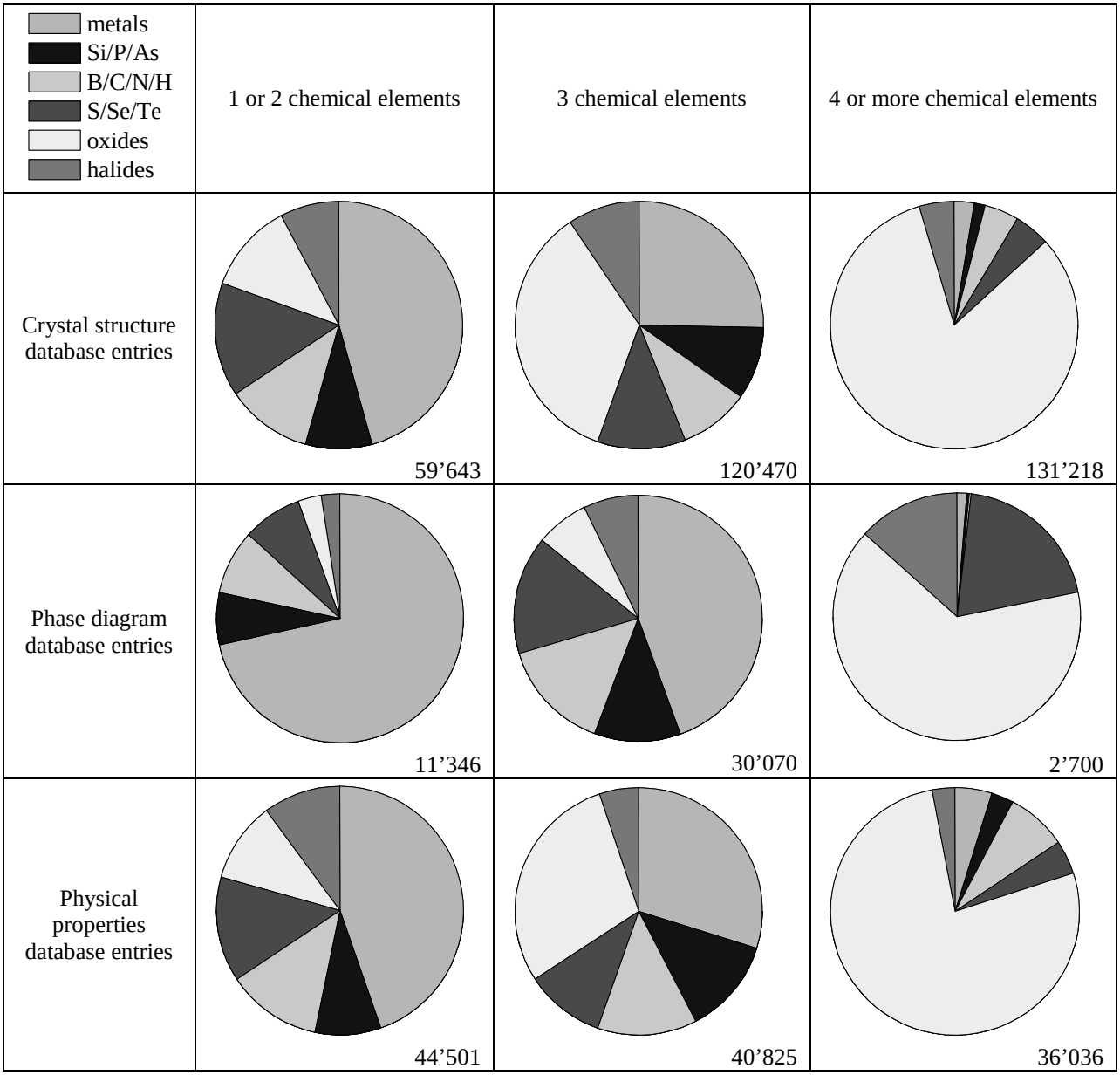


Fig. 9 Distribution of database entries in the PAULING FILE (June 2016) according to the chemical class. The order in the legend corresponds to clockwise order, starting from the top, on the diagrams.

600 K, such as parts of the spinel (solid solution) *phases* $CrFeNiO_4$, $TiFeCoO_4$, and $Zn_{0.5}Mn_{0.5}Fe_2O_4$, or carbides ordering antiferromagnetically above 30 K (*e.g.* several RC_2 and $R_2Fe_{14}C$ compounds, where *R* is a rare-earth metal), or ionic conductors containing Ag and crystallzing with a cubic structure (halides, chalcogenides, including phases adopting the structure type $RbAg_4I_5$,cP80,213).

8.1 Products containing PAULING FILE data

The hundreds of interconnected database fields can be used as LEGO pieces to create different products. PAULING FILE data are included in several on-line, off-line and printed products, of which part are listed below. 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 listed below are limited to PAULING FILE data, whereas others also contain data from other sources.

• **ASM Phase Diagram Database**, ASM International (on-line) [37]
The Phase Diagram Database offers easy viewing of phase diagram details, crystallographic and reaction

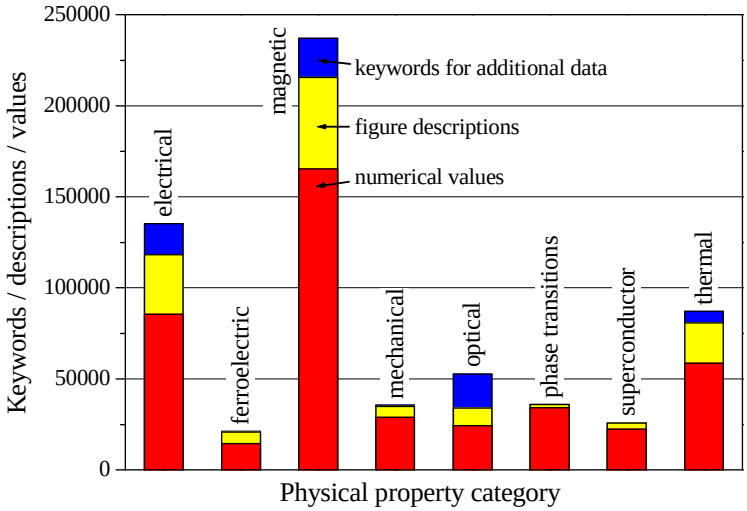


Fig. 10 Number of items in the physical properties part of the PAULING FILE (June 2016) according to the property category and the data category; from bottom to top for each column: numerical values, figure descriptions, keywords for additional data.

data. The content is updated on an annual basis and the 2016 update brings the database to more than 40'000 on-line phase diagrams for binary and ternary systems.

• **Inorganic Material Database (AtomWork)**, NIMS (on-line) [6]
The data part of AtomWork is the result of collaboration between Japan Science and Technology Corporation (JST), the National Institute for Materials Science (NIMS), and Material Phases Data System (MPDS). The Inorganic Material Database aims to cover all basic crystal structure, X-ray diffraction, physical properties and phase diagram data of inorganic and metallic solids from main literature sources. In 2016 (last update of the data part in 2000) AtomWork contains 82'000 crystal structures, 55'000 physical properties and 15'000 phase diagrams. A new release, which will also contain recent data, is under development.

