Towards a Common Format for Computational Materials Science Data

CECAM HQ, Lausanne.

First part, discussion and decisions: January 25 to 27, 2016
Second part, implementation: January 27 to February 5, 2016

State of art in the field

The development of modern commercial products – be it from the health, environment, clean energy, heavy industry, information or communication technology sector – depends strongly on the development and design of new and improved materials. However, identifying the best material or designing a novel and improved material for a specific task/application is a significant challenge. Of key importance are the characteristics of the materials at the atomic and molecular levels, which determine their properties and behaviors at the macroscopic scale. To aid and guide this search, computational materials science employs complex methods and computing algorithms ('codes') to investigate, characterize and predict material properties. Fueled by the "Materials Genome Initiative for Global Competitiveness" [1], announced by President Obama in June 2011, these computational techniques are increasingly and successfully employed also for the "high-throughput screening" of materials [2-4]. In conjunction with techniques from big-data analytics and machine learning, such an approach enables to scan many thousands of compositions for the material with the best-suited properties to predict trends, and to identify potentially (technologically) important candidates [5,6]. So far, however, different technologies and frameworks developed in this context have addressed only very specific aspects, e.g. by focusing on properties relevant to one particular application and/or by supporting only one or very few electronic structure codes.

In practice, this means that computational material scientists produce a huge amount of materials data on their local workstations, computer clusters, and supercomputers using a variety of computer codes that are most commonly developed by European research groups. Though being extremely valuable, this information is mostly unavailable to the community, since most of the data are stored locally or even deleted right away. But even if they are available, a re-use and re-purposing would not be straightforward, given that different codes often use very different file formats and conventions to store the same physical data. Enabling sharing and comparing such data is thus a pressing issue that needs to be addressed to advance this field, as exemplified by multiple European initiatives. For instance, the European Center of Excellence for Novel Materials Discovery (NOMAD-CoE) [7] aims at establishing a unified, code-independent data format, to which the raw data calculated by different electronic structure codes can be converted, so that big-data analytic techniques can then be exploited to obtain unprecedented insight from vast amounts of calculations. In a similar spirit, the Center of Excellence E-CAM, which was recently established by CECAM to build an e-infrastructure for software, training and consultancy in simulation and modeling, is committed to actively support the development and adoption of software libraries and standards within the electronic structure community. One measure aiming at this is CECAM's Electronic Structure Library (ESL) initiative [8], which drives establishing an Electronic Structure Common Data Format (ESCDF)

Strategy of the workshop

There are some differences in the goals of data representations in the two actions:

- The ESCDF provides a standardized data format and an API every code can use. Material science code developers profit from centralizing implementations like efficient parallel IO and hiding file format specific issues. At the same time this allows a certain amount of interchangeability of data, for example for post-processing tools. The data itself is not altered, so that checkpointing and restarting functionalities can be granted without additional data loss.
- The NOMAD-CoE aims at making also the data itself comparable, which involves data transformations ranging from simple unit conversions up to normalizations based on reference calculations and analytics tools.

As both initiatives target the whole electronic structure code community, they are based on the same concepts and codes, thus have a large common ground. So, for a maximal mutual benefit, they jointly organized the present workshop.

Many attempts of standardization fail because initiatives are too small to reach a critical mass or try to impose their solution to a community for which it is not profitable to adapt to it, possibly also because due to shortcomings of the standard. This is avoided by involving all major codes right from the beginning. The workshop attended key experts of ABINIT, BigDFT, CASTEP, CP2k, CPMD, Dmol3, ESPResSo, Exciting, FHI-aims, FHI-98, FLEUR, KKR, LMTO, Octopus, Quantum Espresso, SIESTA, Turbomole, VASP, Wien2k. Each code expresses key magnitudes like wave functions, operators, and density matrices as linear combination of basis functions. There are various types of such “basis” established, some of them of very different character. Some codes describe all electrons in this basis, others use a simplified description of the core electrons. This results on one side in basically different data representation, on the other side also comparing certain magnitudes like energies is not trivial. This is addressed within the NOMAD-CoE, with funding currently granted until October 2018, by developing a conversion layer for normalizing this data and analytics tools for error quantification.

