Discussion note

Exchange between Databases

Please give your opinion on the following (sometimes provocative) statements

State of the Art and outline of the Discussion

Modelling in each of the subfields electronic/atomistic/mesoscopic/continuum has already resulted in many databases, with each their own internal structure and data standards driven by the individual community. In these databases, both simulated and experimental data are gathered. Some individual databases are now well established while others have not yet reached a mature, strongly rooted status. One of the reasons for this is the lack of one (or at most a few only) de facto standard for data (including models, input and output) which makes it hard to maintain such databases and leverage enough momentum for their use. Nevertheless, the underpinning believe is, they should remain as the subfield has designed them to satisfy their purpose. Moreover, each field is still organised as an isolated data repository, largely isolated from other fields, gathering data remains largely unavailable to the other subfields and these subfields are until now not communicating enough. Scientific results are published in written form. Some publishers do allow a certain amount of supporting data. Nevertheless, the raw data itself should be available and should get a digital object identifier, allowing further use and interpretations.

However, the data is not readily available to the other subfields. Furthermore, the different subfields are until now not communicating enough.

The task the EMMC has set itself is to co-ordinate these subfields. Several activities will be discussed related to gathering information on already available data to make them visible to all stakeholders in all subfields.
The database infrastructure can be compared to roads, water supply, electricity and internet connections. The different countries in the EU handle their infrastructure themselves. Similarly to that, it will be an issue of discussion how to handle the maintenance of the future database. But it remains absolutely clear that the authors hold the rights to their data and models (unless licensed to third parties).

The EMMC should elaborate an exchange methodology should be elaborated to interpret and share the information in the different subfields so that a structured exchange is generated and supported. One of the goals is to establish methodologies to collate the data, interface them and exchange information. This should alleviate the current incompatibility or incompleteness of the set of distributed databases:

The long term goal is to establish an exchange/interpreter of databases that for a specific material gives e.g. the electronic structure, atomic nuclei positions, particle/grain and continuum behaviour. This goal will pave the way to establish a holistic modelling approach and thus produce realistic predictions of materials behaviour under real world conditions.

Neither the commercial use of databases nor individual projects to optimise the running and storing output of a particular code are dealt with in this discussion note. The creation of database like "If I want a material with this conductivity and this weight, I need to take this composition" will be left to commercial third parties.

Underlying concepts

There is a distinction between *raw data* and *interpretations thereof*. The latter are often subject to commercial considerations, the former are mostly free. The latter is the result of data mining or data processing techniques (and reflects current interpretations, which might change over time). For example, in the electronic modelling world most codes already include standardised homogenisation techniques (called coarse-graining), while in the mesoscopic and continuum world this is still a vast topic under development.

Design of materials is based on *high-throughput simulations* where a domain of possible parameters is identified, in which a refined search is to be conducted. As each model simulation can take up several weeks or months, this is a time-consuming activity. This initial identification of the search domain, which is a critical factor for a successful design of new materials, could be much faster if raw model output data would be available. This will enable for example, re-analysis of this data, i.e., interpretation, possibly even of raw data from multiple sources, to infer a more optimised initial search domain. It is recognised that electronic models might be able to recreate their data in acceptable times in the near future, while other models can do this already, but it is still believed that the availability of linked data across the electronic, atomistic, mesoscopic and continuum models will be in big industrial demand in the coming 20 years.

Motivation of stakeholders:
Motivation for (academic) modellers and experimentalists to make their data available will be that their data get cited more often than only once in their own scientific paper (see also below the remarks on decoupling data from interpretation) and the modellers and experimentalists can thus attract more attention to their work. Citations will show the quality of the data! Early adopters uploading their data will become visible and this creation of awareness could results in a positive feedback loop resulting in the recognition of the quality of the data. Other advantages include shortening the time to publish, as the data can be published as soon as available.
Industry and software owners are interested in public data to compare their own data with and use them for their own benefit (e.g., expanding their own databases).