• **PAULING FILE – Binaries Edition**, ASM International (off-line) [7]
The Binaries Edition of the PAULING FILE, which is limited to binary compounds, was published in 2002. It contains 8'000 phase diagrams covering 2'300 binary systems, 28'300 structural data sets for more than 10'000 different *phases*, roughly 3'000 experimental and 27'000 calculated diffraction patterns, and around 17'300 physical-property entries (with about 43'100 numerical values and 10'000 figure descriptions) for some 5'000 *phases*. To reach this result, 21'000 original publications had been processed. Even if restricted to binary compounds, the data contained on the CD-ROM equals over 30'000 printed pages, *i.e.* a 20 volume Handbook! The user-friendly retrieval program offers numerous possibilities, including some data-mining options.

• **Pearson's Crystal Data**, ASM International (off-line and on-line) [26]
The tenth release of Pearson's Crystal Data (2016/17) contains 288'846 structural data sets for 130'000 different phases. Differently from the similarly named Pearson's Handbook, the electronic product contains data for all classes of inorganic substances (approx. 50% oxides). All data sets with published coordinates, and 80% of the data sets where only cell parameters were published, have been assigned a structure type: 185'000 data sets with published atom coordinates, 85'000 data sets with assigned atom coordinates, 19'000 data set with only cell parameters. Atomic environments have been defined for the first category. The crystallographic data are presented as published and standardized, and are accompanied by experimental details and remarks. In addition, the product contains: 18'300 experimental and 271'000 calculated diffraction patterns; 40'000 descriptions of cell parameters as a function of temperature, pressure, or composition, 13'000 plots; 100'000 unit cells extracted from plots vs. T or p; links to the publication, PDF4+, ASM Phase Diagram Database, SpringerMaterials. The software offers numerous possibilities to retrieve and process the data.

• **Powder Diffraction File PDF4$^+$**, ICDD (off-line) [38]
Since 1940 ICDD provides tools in the form of experimental and calculated powder patterns for phase analysis based on diffraction methods. PDF-4$^+$ (inorganic solids) and PDF-4 Minerals include also atom coordinates, which can be used to perform

Rietveld refinements. Over two thirds of the structures in the current edition of PDF-4$^+$ originate from the PAULING FILE. PAULING FILE entries, containing more data, replace duplicate reference patterns and citations of other origin.

• **SpringerMaterials**, Springer (on-line) [8]
Based on the well-known series of Landolt-Börnstein Handbooks, SpringerMaterials allows materials scientists to identify materials and their properties by offering access to physical and chemical data in materials science on an on-line platform. The PAULING FILE provides crystal structure, phase diagram, and physical properties entries on a yearly basis to the section *Inorganic Solid Phases*.

• **World Materials**, MPDS (on-line, planned release mid-2017) [9]
World Materials is a web platform, presenting on-line the three parts of the PAULING FILE data. In its first release, planned for mid-2017, it will contain 45'300 phase diagrams, over 400'000 crystal structures, and over 500'000 physical properties entries. About 80% of the data can be requested remotely in a developer-friendly format, ready for external data-mining applications. The remaining 20% can be obtained as references to the original publications. Altogether 265'000 scientific publications in materials science, chemistry, physics, *etc.* serves as starting point for the platform, and this number will steadily increase. Main focus is on the *verbatim* representation of the original scientific data, however, convenient for quick re-use and re-purposing. A web-browser (without any plugins) and internet connection will allow comfortable work with the scientific data, be it for a literature overview, evaluation of hypotheses, or design of new materials.

The *Landolt-Börnstein* handbook series *Inorganic Crystal Structures* [39] and the *Handbook of Inorganic Substances* [35], also contain PAULING FILE crystal structure data. The former describes structure prototypes in space groups 123-230, whereas the most recent edition of the latter lists crystallographic data for 157'000 inorganic *phases*. The software proposed by Materials Design Inc. to perform *ab initio* calculations [40] also contains crystallographic data from the PAULING FILE. The electronic book *Inorganic Substances Bibliography* [41], lists publications selected for processing in the PAULING FILE, ordered according to the chemical systems considered in the papers.

8.2 Holistic overviews based on the PAULING FILE

For any data mining or statistical approach to inorganic crystalline substances, the prototype classification of their crystal structures represents a key-point, since it offers a "window" to view the electronic interactions of the atoms. In 2016, more than 36'000 different prototypes, as defined in the PAULING FILE, have been experimentally established for inorganic compounds.

Several strong patterns have been revealed in maps using as coordinates elemental-property parameters (or expressions of these), based on thousands of data sets for different chemical systems/compounds [42-45]. This proves that the underlying quantum mechanical laws can be parameterized using elemental-property parameters of the constituent chemical elements. An appropriate choice of parameters leads to relatively simple maps with well-defined stability domains, offering excellent overviews of experimentally known inorganic substances. The maps provide, as a direct consequence, some possibilities to predict features of not yet known compounds.

Particularly nice overviews of the phase diagrams of binary systems can be obtained using the Constitution Browser in the PAULING FILE – Binaries Edition [7]. Fig. 11 shows all available binary Mo–*X* phase diagrams in a periodic table representation. It can be seen that the systems exhibit certain regularities, *e.g.* all Mo–s^1 (s^1 = H, Li, Na, K, Rb, Cs, Fr) systems are non-formers, which means that no true binary compounds form under ambient conditions.

Fig. 12 shows an "Inorganic Solids Overview – Elemental Property Parameter Map" in the form of a generalized Atomic Environment Type (AET) matrix using as coordinates PN_A *vs.* PN_B, where PN_A and PN_B are the periodic numbers (to a first approximation, the periodic number runs from top to bottom and from left to right, column by column, through the periodic system: Li: 1, Na: 2, K: 3, and so on) of the chemical elements *A* and *B*, respectively [44]. The map on the left hand-side is based on experimental data, whereas the equivalent map on the right hand-side shows simulated (or extrapolated) data, making it possible to estimate in one glance the agreement or disagreement between experimental and simulated data.

8.3 Principles defining ordering of chemical elements

Before initiating the PAULING FILE project, in 1994 one of us reviewed the world literature, focusing on intermetallics and alloys on the topic *"Factors Governing Crystal Structures"* [46], and came up with 9 quantitative principles. The conclusions were based on the second edition of *Pearson's Handbook of Crystallographic Data for Intermetallic Phases* [36], which covers about 28'000 intermetallics and alloys (including a few oxides). 20 years later, having now access to Pearson's Crystal Data [26], release 2013/14, with structural information for over 165'000 distinct *phases* (intermetallics, but also oxides,

Fig. 11 Example of the Constitution Browser in the PAULING FILE – Binaries Edition [7], showing phase diagrams of binary systems containing Mo.

