EMMC-CSA

European Materials Modelling Council

H2020-NMBP-CSA-2016

NMBP-24-206

Network to capitalize on strong European position in materials modelling and to allow industry to reap the benefits

Start date of the project: 01/09/2016
Duration: 36 months

DELIVERABLE REPORT

Deliverable ID	**D2.11**
Deliverable name	**Report on first horizontal workshop on Metadata/Interoperability and Roadmap contributions**

WP	2	Interoperability and Integration
Task	2.6	Interoperability and Open Simulation Platform session for horizontal workshops

Dissemination level[1]	PU
Nature[2]	Report

Delivery date	11/07/2017

Lead beneficiary	GCL
Contributing beneficiaries	Fraunhofer, ACCESS, SINTEF, EPFL

[1] Dissemination level: **PU** = Public, **PP** = Restricted to other programme participants (including the Commission Services), **RE** = Restricted to a group specified by the consortium (including the Commission Services), **CO** = Confidential, only for members of the consortium (including the Commission Services).

[2] Nature of the deliverable: **R** = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other.

Consortium		
TU WIEN	Technische Universität Wien	Austria
FRAUNHOFER	Fraunhofer Gesellschaft	Germany
GCL	Goldbeck Consulting Limited	United Kingdom
POLITO	Politecnico di Torino	Italy
UU	Uppsala Universitet	Sweden
DOW	Dow Benelux B.V.	Netherlands
EPFL	Ecole Polytechnique Federale de Lausanne	Switzerland
DPI	Dutch Polymer Institute	Netherlands
SINTEF	Stiftelsen SINTEF	Norway
ACCESS e.V.	ACCESS e.V.	Germany
HZG	Helmholtz-Zentrum Geesthacht Zentrum für Material- und Küstenforschung GMBH	Germany
MDS	Materials Design S.A.R.L	France
QW	QuantumWise A/S	Denmark
GRANTA	Granta Design LTD	United Kingdom
UOY	University of York	United Kingdom

EC-Grant Agreement	723867
Project acronym	EMMC-CSA
Project title	European Materials Modelling Council
Instrument	CSA
Programme	HORIZON 2020
Client	European Commission
Start date of project	01 September 2016
Duration	36 months

Coordinator – Administrative information	
Project coordinator name	Nadja ADAMOVIC
Project coordinator organization name	TU WIEN
Address	TU WIEN | E366 ISAS | Gusshausstr. 27-29 | 1040 Vienna | Austria
Phone Numbers	+43 (0)699-1-923-4300
Email	nadja.adamovic@tuwien.ac.at
Project web-sites & other access points	https://emmc.info/

The EMMC-CSA project has received funding from the
European Union's Horizon 2020 research and innovation programme
under Grant Agreement No 723867

TABLE OF CONTENT

1. Executive summary

1.1 Description of the deliverable content and objectives

In line with the objective of Task 2.6, WP2 organised one plenary presentation and three discussion sessions at the EMMC International Workshop 2017. The plenary presentation was given by Emanuele Ghedini, University of Bologna, Italy, entitled: *MODA, a common ground for MOdeling DAta generalization: introduction, use case and possible improvements*. The three session topics were:

- *Vocabulary and taxonomy for improved community integration, communication and interoperability* (Session2)
- *Ontologies and metadata schema and their implementations* (Session 5)
- *Pragmatic approaches to interoperability: implementations, realisations and scenarios of practical relevance* (Session 8)

Discussion Notes were prepared and circulated to all participants in advance of the workshop. The questions in the Discussion Notes were also circulated via an online survey, enabling stakeholder feedback before and following the workshop.

This Deliverable provides an overview of the WP2 sessions and the feedback received from stakeholders answering the survey. Recommendations for the next edition of the EMMC Roadmap are given.

1.2 Deviation from objectives, corrective action (if applicable)

None

1.3 Major outcome

Wide stakeholder endorsement for a community led agreement on terminology and the need to establish semantically based interoperability was obtained. The pivotal role of establishing taxonomies and ontologies became clear as a result of the workshop. Also, the workshop showed the need to support the community in the short term with improved guidance about practical integration and interoperability approaches.