Layers of specification and time line of EMMC activities
There are many layers in the specifications of a database. This chapter deals with the end user specs (the SWO mostly and of course in the plenary meeting). The second level is the scientific requirements and software requirements and these are discussed below (and will be thoroughly discuses in the GEN and plenary meetings). The third layer "technical specifications of the database"

based on these requirements will be discussed in coming EMMC activities (mainly through the coordination of the Material Modelling Market Place (MMP) EMMC working group).

Chapter 1:End-user requirements on Materials Modelling Database

Activity 1: Preparation of "Sharing across modelling fields"

The overarching goal of the EMMC in this domain is to establish leadership in "sharing data across modelling fields"

Within EMMC in 2015:

- The EMMC could start this activity with an investigation of what databases already exist in each subfield.
- Workshops with experts who linked different "modules" should be held to hear their lessons learnt from existing databases. (E.g. experts who developed STEP connecting databases on the different parts of an airplane).
- Methodologies for exchange of information should be agreed. (This is related to the work done in the WG Coupling and Linking of models and the Open Simulation Platform, who resp. elaborate the physics of the information transferred and the IT background and standards.)This will prepare an exchange mechanism, rather than a new database. It will map substructures into a common, invisible layer. This interpreter would interpret the vocabulary of the different fields, but not necessarily change the vocabulary in use within a particular field.
- The EMMC needs to develop a structure which will reward participating scientists and allows new business models for companies to add the necessary driving force and sustain their motivation. The open data exchange needs to become part of the scientific culture.
- There should be a 10 year development plan with a planning for when to bring the industry on board.

The results should be sharing and understanding between subfields based on an accepted standards and protocols. The EMMC will thus provide added value and thus leadership!
Without this leader ship, there is no common thread to weave all these fields together.

Activity 2: Database of simulations

The approach is an incremental one, where small realistic steps are taken at a time.

The current state is that a lot of simulation data is stored by many stakeholders.
This Activity 2 addresses the technical task of cataloguing what exists, without the wish to restructure any subfield!
Although many databases are around, there is still a need to enlarge the set of materials simulated. E.g. EU-LEIT projects generate data that is for a large part confidential, but that

there will be modelling data that the owners are willing to share (this will remain a free choice for participants). Also outside EU projects a lot of simulations are performed.
Most modelling is done by academics and the strong believe is that there is a wealth of data that could even be made public after an initial short publication period.
There is conflicting opinions on the number of modellers who would want to participate. Some say we will be snowed under, others state nobody will share anything.
In any case the first set of activities will be limited to proprietary and open source codes licensed on the internet to third parties.
If need be this can be restricted to codes that have a support and maintenance service.
If need be it can be restricted to simulations that are documented in textual articles.

The idea is thus NOT to do a peer review, but to this should implicitly be done by the users of the data. The user should decide whether they want to use this simulation or not, like in the following scenario:

> *If user1 uploads some results to the database front end user1 will still be owner of this data, Suppose user2 is interested in similar problems so *she* does a very similar but independent simulation using perhaps slightly different models or applications. Nevertheless when she compares to user1 results, she finds agreement! Then she will upload this data and tag it as in agreement to user1. Then the "confidence/acceptance level" of both data sets may increase. I know of many good papers that have been published even in Science or nature having results that no one could reproduce yet (though everything is scientifically correct), simply because of possible misinterpretation or analysis step not known to any one before. So having the possibility to access raw data will help correlate different simulations or experiments to enhance previous interpretations and even generate more correct results by data mining alone!*

It may be very difficult to bring existing data into a new standard form. New data on the other hand may be readily, either at the creation stage, or shortly after, be casted into standard forms enabling their timely and rapid integration into the new databases. It is reasonable therefore to expect that the EMMC databases will, at least initially target new data primarily, with older, available, and valuable data migrated slowly, and continuously into the new databases.