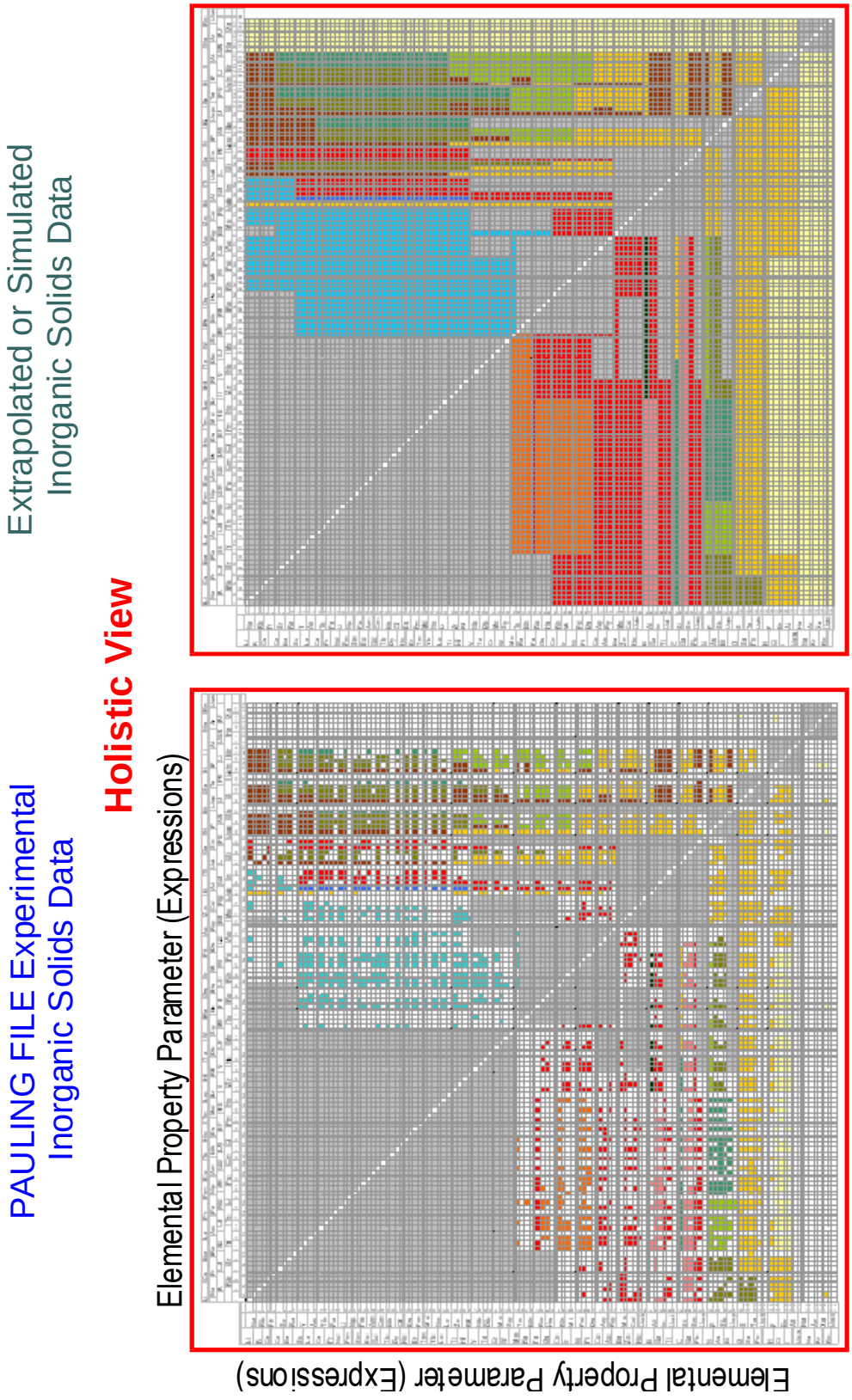


Fig. 12 A generalized Atomic Environment Type (AET) matrix PN_A *vs.* PN_B, which is independent of the stoichiometry and the number of chemical elements in the inorganic solid. The element occupying the center of the AET is given on the *y*-axis and the coordinating element on the *x*-axis. Different colors represent different AETs, gray fields correspond to non-former systems. The results for experimentally determined data are given on the left hand-side, simulated or extrapolated data on the right hand-side [44].

halides, *etc.*), *i.e.* almost six times more experimental data than Pearson's Handbook, a new study was undertaken [45]. Most of the examples given below are based on the content of Pearson's Crystal Data [26], release 2016/17, hereafter referred to as PCD-2016/17.

When chemical elements combine to form solid compounds, their crystal structures are beautifully rich, yet systematic patterns underlie this process. The most striking manifestation of this fact is the existence of crystal structure prototypes, which can be understood as geometrical templates adopted by large groups of compounds, *e.g.* the prototype NaCl,cF8,225 (rocksalt) is adopted by 1'392 *phases* in PCD-2016/17. Different compounds crystallizing in the same prototype are either geometrically identical, or very similar to each other. The work from 1994 focused on the 1'000 most populous prototypes and their representatives. The about 1'000 most frequent prototypes in PCD-2016/17 (987 prototypes adopted by at least 28 *phases*) cover about 70% of all the entries.

The four statistical plots shown below quantitatively illustrate "*the core principle that defines ordering of chemical elements within a prototype*".

(1) Simplicity principle

Fig. 13 shows that the majority of the phases crystallizing with one of the about 1'000 most frequent prototypes have less than 40 atoms per unit cell, with the maximum at 12. This principle was formulated in 1994 [46] as follows*:* "*The vast majority of the intermetallic compounds have less than 24 atoms per unit cell*". Considering all inorganic substances as defined in the PAULING FILE (no C-H bonds) the 24 atoms per cell have become 40, nevertheless the maximum remains near 10 atoms per unit cell. "*In addition the majority of the crystal structures have three or less Atomic Environment Types (single-, two-, and three-environment types)*". This statement is still supported and an analog observation can be made focusing on the number of point sets (atom sites) per prototype. Fig. 14 shows that the majority of the 1'000 most common prototypes (and therefore also their representatives) have 6 or less different Atomic Environment Types, with a maximum at 3 different AETs. The number of point sets per prototype (see the same figure) is for the majority of the *phases* also less than 6, with a maximum at 3 point sets per prototype.

(2) Symmetry principle

Fig. 15 shows the distribution of the *ca* 165'000 *phases* in PCD-2016/17 according to the space group number. The symmetry principle was initially formulated as: "*The vast majority of all intermetallic compounds and alloys crystallize in one of the following 11 space groups: 12, 62, 63, 139, 166, 191, 194, 216, 221, 225, and 227*". Extending the statistics to all classes of inorganic compounds considered in the PAULING FILE, a few more space groups have to be added: 2, 14, 15, 123, 129, 136, 140, 148, 164, 167, 176, and 229. It appears from the figure that, within each crystal system, certain space groups of high symmetry are preferred. As a consequence, 10% of the 230 space groups count for 67% of the entries in PCD-2016/17.