2. Progress report (main activities)

The path from terminology and MODA to Ontologies

The topics of the Plenary, Session 2 and Session 5 of the workshop covered the arch from the establishing a common terminology and a framework for capturing the information about a user case (MODA) to potential future taxonomies and ontologies for Materials Modelling. A wide range of stakeholders provided input via Impulse presentations, survey and workshop discussion contributions. Stakeholders included

- Experts in taxonomies and ontologies with experience from related/adjacent domains (Chemistry, Analytical Science)
- Experts in Chemistry and Materials Science repositories and exchange formats (in particular CIF and Pauling File)

- Modellers (discrete and continuum), Software Owners (academic and commercial), Translators and Manufacturing Industry

The strong need for lowering the barrier to utilising materials modelling was emphasised. This requires

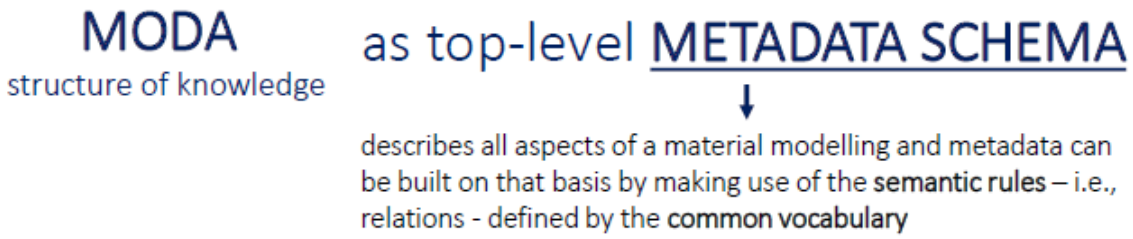
- efficient use and re-use of modelling data
- efficient communication across the field of materials modelling
- interoperability
- a common language, taxonomy and ontology.

A starting point is defined in an ongoing CEN standardisation workshop.

Recommendations and improvements for the terminology and MODA

In his plenary presentation Emanuele Ghedini introduced the MODA and discussed the NanoDome project as a User Case providing MODA and workflow examples. It was summarised that MODA can be used to lay out the top-level (upper) ontology of materials modelling. MODA will allow harvesting specific vocabularies and bridging over vocabulary barriers between different communities by harvesting semantic communalities from different MODA describing essentially the same model schemes.

Future steps should target:



- Exchange of information between materials modelling codes
- Putting data in a form that allows models to properly recognise it along with its meaning
- Deal with the complexity of sharing data between multiple tools (in-house and commercial; proprietary and open)
- Code generation (meta-programming of classes and structures)

As possible improvements, he suggested the following:

- Provide a selected small set of MODA examples for basic user cases for different fields of applications to be used as reference point (easier to navigate than the ones that are published in the RoMM)
- Distinguish between free text field entries (e.g. description) and fixed options (e.g. model entities)
- Provide a first set of standards PE and MR for the most common models
- MODA online form for easy compilation, catalogue and formatting

These points were largely echoed by the survey input and discussion. It was requested to make the use, re-use and interfacing of MODA easier. It was noted that MODA were primarily meant for humans reading but that they could/should be developed into machine-readable form as well. To that end, MODA should be encoded in a standards-based representation where both the schema and the data are machine interoperable. Examples are (1) XSD Schema and XML Data, (2) JSON Schema and JSON Data, or (3) JSON-LD and JSON Data.

MODA repositories and publishing based on MODA were also discussed and would widely be regarded as a useful development. MODA should be suggested to be attached to publications. In this context, metadata should be designed to provide the possibility of data indexing. Also, a system for unequivocal data citation should be developed, analogous to DOI identifier. This would help data mining and dissemination. Making

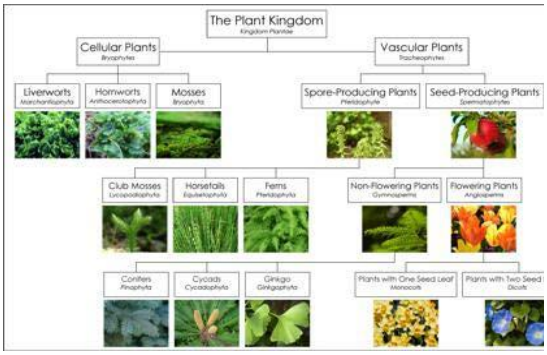
MODAs publishable would also encourage people to spend more effort on them. However, publishers would need to be convinced and the MODA scheme will probably have to be further developed and formalised. It is the requirements of the publisher that will count in the end.

Taxonomy and Ontology overview

An overview/introduction to taxonomy, ontology and semantic system was given by Geoff Gross in his Impulse presentation:

Taxonomy:

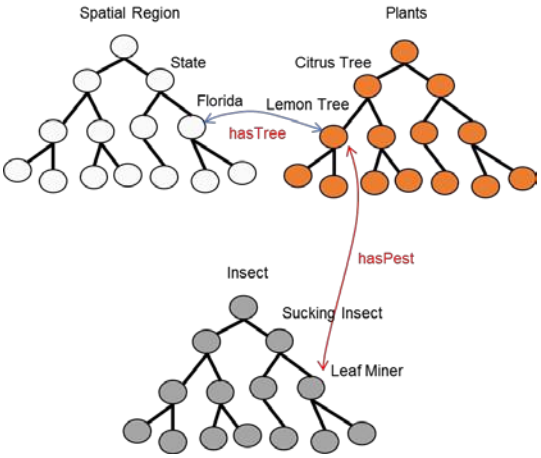
- can only capture subclass/superclass relationships
- no complex relationships
- hierarchical
- represent data as acyclic graphs
- used for grouping information into types



Taxonomy: only parent/child relationships.