Everybody who uses the data will need to cite the (specific entry in) the database.
Motivation for academic modellers to list their data will be that their data get cited more often than only once in their own scientific paper (see also below the remarks on decoupling data from interpretation) and the modellers can thus attract more attention to their work. And this is from the minute their data is available on the database. Furthermore, the data will be evaluated by users who have a genuine interest in it. Software owners will not populate the database, this will be done by the users, but the SWO will be very interested to see how their code is used.

The database will have a legal disclaimer for non-liability when the data is used and the warning that if modellers want to execute the runs themselves they should have a valid licence for the respective software code.
Whether the data can be used for benchmarking is to be decided by the code and data owners.

Step 1: Database of simulations (metadata only: code and input)

Within EMMC in 2015:

First the simulation data is to be identified and a central directory could register where the output data can be found and how it may be accessed. This activity is for licenced codes only and all that is thus needed to be registered is the code name, its version, and the owner or author. Furthermore the input data will be stored and where the output data can be obtained and under what conditions as the uploader chooses the licence for the output. Parameters like low or high accuracy runs are given in the input file, but could be clearly presented).The design of this database is rather simple as it only lists the code name and version and the input data (which is defined by the code) and who owns the raw output data and how it can be obtained. Also scientific articles which describe the simulations should be hyperlinked.
The data and IPR remain with the data owners (= the people who did the simulation). Also the input can be IP protected by the modeller (e.g. the morphology of the molecule, if patented).
The uploader has to decide how much detail is given on the material simulated. If the user is interested the user can go and talk to the simulation owner. There will be no output available. It resides with the simulator! The simulator determines under what conditions the output is released.
The database in this first phase would thus be a distributed database (called 'green' repository).
The database itself **need not be protected**.
The code owners should be involved in the **design** of this simple database also because the user interface could write the input files.

The meta-database will be kept by EMMC Market Modelling Place and is believed not to need too much effort. Surely, the catalogue of materials needs to be organised. But catalogue systems are already existing!
The data can be linked through a DOI or a unique link which does not change with time.
Quality and curation are not an issue for this data. Maintenance support is limited and can be provide over existing EC projects as networking activities. Note that the linked databases with output data will all be different, but this will not cause any problems, other than for the search engine.

Step 2 Gold Databases of simulations (input, code, raw output data)

In eventual CSA and/or and dedicated EU projects:

The basic idea is to gather raw output data from the simulations above, so called "gold" repositories. The design of this database will be the same as the database above with the addition of the raw output data. The structure of the last addition is determined by the code and these databases need to be created for specific niches and there will thus be many of them. As this will need a huge storage space this activity is delayed until it can be funded over dedicated projects. An eventual CSA could help establish a road map for these application fields

The database projects should have a strategy for the upkeep of the databases after the database projects end.

Databases for different models should be linked for multi-modelling purposes.

It should be decided whether the code-owners will gather data for simulations with their code, or whether there should be a central organisation.

Recommendation to the EC

New projects should be required by EC to produce data whenever feasible in common universal/unified data formats, since this will enhance the efficiency and lower maintenance costs of the database. Such CUDS will be elaborated in the OSP and MMP EMMC activities.

Timeline: 2017 at the earliest

Step 3: Database of Materials properties

In eventual CSA:

Adding materials properties derived from the raw simulation data sets listing the homogenisation techniques (and the physics/chemistry involved) by which they are derived.

Commercial aspects of these processed data should be discussed and the consequences for the database.

Timeline: 2017 at the earliest

Activity 3: Website facilitating beta testing of codes under development

Often the software developed by PhD students in EU and other projects disappear after the project is over. The codes remain undocumented and can not be used by third parties.

It is often tedious to install new codes on computers, while the developers would like to see their codes beta tested. It is thus proposed to make a website where anybody can upload academic codes in development. (See also activities to bring these codes to the industry).

Ofcourse also the code itself should be stored and documented so that it can be tested.