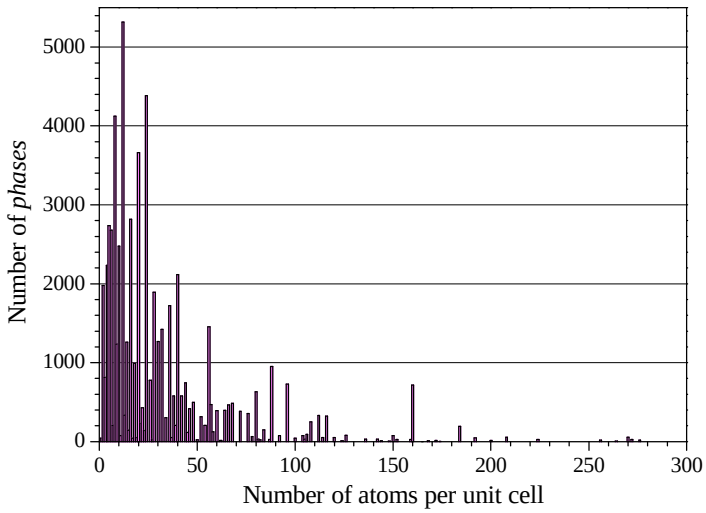


Fig. 13 Number of *phases vs.* the number of atoms per unit cell, considering the representatives of the about 1'000 most common prototypes in PCD-2016/17.

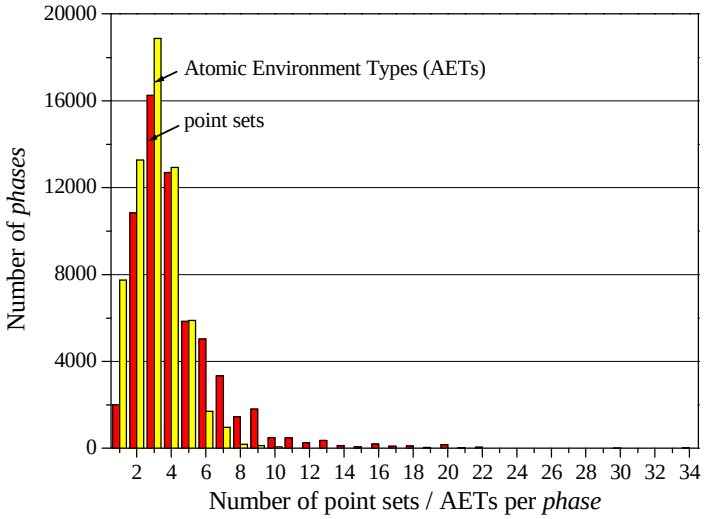


Fig. 14 Number of *phases* according to the number of different AETs (right column), respectively number of point sets (left column), in the structure, considering the representatives of the about 1'000 most common prototypes in PCD-2016/17**.**

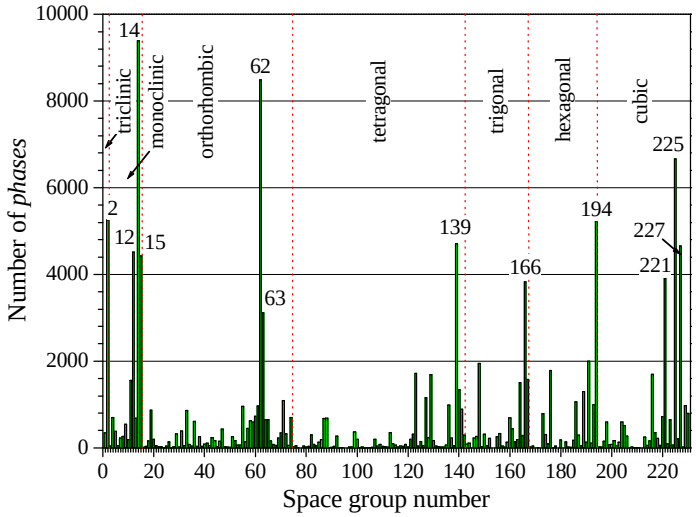


Fig. 15 Number of *phases* according to the crystal system and space group number in PCD-2016/17.

(3) Atomic-Environment principle

Fig. 16 shows a frequency plot for the 18 most often observed AETs, considering the representatives of the about 1'000 most frequent prototypes in PCD-2016/17. In 1994 the Atomic-Environment principle said: "*The vast majority of all atoms (point sets) in intermetallic compounds have as Atomic Environment one or several of the following 14 AETs: tetrahedron, octahedron, cube, tricapped prism, fourcapped trigonal prism, icosahedron, cuboctahedron, bicapped pentagonal pyramid, anticuboctahedron, pseudo Frank-Kasper (CN13), 14-vertex Frank-Kasper, rhombic dodecahedron, 15-vertex Frank-Kasper, and 16-vertex Frank-Kasper*". The statement, made based on intermetallics, that certain AETs are highly preferred, is still correct, but by considering also other classes of inorganic compounds the order has changed, and the following low-coordination AETs: single atom (CN = 1), collinear (CN = 2), non-collinear (CN = 2), coplanar triangle (CN = 3), non-coplanar triangle (CN = 3), and square antiprism (CN = 8) have appeared among the most popular AETs. These 18, out of the 100 Atomic Environment Types distinguished in the PAULING FILE, are found for 90% of the point sets in the representatives of the most common prototypes. One may conclude that Nature strongly prefers certain AETs, most of them highly symmetrical.

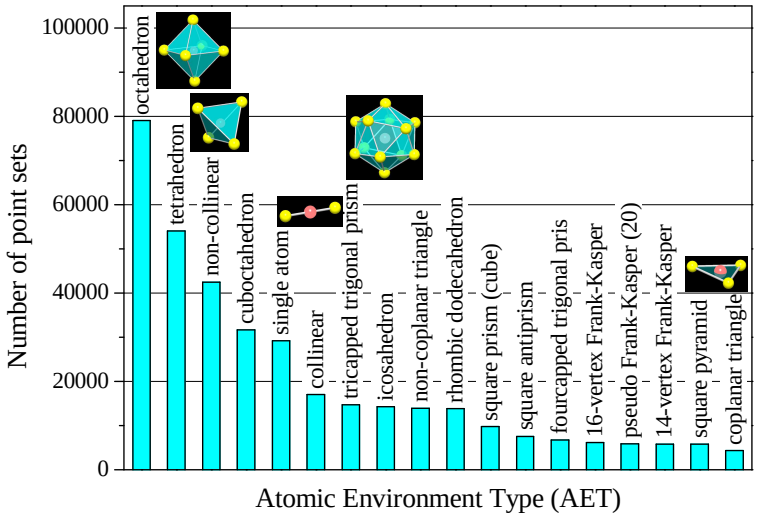


Fig. 16 Total number of point sets observed for the 18 most frequently occurring AETs, considering the representatives of the about 1’000 most common prototypes in PCD-2016/17.

Table 7 Number of distinct *phases* in PCD-2016/17 comparing the number of chemical elements in the type-defining database entry (rows) and the number of chemical elements in all representatives (columns), considering the about 1’000 most common prototypes. The diagonal corresponding to the same number of elements as in the type-defining entry is emphasized.