The workshop provided a unique platform to discuss and decide the fundamental paradigms needed to establish a common framework that supports several different electronic structure and force field codes and that is prepared to interface with the newly emerging field of data-driven material discovery in the European research landscape. In this view, a common purpose of the NOMAD-CoE and the CECAM-supported ESL is to integrate the computed results from leading electronic structure codes. Defining a common code-independent representation for all relevant quantities, e.g., structure, energy, electronic wave functions, trajectories of the atoms, etc., is challenging, as the codes differ, for example, in their choice of basis sets and treatment of the core electrons (e.g. usage of pseudopotentials). To tackle these challenges from a technological point of view, an envisioned strategy is to build on the experience gained during previous community projects with somewhat narrower focus but with similar philosophy. For instance, one of the most consistent and successful efforts was the development of the (NetCDF [9] based) ETSF file format [10] by the ETSF [11]. Similar standardization efforts are currently under way within the EUSpec [12] network. In this context, it is also planned to extend and modify the ETSF file format, in particular for greater flexibility for parallel I/O.

The key players in the electronic-structure and force-field code development were thus brought together, in order to discuss and implement the aforementioned code-independent representation of materials science data. The workshop was divided into two parts: a 2.5 days discussion on the file format specifications, followed by an 8.5 days coding effort.

Result of the discussion

In the first part, each session was followed by an extended discussion, which led to actual guidelines for the future common-format storage. In the following, we list the topic of the sessions and the decisions taken in the respective discussions:

- *A common energy zero for total energies.* To make (total) energies stemming from different codes comparable, it is necessary to define a reference energy scale. To achieve this goal, it was concluded that a simple, pragmatic computational prescription viable for all codes is necessary. To bridge the gap between periodic and non-periodic codes both free atoms and simple bulk systems shall be used as reference systems.
- *Compact representation of scalar fields: Density, Wavefunction, xc potentials, etc.* The comparison of scalar fields across methodologies and codes requires to translate the internal, code and basis set specific representation of these fields into a common format. For such a representation, it was concluded that an all-electron formalism is desirable, since it allows to evaluate additional properties such as electric field gradients and NMR shifts. Which specific all-electron basis set (Numeric Atomic Orbitals, Gaussians, or APW+lo /FLAPW type basis sets) is best suited for this purpose needs to be evaluated in detail.
- *Quantities related to excited-state calculations.* Advanced many-body perturbation theory (MBPT) calculations (GW, BSE,…) currently output only few properties (spectra, self-energies, etc.) that need to be parsed and stored. To facilitate the analysis of this kind of calculations, it is essential to develop and store a detailed classification of all approximations used in the MBPT calculation in the metadata, given that many different numerical formalisms are implemented in different MBPT codes.
- *Molecular dynamics related common format.* The fundamental information generated during molecular mechanics calculations are the geometric configurations and trajectories. Accordingly, these are also the most useful quantities to store. However, trajectories from specific approaches (Metadynamics, Replica Exchange, …) have to be handled with care. It is thus crucial to store respective metadata and settings. If possible, the original submission scripts should be retained as reference.
- *Metadata for a code independent format.* For the properties desirable for the metadata ontology, it was concluded that both human and machine readable formats are needed. An unambiguous conversion (“translation”) script is required for this purpose. Since multiple ontologies are currently under development (NOMAD, TCOD, ESL), discussion among these different communities should be encouraged to establish a common language/wording. Besides having a standardized central definition (reference), customization options for local users are desirable.
- *Establishing error bars and uncertainties.* Clearly, quantifying the errors and uncertainties of the data included in computational materials’ databases is an essential step to make this data useful at all. Challenges in this field arise, since the errors are code, property, and material specific. Also, the dependence of different errors on each other needs to be taken into account. A first step in this direction is to establish unique identifiers for structures through “similarity recognition”. Furthermore, a systematic investigation of numerical errors is required across codes for both simple and complex properties, which also requires a clear definition of

errors/deviances, e.g., for continuous functions. With respect to errors arising from the use of approximated xc-functionals, (more) test sets are required as a reliable, high-level reference. In this context, using experimental benchmarks can be tricky, since they hardly allow for error analysis.