Support by the code owners should be provided.

Benchmarking aspects have to be discussed. And it should be discussed whether this is for academic and commercial codes.

Alternatively, virtual machines could be stored where the environment and the code itself are guaranteed to work even after drastic changes in hardware or software environments.

Recommendation to the EC

Codes developed in EU projects should be accompanied by wrappers to standard data formats, or have built-in support for the CUDataStandard. This will enable a universal interface to all new codes, and will ease their installation and use.

Timeline: earliest 2018-2019
CSA or other EU projects

Activity 4: Characterisation data (raw data)

This activity is concerned with the creation of databases with experimental data. Also in this field there are the concepts of raw data and of processed material properties. Raw data should be the subject of this first step. (E.g. the interest in the underlying displacement data might attract more interest than the deducted strain value.) Raw data need also be stored in connection with scientific publications by linking the data to the paper and vice versa.
These databases could be initially distributed, but in a later stage be supported by the characterisation machine vendor.
Up till now experimentalists measure what they deem is interesting or what has been agreed on by the community over the last tens of years. Modellers, if given the chance, might want to provide guidance on what measurements they need to validate their constitutive equations (instead of tuning their model to what exists). This interfacing between experimentalists and modellers does not yet exist except maybe in thermodynamics. This process will be to the mutual benefit of both modellers and experimentalist as modellers can validate their models, while experimentalist can get a physics/chemistry based explanation of phenomena.

STEP 1 Interface modellers-experimentalists

Within EMMC in 2015

a) Inventory of existing characterisation database for interest for validation of models (database types and content)
b) Modellers define guidance for data needed for validation of their models
c) This is discussed with characterisation experts and database IT experts
d) The EMMC could establish a database and define the necessary metadata on raw characterisation data observed by a specific characterisation technique. The EMMC characterisation database should list who owns the raw data and how it can be obtained. As the structure of experimental data is not obvious, the design of this database is to be determined per characterisation technique in co-operation with the Characterisation Council and IT specialists. It is of utmost importance to negotiate the necessary metadata to prevent raw data from losing its usefulness for both, modellers and experimentalists. In the near future, uploads to the data repository need to be reviewed by the EMMC and the metadata discussed between modellers and experimentalists to agree on the metadata. Quality and

curation issues are not an issue for this data. Maintenance support is limited and can be provide over existing EC projects as networking activities.

STEP 2 Elaboration of the characterisation database

In eventual CSA

If a CSA is launched they can take over the elaboration and maintenance and the CSA should have a strategy for the upkeep of the db after the CSA ends.

d) Expiration dates for confidentiality of data could be discussed.

e) definition of the database, the metadata and the interface to the models for new cutting edge characterisation techniques

Activity 5: Interpretations of raw characterisation data

In eventual CSA

Linking the (different) interpretations to the characterization database as they occur in scientific papers. (The data can be very good, while the interpretations might change over time). In this way the quality of the experimental data and their interpretation is decoupled.

Activity 6: Validation of constitutive equations

In eventual EU Constitutive Equation validation projects

Note that *"model"* means "the equation with which the behaviour of the entity is described". Phenomenological relations are often also called models, but we prefer to call them *"constitutive equations".*

SWO of materials models are not in great demand of data for the validation of the equations used in their models. The users are though in great demand of validation of their constitutive equations (tuning of parameters to experimental data).

Last but not least…….

Activity 7: Examples for dissemination

Examples on how these databases function on three new materials in order to demonstrate the benefit of these public databases spanning phenomena at all length and time scales would be of great interest.

SMEs as well as large enterprises need success stories to be convinced to change their development and rely on new processes!!!

Within EMMC in 2015

Choice and Outline of examples

Creation of examples/case studies with help of PTA and EC case study support

Chapter 2 Second layer of database design: scientific and software requirements

There are many layers in the specifications. The above are the end user specs from the SWO. The second level is the scientific requirements and software requirements and these are discussed here. The technical specifications of the database based on these requirements will be discussed in coming EMMC activities.