Number of elements / prototype	Number of phases according to number of chemical elements							Total number of entries	Total number of *phases*
	1	2	3	4	5	6	> 6		
1	**242**	1’312	543	55	30	15	5	6’797	2’202
2	3	**6’454**	9’877	1’798	298	65	54	62’282	18’549
3	0	65	**17’410**	7’980	2’667	454	130	83’143	28’706
4	0	0	433	**5’958**	2’721	508	147	23’197	9’767
5	0	0	16	286	**1’038**	385	104	4’272	1’829
6	0	0	0	47	78	**225**	118	1’492	468
7	0	0	1	14	12	35	**205**	843	267

(4) Ordering tendency principle

Table 7 gives the number of *phases* crystallizing in one of the about 1’000 most popular prototypes in PCD-2016/17 in a representation showing the number of chemical elements in the type-defining entry (rows) *vs.* the number of chemical elements in all isotypic *phases* (columns). High numbers are found along the diagonal, where the numbers of chemical elements are identical. Relatively high numbers are also observed for *phases* containing more chemical elements than the type-defining entry. However, cases where the number of elements is lower than in the type-defining entry are rare, in part due to the definition of a structure type used here, where different ordering variants (substitution derivatives) are distinguished.

It is interesting to note that the some 1’000 most popular prototypes are represented by 182’026 entries in PCD-2016/17, but only by 61’788 distinct *phases*. Among the entries, only about one half have no sites with mixed occupation. This means in general structures where the number of chemical elements is the same as for the prototype. The structures of the remaining database entries do contain mixed sites. Such database entries are likely to be part of solid solutions and are in this case not true distinct phases in the commonly accepted sense. For example, partial

replacement of the chemical element *A* in a compound *ABC* by a few at.% of a closely related chemical element *A*' may lead to a quaternary representative (*A*,*A*')*BC*, which in the PAULING FILE will be considered as a distinct *phase*.

The systematic patterns described above lead to restraint conditions expressed in the below listed four principles, summarizing the preference of Nature for:

• simplicity;
• particular overall symmetries;
• high local symmetry (symmetrical AETs);
• ordering of the chemical elements (distinct chemical elements occupy distinct atom sites).

The combination of the above given experimental observations reduces the number of potential prototypes for an unknown inorganic compound to a few hundred of the most common prototypes, *i.e.* approximately 1-2% of the experimentally known prototypes. Fig. 17 shows a frequency plot for the representatives of the 100 most frequent prototypes in PCD-2016/17. Seen from the opposite point of view, the large majority of the 36'080 prototypes in PCD-2016/17 (near 80%) have less than four representatives. One of the reasons for the high number of prototypes is the increasing number of refinements revealing a high degree of disorder and sites with low occupancy. For example, most structure refinements of ionic conductors, complex minerals, zeolites, or hydrides, represent distinct prototypes.

9. Lessons to learn from experience

Even if it takes half a century to reach the critical size for database sustainability it is worthwhile to re-examine the startup thoughts to get a holistic view on materials. As comparative cases, the PAULING FILE and VEMD projects were selected as BUA (bottom-up approach) and TDA (top-down approach), respectively. For the case of developing the PAULING FILE, some key-points on how to overcome the complexity and diversity of materials to transform them into values are listed below:

(1) Define core scientific principles for target materials explicitly. At least one holistic view as a set of digital knowledge is required, which enables continuous quality refinement of newly added data logical deductive confirmation. Geometric group theory: crystallography is the principle as described in detail in the preceding sections. Other data difficult to refine deductively can be refined inductively through *ad hoc* holistic views generated from the compiled digital data.

(2) Implement systematic and graded procedures for highest data quality – digital data compilation, escaping from a conventional way of "manual" industry. Digital data have been produced as products of digital manufacturing industry.

(3) Follow business to business (B to B) models with copyright contracts, as well as open strategies.

(4) Keep human resources with sufficient knowledge for (1) and (2), and, most importantly,

(5) Link the data *via* a closed space defined by crystallography. So as to reduce uncertainties due to inductive evaluations, all the data are linked to deductively evaluated digital data. IDs for phases are linked for such data as phase diagrams and intrinsic properties.

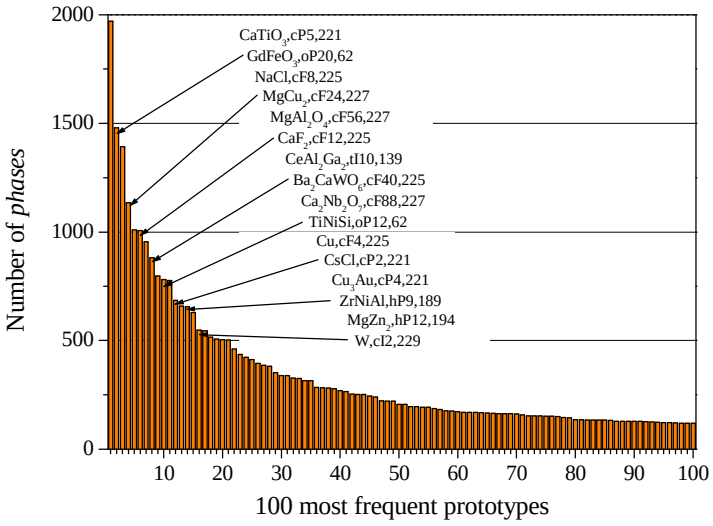


Fig. 17 Number of representatives (*phases*) of the 100 most common prototypes in PCD-2016/17.

These are the prerequisites for a holistic view in the way where the whole should be greater than the sum of its parts. And, consequently, the whole becomes a window on the diversities of materials decorated by impurities, alloying elements and complicated defects.

However, for engineering applications additional work is required to bridge other holistic views, namely, design windows in terms of materials properties, performances and functions and the corresponding holistic views created by materials scientists and materials producers. The former design window is used as a starting point for backcasting from the requirements, and the latter is ideally created as a summary by materials scientists and producers. Basic design windows about engineering materials were compiled and systematized into a digital system, CMS (Cambridge Materials Selectors), for all engineering materials by M.F. Ashby (1980s), and such design windows are shared as a set of "common sense" by materials engineers, even if not explicitly, in the first phase of materials development. *Ad hoc* articulations for materials selection problems have been carried out by resolving each problem into a set of sub-problems, where the above guidelines can be applied. The VEMD project [4], carried out during the period 1995-1999, aimed to show exemplars of such bridging and converging procedures as a new approach to materials design. So the VEMD project is one of the prototyping projects integrating elementary technologies, such as numerical simulation, knowledge information processing, database, and human interface technology. The main objectives of the VEMD project were:

VEMD-1: Make a digital copy of the engineering materials world in terms of data and models.

VEMD-2: Acquire practical knowledge from the copy.

VEMD-3: Create design scenarios to answer requests from materials users.

VEMD-4: Add necessary data by simulation and/or experiment to follow the design scenario.

VEMD-5: Evaluate the designed materials from a viewpoint of material users.

Fig. 18 shows a schematic overview of the project [47].

As design targets, two groups were selected, namely, high-temperature superalloys and electronic materials. In the former case, it appeared too complex and too complicated to design materials, due to their time-, space- and temperature-dependent features developed in open space, and the design strategy was reduced to:

Design Tactic 1: Identify the most promising exemplar in the past.

Design Tactic 2: Analyze the selected exemplar and resolve the problem into a set of sub-problems in terms of intrinsic properties and extrinsic structure-sensitive properties, following the resolution principle by Robinson [48].

