Materials Platform for Data Science: A 10 Years Success Story

Evgeny Blokhin[1, 2, 3]
[1]Tilde MI, Straßmannstraße 25, 10249, Berlin, Germany
[2]MPDS, Sepapaja 6, 15551, Tallinn, Estonia
[3]Absolidix, 36FQ+RR, Dubai, UAE

eb@tilde.pro

In this contribution I will elaborate on the technology, acceptance, and growth of the online edition of the Pauling File project, called Materials Platform for Data Science (MPDS), being developed since 2015. Today MPDS is a well-established online platform, delivering the curated materials data, both open-access and commercially. In 2025, there are about 20 thousand monthly active users, who are materials scientists from all over the world, and about 10 thousand registered accounts. These numbers are growing steadily.

I will start from the challenges ahead of the materials community 10 years ago, the greatest anachronisms of the scientific communication medium, still actual today, and the published data excerption and curation, funded commercially. The Pauling File project materials data layout will be presented, namely, the distinct phase concept, as formulated in 1993 by Dr. Villars and Prof. Iwata (and later implemented by Materials Project), and the Materials Genome trademark, as registered in the USA in 1990-es by Dr. Villars and Prof. Liu. Importantly, we link together the crystalline structures, physical properties, and phase diagrams via the distinct phase concept, which makes the holistic view on the materials science from the point of view of computer science. In 2019, Dr. Villars was acknowledged by the NIMS Award for realization of this concept in the Pauling File.

I will briefly review the offline vs. online software paradigms, mention implementation details of the MPDS, and elaborate why the online product must be fast. Then the relational, document, and graph databases, semantic technology stack, neuro-semantic approach, and strong AI with respect to the new materials design will be in the center of my attention. I will also discuss human-readable vs. machine-readable approaches and how they converge to the symbolic AI (cf. ChatGPT o3 as of December 2024). The role of the open source, open access, and open standards will be also emphasized, not to forget why exactly the Python programming language is so popular, also being the most accepted tool in the materials informatics community.

Then I will discuss the virtual materials design laboratories (cf. lights-out manufacturing) and the fact they must be based on the fundamental very rigid systematic framework. While implementing the MPDS, as a by-product, we have discovered several quantitative trends via the systematic *ab initio* simulations and machine learning [1, 2], serving as a perfect example. I will also share my (very limited) experience on synthesizing and productization of the virtual materials in such companies as BASF SE (Germany) and Hitachi Energy (Sweden).

To summarize, my success recipes are: on-demand cloud computing, AiiDA computational workflows and graphs [3], universal access to materials databases called Optimade [4], materials ontologies lingua franca [5], CALPHAD vs. multiscale modeling, and, finally, a synergy of data-driven and physics-driven modeling.

References

[1] Blokhin, Villars. arXiv, **2018**, 1806.03553
[2] Caputo, Villars, Tekin, et al. J. Alloys and Compounds, **2024**, 172638
[3] Pizzi, Cepellotti, Sabatini, et al. Comp Mat Sci, **2016**, 218
[4] Evans, Bergsma, Merkys, et al. Digital Discovery, **2024**, 3, 1509
[5] De Baas, Del Nostro, Friis, et al. IEEE Access, **2023**, 120372

Figures

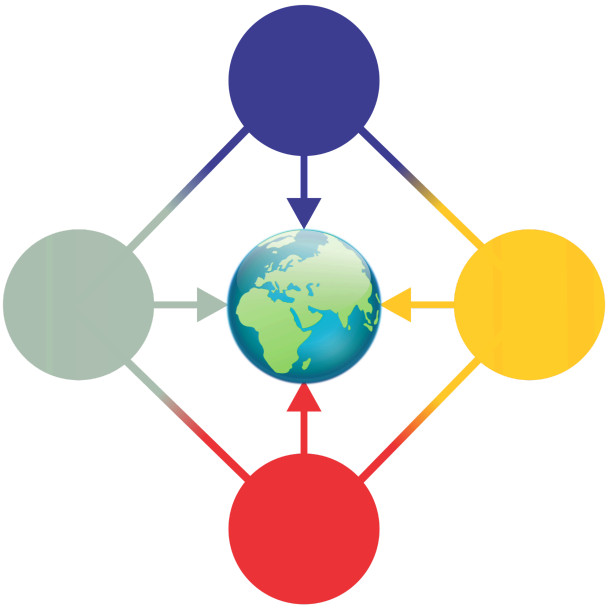

Figure 1. The MPDS logo: phase diagrams, crystal structures, and physical properties, interlinked from the world's published literature.