Workflow in multi-model simulations

Linking/coupling codes C1 and C2 consists of five steps
1) Code 1 is run and generates output O1 according to a standard of this code
2) The data in O1 needs to be reorganised by a wrapper tool into a format for the linking=interfacing=homogenisation tool.
3) The linking/coupling tool is operated and the data is written according to the linking tool standard L1
4) The next wrapper rewrites the L1 data into an input form I2 that code C2 requires.
5) Code C2 is run generating an output O2 according to a standard of this code

This scenario enables reusing previously developed and well established codes and linking or coupling codes. As soon as a wrapper developed, the information can be shared.

Activity within EMMC

The first step would be NOT to touch the input and output standards of codes (I1, I2, O1, O2) The proposal is that the EMMC Work Group Market Modelling Place (MMP) will concentrate on assembling and coordinating existing activities related to the wrapper operations 2 and 4. The EMMC Coupling & Linking (C&L) WG will concentrate on assembling and coordinating existing activities related the linking/coupling step 3.

Timeline: 2015

Open simulation platform

An “open simulation platform” will enable the linking and coupling of codes. It will be a tool to orchestrate a number of other tools. These tools are software codes/packages for a multiple of purposes:

- Representations of models (defined as approximated physics eqs)
- Solvers used to solve the physics equations
- Data processors like homogenisation tools incl volume averaging
  - E.g. a “Homogenisation tool” would use “ wave functions” as input and provide ”one diffusivity” as output
  - “Estimator tools" would estimate e.g. the diffusivity on the basis of the wave functions even without yet having a precise model for this. These tools can be used instead of homogenisation tools)

- Wrapper tools generating data to any type of tool.
- Tools to steer/control other tools (i.e. “platform tools”) including the steering of hardware tools (e.g. grid computing).
- Tools to store and retrieve data in and from databases.

The OSP will comprise facilities to edit and manage a workflow and conversion tools (wrappers) from and to the platform standard.
The term “open” in this context refers to the platform tool and not to the tools being orchestrated by this tool.

Standardisation

All input and output could in principle be standardized and there are EU activities to design common universal/unified data structures (CUDS) that include all steps. The specification of “communication standards" will be essential to facilitate the orchestration of tools in an open platform. If standard formats are supported the wrappers become less critical. This will also facilitate the integration of the CUDS data in highly efficient, reusable, and well-structured databases.

Activity within EMMC

The EMMC MMP and Open Simulation Platform will coordinate activities to standardise input required to the tool and the output provided by the tool
10 year planning to be broken down as follows:

1. creation of technical basis for CUDS
2. publication of the CUDS specifications
3. choosing a set of software tools covering all scales and developing wrappers for them
4. developing the metadata standard (e.g., keyword bases) for codes for models (defined as approximated physics eqs)
5. assemble solvers used to solve the physics equations
6. assemble data processors like homogenisation tools incl volume averaging etc.
7. assemble “estimator tools"

Timeline: 2015-2025

Wrappers

Future development (beyond 2017-2018), will allow wrappers to be integrated so that the code input files can be converted to any format. This may be very important if the software used originally for the simulation does not exist anymore, so that a conversion of the input to a newer format is needed. This will guarantee the longevity of the knowledge associated with the particular simulation, even if the programs are no longer available for any reason.
These wrappers may be integrated into the databases and allow a simulation to re-use previously published data, thus achieving an accelerated development process. This reuse of resources will also contribute to a saving in efforts and resources.

Activity by ???

Timeline: 2017-2018

Database technology

The meta-database linking existing databases need to be tested, and the best tool chosen. Especially if the linked databases use different API, the meta-data base must be programmed and tested with respect to each. There will also be a need for some code programming and maintaining efforts associated even with this simple and basic meta-database.

Activity by EMMC

Timeline: 2015