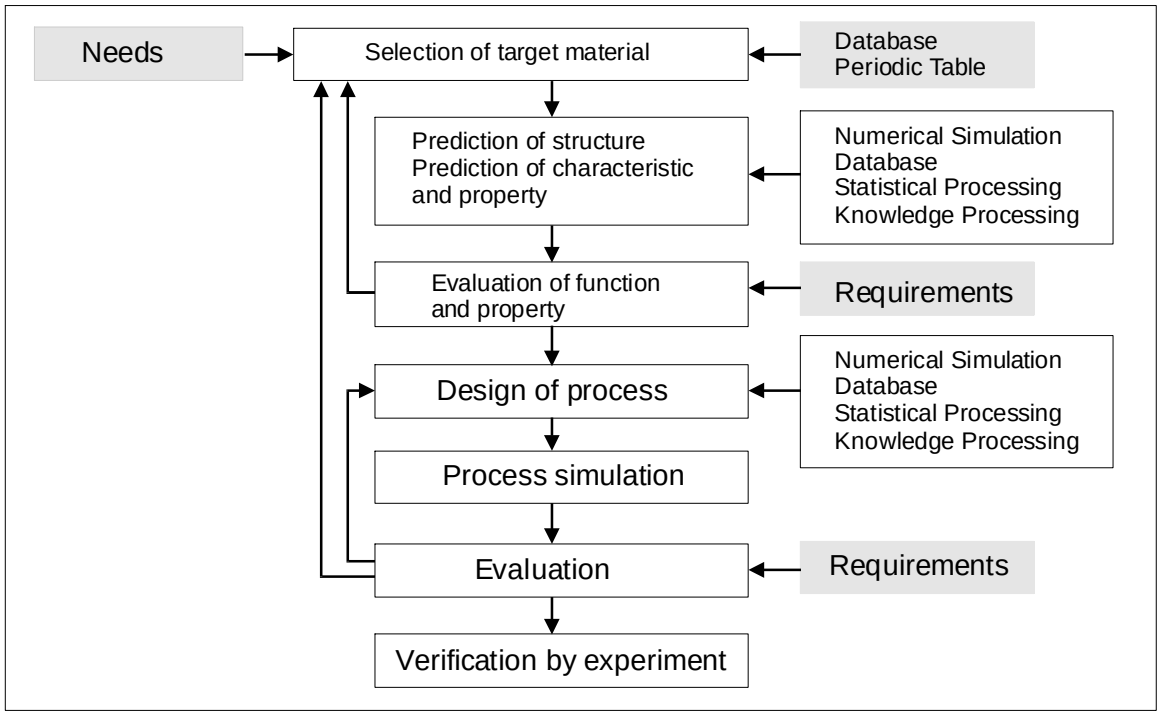
Design Tactic 3: Carry out comparative studies in accordance with intrinsic properties to discover candidate materials, following the way of dealing with complexities as summarized by Masahiko [49].

Design Tactic 4: Evaluate the candidate materials by experiments.

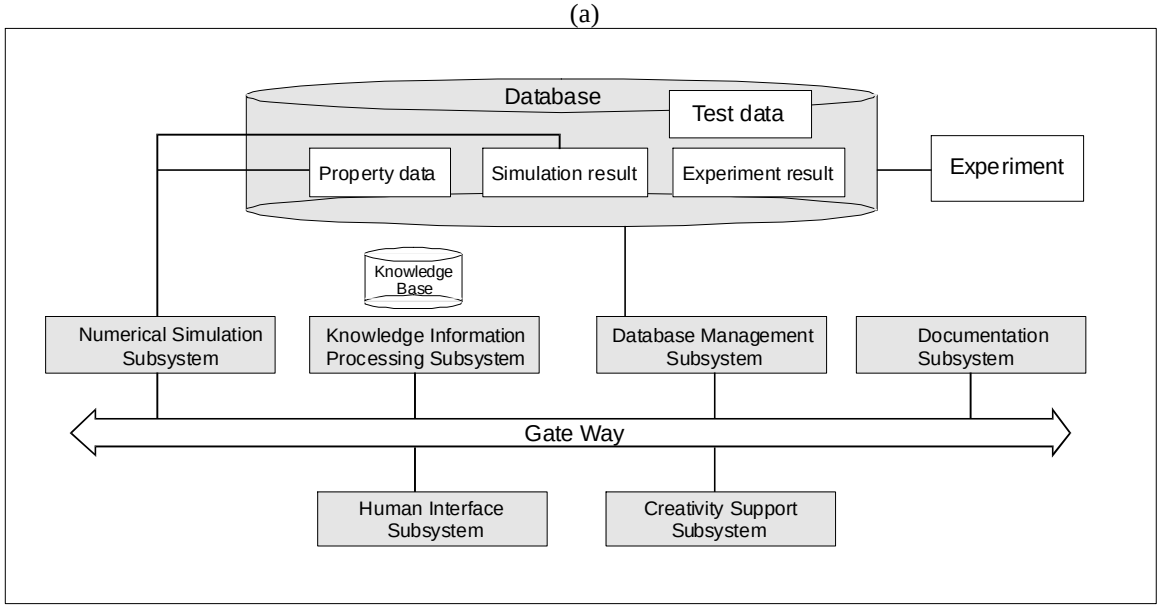
For each Design Tactic we need to assume a holistic view not to miss potential solutions, which is combined into a design scenario balanced by another holistic view on the designed material for each engineering system. For the electronic materials, only results of first principles calculations were used to select the candidate materials for electronic devices, so that density of states (DOS) and electron mobility data were mainly used and defects and processing data were not taken into account. For the structural materials design scenarios were made as a hybrid of two parts, namely, one scenario consisting of intrinsic properties derived by calculation as well as experiment, another scenario about structure-sensitive properties rewritten as a combination of qualitative causality and/or statistical correlation calculated from experimental data. As the latter structure-sensitive part has extremely rich semantics, due to the strongly correlated dynamics of defects under stress with severe thermal and chemical environments, each design scenario cannot be derived deductively and/or inductively from available data. The data were not enough to make a design scenario by deduction and/or induction. So design scenarios of the VEMD were derived abductively, in other words, predetermined, and experts were expected to explain and refine the predetermined design scenario partly according to their own scientific domains.

The VEMD project was thus conceived as a conventional research-oriented project on materials – mainly explaining what happened and why it happened. It was not a mission-driven strategic project "designing new materials", but the aim was to re-write the predetermined design scenario in terms of available data and models. Consequently, the priority of the project was shifted to writing original scientific papers on calculated data and models, rather than to developing materials by narrowing gaps between experimentally obtained data and the requirements of materials users. Iterative dialogue/communication between materials experts and system designers as users of materials are required ideally to reach a common design scenario, and the calculated data need to be used to converge into a set of design solutions. They should not be just a set of output digits from selected programs. Connections revealed by first principles calculations and molecular dynamics simulation *via* interatomic potentials were expected to explain atomic microscopic scale dynamics of materials, but the real dynamics of the microstructural evolutions were different. The results of the calculations could only be used to explain a particular observation in terms of models. The situation was the same for microstructural mesoscopic simulations, and also for macroscopic simulations, even if taking advantage of well developed models

such as the Finite Elements Method (FEM). They were not holistic but fragmented, or just a collection of parts. Each time careful parametric tuning of computing parameters was required to bridge the gaps among models and also between the results of the calculations and real world data. In practice VE (Virtual Experiment) followed real experiments and gave some explanations about the experimental data, but not the inverse. In short, one of the authors (S.I.) now concludes that the tasks of VEMD were not to design new materials, but to explain each fact in terms of established models and substantial calculated results, although it was a challenging feasibility study for dealing with complexities of materials and describe their dynamics as an interplay of data and models.



