The Novel Materials Discovery (NoMaD) Laboratory

Matthias Scheffler (Fritz Haber Institute of the Max-Planck Society, Berlin, project coordinator), Arndt Bode (Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Garching), Claudia Draxl (Humboldt-Universität, Berlin), Daan Frenkel (University of Cambridge), Stefan Heinzel (Rechenzentrum Garching of the Max Planck Society), Francesc Illas (University of Barcelona), Kimmo Koski (CSC – IT Center for Scientific Computing, Helsinki), Jose Maria Cela (Barcelona Supercomputing Center (BSC)), Risto Nieminen (Aalto University, Helsinki), Angel Rubio (Max Planck Institute MPSD, Hamburg and UPV/EHU, San Sebastián), Kristian Thygesen (Technical University of Denmark, Lyngby), Alessandro De Vita (King's College London)
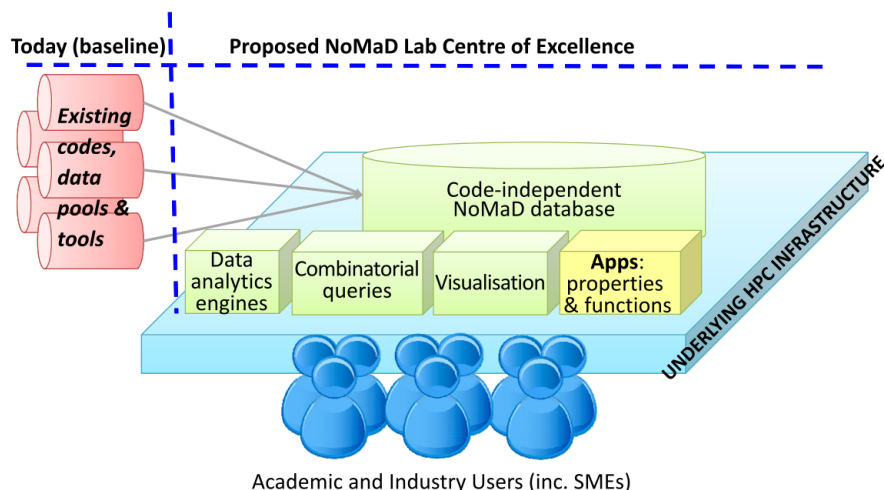
Figure: From various open-access repositories (present) to the Laboratory (this proposal).

A *code-independent* data base and properly stored data enable novel combinatorial queries, big-data analytics, and more.

We will be surprized what people will do with the data once they are made available in a common format.

Abstract

Essentially every new commercial product, be they smart phones, solar cells, batteries, transport technology, artificial hips, etc., depends on improved or even novel materials. Computational materials science is increasingly influential as a method to identify such critical materials for both R&D. Enormous amounts of data, precious but heterogeneous and difficult to access or utilise, are already stored in repositories scattered across Europe. The NoMaD CoE will open new HPC opportunities by enabling access to this data and delivering powerful new tools to search, retrieve and manage it.

NoMaD will foster sharing of all relevant data, building on the unique CECAM, Psi-k and ETSF communities, putting Europe ahead of materials science in other continents. Unprecedented, already initialised networking with researchers, with industry, with students and with other stakeholders will guarantee relevance and end-user value. NoMaD will become a crucial tool for atomistic simulations and multi-scale modelling in the physical, materials, and quantum-chemical sciences. This field is characterised by a healthy but heterogeneous eco-system of many different codes that are used at all HPC centres worldwide, with millions of CPU hours spent every day, some of them at petascale performance. NoMaD will integrate the leading codes and make their results comparable by converting (and compressing) existing inputs and outputs into a common format, thus making these valuable data accessible to academia and industry.

NoMaD will develop "big-data analytics" for materials science. This will require novel algorithms, e.g., for statistical learning based on the created materials encyclopaedia, offering complex searches and novel visualisations. These challenges exploit the essential resources of our HPC partners. Without the infrastructure and services provided by the NoMaD CoE, much of the information created with the above mentioned petascale (towards exascale) computations would be wasted.