- *The Electronic Structure Common Data Format (ESCDF).* A standardized data format for electronic structure calculations must provide a framework for saving and reading data without enforcing a specific physical representation, while providing means to store different types of data. This can be achieved with self-describing formats like HDF5 or NetCDF, which are extendable and allow the inclusion of metadata needed to interpret them. The first version of the ESCDF must include specifications to read/write the following type of data: geometry/structure of the system, basis sets, densities, potentials, and wavefunctions. The associated software library and corresponding API will focus on flexibility, extensibility, and performance in order to maximize its usefulness and adoption by the community of code developers.
-

The future

There are some differences in the goals of data representations in those two actions:

The ESCDF provides a standardized data format and an API every code can use. Material science code developers profit from centralizing implementations like efficient parallel IO and hiding file format specific issues. At the same time this allows a certain amount of interchangeability of data, for example for post-processing tools. The data itself is not altered, so that checkpointing and restarting functionalities can be granted without additional data loss.

The NOMAD-CoE aims at making also the data itself comparable, which involves data transformations ranging from simple unit conversions up to normalizations based on reference calculations and analytics tools.

As both initiatives target the whole electronic structure code community, they are based on the same concepts and codes, thus have a large common ground. So, for a maximal mutual benefit, they jointly organized the present workshop.

Many attempts of standardization fail because initiatives are too small to reach a critical mass or try to impose their solution to a community for which it is not profitable to adapt to it, possibly also because due to shortcomings of the standard. This is avoided by involving all major codes right from the beginning. The workshop attended representatives of of the most important codes (see above). Each code expresses key quantities like wave functions, operators, and density matrices in terms of basis functions. There are various types of such “basis” established, some of them of very different character. Some codes describe all electrons in this basis, others use a simplified description of the core electrons. This results on one side in basically different data representation, on the other side also comparing certain magnitudes like energies is not trivial. This is addressed within the NOMAD-CoE, with funding currently granted until October 2018, by developing a conversion layer for normalizing this data and analytics tools for error quantification.

The conclusions of the discussion, as reported above, will be implemented in the data format of the NOMAD Archive. A key to success, here, may be that NOMAD is not “ imposing” the

common format to code developers, but rather to convert existing output into the common format.

However, a direct standardization of the code outputs would greatly facilitate the maintenance of useful big –data storage. To this end, an API definition, which is directly usable for all represented codes, has been developed in the second part of the workshop, and a library implementation will follow.

Funding

Three Centres of Excellence, in the framework of the Horizon-2020 call for e-infrastructure, have been funded in the materials science field, with several millions euros and therefore several academic positions: Materials design at the eXascale (MaX), E-CAM, and Novel Materials Discovery (NOMAD). With different focuses, all three centers share the goal of further establish high-performance computation for novel materials design and discovery.

Such initiatives will benefit of a second funding period, before being able to be fully self-funded, in particular by offering services to the industrial sector. They can further be supported by providing computational resources, e.g. by the PRACE initiative as it happens already now.

The NOMAD Project has additional demands on storage and server infrastructure, as besides the raw data also data generated by normalization and analytics tools needs to be stored and provided to the scientific community.

Additional information and resources:

Program and Abstracts:
http://www.cecam.org/workshop-2-1290.html
List of participants:
http://www.cecam.org/workshop-1-1290.html
Pictures of the Workshop:
http://th.fhi-berlin.mpg.de/th/photoalbum/FCMSD2016/

Bibliography

[1] Materials Genome Initiative for Global Competitiveness (President Obama, June 2011): http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf
[2] B. C. Wood and N. Marzari, Phys. Rev. Lett. 103, 185901 (2009). DOI:10.1103/PhysRevLett.103.185901
[3] S. Curtarolo, et al., Nat. Mat. 12, 191 (2013). DOI: 10.1038/nmat3568
[4] S. Kang, et al., Nano Lett. 14 , 1016 (2014). DOI: 10.1021/nl404557w
[5] Y. Ritov, et al., Statistical Science 29, 619 (2014). DOI:10.1214/14-sts483
[6] L. M. Ghiringhelli, et al., Phys. Rev. Lett. 114, 105503 (2015). DOI: 10.1103/PhysRevLett.114.105503
[7] European Center of Excellence for Novel Materials Discovery (NOMAD-CoE), http://nomad-coe.eu
[8] The Electronic Structure Library, http://esl.cecam.org
[9] R. K. Rew and G. P. Davis, IEEE Computer Graphics and Applications 10, 76 (1990). DOI: 10.1109/38.56302

[10] ETSF File Format Standardization Project: http://www.etsf.eu/fileformats
[11] European Theoretical Spectroscopy Facility: http://www.etsf.eu
[12] COST Action MP1306: http://euspec.eu/