(a)



(b)

Fig. 18 Outline of the VEMD project (reproduced from [47]).

Due to the complexity of materials, the above statements are more or less true for all materials projects. Mismatches between the objectives and the obtained results are usually explained in terms of weakness of the theory, models, algorithms, and/or computational power, and also by the shortage/inaccuracy of experimental data. Historically all the projects were carried out properly and the number of original papers was in general sufficient to get good scores of evaluation, thanks to the very complexity of materials. New fascinating keywords were proposed, which are now "Big Data" and IoT coupled with the third AI wave "deep learning from data" toward "Industry 4.0" [50] in the cloud environment of "Collective Knowledge". A single flight generates 500 GB data for a jet engine of several hundreds of components made by different materials, and an automatic driving vehicle may produce the same or larger amounts of data on materials. Not only devices like X-ray diffractometers, but almost all engineering products and parts are monitored by numerous sensors. This is a characteristic of the IoT era, and the total amount of data easily reaches exabytes (1 EB = 10^{18} bytes). Data on what is happening are flooding. Now, even if everyone feels that huge obstacles still exist, we are expecting data-centered sciences and engineering, so that everyone is convinced that something will change.

How to bridge the gap between data producers and data users? This is an old question repeated many times, but data producers and users are changing. Digital devices and engineering systems are joining as users to such human professionals as members of manufacturing companies, scientists/engineers, data editors. In such a digital data era, it seems more realistic to quickly put a digital prototype product into the hands of potential customers, than thinking and spending a lot of time on hypothetical business forecasting and planning of attractive products. Here the system preparedness for further digital processing becomes crucial. The state of collective knowledge needs to be switched on, taking advantage of a minimum of viable products. This should be the first step in a build-measure-learn feedback loop, endorsed by such a system as the PAULING FILE, of high-quality data with traceability. Organization of serious users' groups is important to increase the quality of collective knowledge, even more than studies of mission-driven data projects (risk, climate change). Collaboration frameworks to share data can be established step by step through activities like the Research Data Alliance (RDA). Data projects focusing on particular methods (neutron cross section, NMR, X-ray diffraction, spectroscopy, beam technology, and so on) will be organized as exemplars of IoT, outputs from which are expected to be used to link associated data as in the case of the PAULING FILE project. Business models around data will emerge by harmonizing various initiatives, inspired by innovative projects, so-called Complex Design, or initiatives of researchers to explore new dimensions, such as David Baker [51]. The key point is to hybridize databases, *ab initio* calculations, and evolving algorithms, as in the case of Artificial Life. A crystallographic embryo is created in the hybridized cloud environment taking advantage of PAULING FILE data, and evolves there driven by *ab initio* calculations based on its boundary conditions and so on in accordance with multiscale modelings finally associated with materials requirements. It is a big challenge to fill the gaps between thermodynamics and quantum mechanics, but this encourages PAULING FILE customers to use phase diagrams to deal with time-, temperature- and pressure-dependent dynamics with intelligent formulations of each dynamic phenomenon. Self-organization of microstructural evolution can become realistic thanks to the recent development of image processing. A first descriptive analysis of the recorded data is followed by a diagnostic analysis of why it happened. The next steps consist in producing predictive analytics on what will happen next, and prescriptive analytics on what should be done. There is no shortage of data, but a lack of intelligence and knowledge on how to link key data and carry out clustering of data for these analytics. Powerful tools are needed to do so and the evaluation of structural stability of component materials may be performed by taking advantage, in the beginning, of various learning methods, which may be obtained by refining and reorganizing important tools and data developed through the PAULING FILE project. Then a digital eco-system is created as a kind of rice nursery, parenting new materials to emerge, where all models and data are categorized and encapsuled into a set of active agents. *Ad hoc* tactics as used in the VEMD project are regarded as one of many knowledge chunks, and many lessons from the PAULING FILE project can be developed as core guiding principles for the digital eco-system.

10. Conclusions

Several factors must be taken into consideration for the development of new materials and it is essential to build up a holistic view on inorganic substances by giving rapid access to different kinds of critically evaluated experimental data published in the world literature over the last 100 years. The PAULING FILE project was launched in 1993, and 26 years later this world-largest materials database contains over 500'000 database entries for inorganic crystalline solids, summarizing over 160'000 scientific publications. The linkage between the three different groups of data (crystal structures, phase diagrams, physical properties) is achieved by linking each database entry to one of the distinct *phases* defined based on the chemical system and the crystal structure.

With the help of several examples, we have shown that it is possible to gain a better view on inorganic

substances, effective or potential materials, by looking at large amounts of different data in an appropriate way. Data mining applied to the PAULING FILE provides good examples of holistic views, showing that "*the whole is greater than the sum of its parts*".

Note added in proof

The present manuscript was prepared a few years ago, for a handbook that was never published, and some of the numbers given here are consequently not up-to-date. All the electronic products that include PAULING FILE data have in the meantime been updated according to the plans. The platform "*World Materials*" was effectively launched in 2017, but under the final name *Materials Platform for Data Science*, MPDS (http://www.mpds.io/) and contains 64'927 phase diagrams, 405'040 crystal structures, and 765'578 physical property entries.

References

[1] P. Villars (Editor-in-chief), *PAULING FILE*; *http://www.paulingfile.com*/

[2] P. Villars, *J. Less-Common Met.* 110 (1985) 11-25.

[3] J. Rodgers, P. Villars (Eds.), *Proc. Workshop on Regularities, Classification and Prediction of Advanced Materials*, Como, April 13-15, 1992, J. Alloys Compd. 197 (1993) 127-307.

[4] N. Nishikawa, M. Nihei, S. Iwata, *Lect. Notes Comput. Sci.* 2858 (2003) 320-329.

[5] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, S. Iwata, *J. Alloys Compd.* 367 (2004) 293-297.

[6] *Inorganic Material Database (AtomWork)*, National Institute for Materials Science (NIMS), Japan; *http://www.crystdbnims.go.jp/index*

[7] P. Villars, K. Cenzual, F. Hulliger, H. Okamoto, J. Daams, K. Osaki, A. Prince, T. Massalski, S. Iwata, *PAULING FILE – Binaries Edition*, on CD-ROM, Materials Park (OH): ASM International, 2002.

[8] P. Villars (Editor-in-chief), F. Hulliger, H. Okamoto, K. Cenzual (Section editors), *SpringerMaterials, Inorganic Solid Phases*, Heidelberg: Springer, *http://www.Springerlink/SpringerMaterials*

[9] *World Materials*, MPDS, Switzerland (in preparation).

[10] E. Parthé, L.M. Gelato, *Acta Crystallogr. A* 40 (1984) 169-183.

[11] E. Parthé, L.M. Gelato, *Acta Crystallogr.* A 41 (1985) 142-151.

[12] L.M. Gelato, E. Parthé, *J. Appl. Crystallogr.* 20 (1987) 139-143.

[13] M. Berndt, Thesis, University of Bonn, 1994; updates by O. Shcherban, Scientific Consulting Company "Structure-Properties", Lviv.

[14] G.O. Brunner, D. Schwarzenbach, *Z. Kristallogr.* 133 (1971) 127-133.

[15] J.L.C. Daams, I.H.N. van Vucht, P. Villars, *J. Alloys Compd.* 182 (1992) 1-33.

[16] J.L.C. Daams, *Atomic Environments in Some Related Intermetallic Structure Types*, in J.H. Westbrook, R.L. Fleischer (Eds.), *Intermetallic Compounds, Vol. 1: Principles*, New York: John Wiley and Sons, 1994, pp. 363-383.

[17] K. Cenzual, M. Berndt, K. Brandenburg, V. Luong, E. Flack, P. Villars, *ESDD Software Package*, copyright: Japan Science and Technology Corporation, 2000; updates by O. Shcherban, Scientific Consulting Company "Structure-Properties", Lviv.

[18] T. Hahn (Ed.), *International Tables for Crystallography*, Vol. A, Dordrecht: D. Reidel, 1983 and more recent editions.

[19] P.P. De Wolff, N.V. Belov, E.F. Bertaut, M.J. Buerger, J.D.H. Donnay, W. Fischer, T. Hahn, V.A. Koptsik, A.L. Mackay, H. Wondratschek, A.J.C. Wilson, S.C. Abrahams, *Acta Crystallogr. A* 21 (1985) 278-280.

[20] P.P. Ewald, C. Hermann (Eds.), *Strukturbericht*, Leipzig: Akad. Verlagsgesellschaft M.B.H, 1931.

[21] W.B. Pearson, *Handbook of Lattice Spacings and Structure of Metals*, New York: Pergamon, 1967.

[22] E. Parthé, L. Gelato, B. Chabot, M. Penzo, K. Cenzual, R. Gladyshevskii, *Gmelin Handbook of Inorganic and Organometallic Chemistry, 8th Ed., TYPIX - Standardized Data and Crystal Chemical Characterization of Inorganic Structure Types*, 4 vols., Heidelberg: Springer, 1993, 1994.

[23] E. Parthé, K. Cenzual, R. Gladyshevskii, *J. Alloys Compd.* 197 (1993) 291-301.

[24] Y. LePage, *J. Appl. Crystallogr.* 21, 983-984 (1988).

[25] K. Cenzual, L.M. Gelato, M. Penzo, E. Parthé, *Acta Crystallogr. B* 47 (1991) 433-439.

[26] P. Villars, K. Cenzual (Eds.), *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds*, Materials Park (OH): ASM International, 2016, on DVD; *http://www.asminternational.org/AsmEnterprise/PCD*

[27] P.I. Kripyakevich, *Structure Types of Intermetallic Compounds*, Moscow: Nauka, 1977 (in Russian).

[28] *GetData Graph Digitizer*; http://www.getdata-graph-digitizer.com

[29] T.B. Massalski (Editor-in-chief), H. Okamoto, P.R. Subramanian, L. Kacprzak (Eds.), *Binary Alloy Phase Diagrams*, 2nd Ed., Materials Park: ASM International, 1990.

[30] G. Petzow, G. Effenberg (Eds. of vols. 1-8), *Ternary Alloys : A Comprehensive Compendium of Evaluated Constitutional Data and Phase Diagrams*, Weinheim: Wiley-VCH Verlag, 15 vols., 1988-1995.

[31] D.R. Lide (Editor-in-chief), *CRC Handbook of Chemistry and Physics*, Boca Raton (FL): CRC Press Inc., 1997-1998 and more recent editions.

[32] *Database of Zeolite Structures*, IZA Structure Commission; *http://www.iza-structure.org/databases/*

[33] H. Strunz, E.H. Nickel, *Strunz Mineralogical Tables*, 9th Edition, Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung (Nägele u. Obermiller), 2001.

[34] *IMA Database of Mineral Properties*; *http://www.Rruff.info/ima/*

[35] P. Villars, K. Cenzual, R. Gladyshevskii, *Handbook of Inorganic Substances*, Berlin: De Gruyter, 2016.

[36] P. Villars, L.D. Calvert, *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, 2nd Ed., vols. 1-4, Materials Park (OH): ASM International, 1991.

[37] P. Villars (Editor-in-chief), H. Okamoto, K. Cenzual (Section editors), *ASM Alloy Phase Diagram Database*, Materials Park (OH): ASM International, 2016; *http://www.asminternational.org/AsmEnterprise/APD*

[38] *PDF-4+*, International Centre for Diffraction Data (ICDD), Newtown Square (PA), 2016.

[39] P. Villars, K. Cenzual (Eds.), J. Daams, R. Gladyshevskii, O. Shcherban, V. Dubenskyy, V. Kuprysyuk, I. Savysyuk, R. Zaremba (Contributors to vol. 11), *Landolt-Börnstein, III-43, Crystal Structures of Inorganic Compounds*, 11 vols., Heidelberg: Springer, 2004-2012.

[40] *MedeA*, Materials Design Inc.; *http://www.materialsdesign.com/*

[41] P. Villars, K. Cenzual, M. Penzo, *Inorganic Substances Bibliography*, Berlin: De Gruyter, 2016 (e-book).

[42] P. Villars, K. Cenzual, J. Daams, Y. Chen, S. Iwata, *J. Alloys Compd.* 317-318 (2004) 167-175.

[43] P. Villars, J. Daams, Y. Shikata, K. Rajan, S. Iwata, *Chem. Met. Alloys* 1 (2008) 1-23; *http://www.chemetal-journal.org*

[44] P. Villars, J. Daams, Y. Shikata, Y. Chen, S. Iwata, *Chem. Met. Alloys* 1 (2008) 210-226; *http://www.chemetal-journal.org*

[45] P. Villars, S. Iwata, *Chem. Met. Alloys* 6 (2013) 81-108; *http://www.chemetal-journal.org*

[46] P. Villars, *Factors Governing Crystal Structures*, in J.H. Westbrook, R.L. Fleischer (Eds.), *Intermetallic Compounds, Principles and Practice, Vol. 1: Principles*, New York: John Wiley and Sons, 1994, pp. 227-275.

[47] N. Nishikawa, C. Nagano, H. Koike, *Integration of Virtual Experiment Technology for Materials Design*, in S. Nishijima, S. Iwata (Eds.), *Computerization and Networking of Material Databases*, ASTM STP 1311, West Conshohocken (PA): ASTM, 1997, pp. 21-27.

[48] J.A. Robinson, *J. Assoc. Comput. Mach.* 12(1) (1965) 23-41.

[49] A. Masahiko, *Towards a Comparative Institutional Analysis*, Cambridge (MA): MIT Press, 2001.

[50] *Plattform Industrie4.0*, Bundesministerium für Wirtschaft und Energie & Bundesministerium für Bildung und Forschung; *http://www.plattform-i40.de*

[51] R.F. Service, *Science* 353 (2016) 338-341.