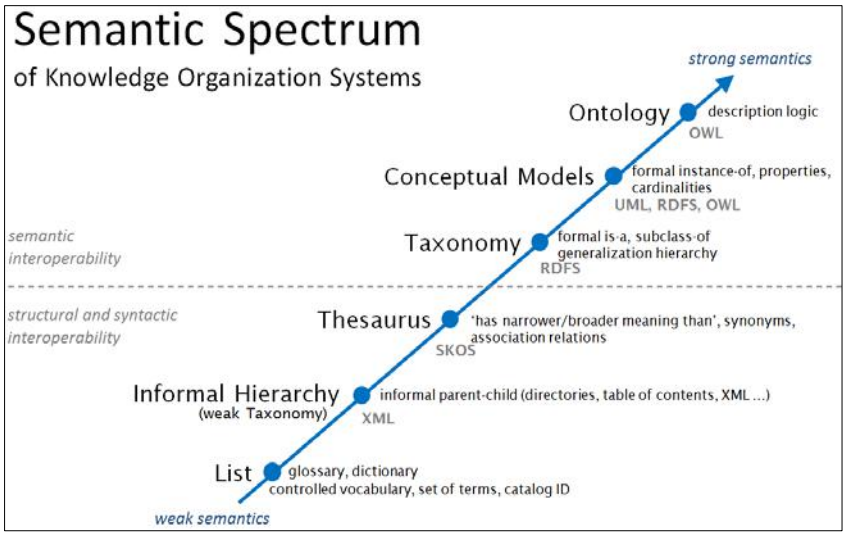
Ontology:

- captures complex relationships
- captures deeper relation types
- captures connected taxonomies
- used for grouping information into types



Ontology: complex relations (between taxonomies)

Evolution of semantic systems



Also, in the survey Piotr Macioł commented that simplified, but working solutions are better than sophisticated, but not working (e.g. in IT industry - REST APIs displace better, but more complex web services). Also, he recommended working less with natural language, more with formal description.

On the other hand, Merja Sippola from VTT commented:

- Agreeing on terms is useful. But it is not easy to make a taxonomy the standard. As close to layman talk as is possible without losing too much accuracy is the best. Lots of pictures will help.
- Do not make it from one field viewpoint only. A group of developers from each major engineering field. They could make the taxonomy for their own field first and then study the similarities and contradictions between fields.

Going forward it was recommended that:

- Vocabulary and taxonomy should be flexible and continually accept new terms and definitions.
- Vocabularies and taxonomies should be hosted on a server, where each concept is an entry in a database.
- Communities should be allowed to enter new terms and definitions and experts will define links between related or synonymous terms.

Experiences and lessons learnt

The first question pertains to the experiences and 'lessons learnt' in establishing and maintaining common terminologies, classification/taxonomy, metadata and ontologies.

The experience from the EC as outlined by Anne de Baas is that both in their grant proposals and reports people cannot find it very difficult to describe their modelling according to basic physics principles. The RoMM and MODA have been an enormous effort. The benefit is clear. The MODA can be understood much more widely and have been a big success in communicating what the field does to outsiders (e.g. non-scientists in the EC, other projects etc.).

Responses to the survey question pointed out that

- There are a lot of ontologies, which are not in common use (It is quite difficult to promote common standard). There should be rather more ontologies focused on particular topics than one huge ontology. However, all dedicated ontologies must be coherent with other ones.
- Physicists, chemists, materials science people, engineers have all different slangs; use different words for same things and same words for different things. It is not easy to uniform it. People are not so flexible.
- You cannot expect people or communities to change an established vocabulary, but that translation is key.
- Agreeing on terms is useful. But it is not easy to make a taxonomy the standard. As close to layman talk as is possible without losing too much accuracy is the best. Lots of pictures will help.
- People really mean different things with same words and same things with different words. The people making the ontologies should come from different fields to avoid bias.

Indeed, we have since also learnt from a key textbook on ontology [Barry Smith, BFO] that in terminology selection it is important to stay as close as possible to the usage of actual domain experts and to incorporate community-specific " synonyms " into the ontology alongside the preferred labels.

In terms of approaches to building an ontology, it was also proposed to start with folksonomy (which is a kind of collaborative tagging approach) and run that alongside formal, top down approach. This could be useful in the context of the above mentioned need to build up community-specific synonyms for example.

The experience of the Crystallography field was also pointed out. The Crystallographic Interchange Framework (CIF) is a common, well maintained data standard which greatly facilitates exchange of information in the field of crystallography. The advantages of the CIF framework can be leveraged in related fields of research. In the discussions it was pointed out by Saulius Grazulis (COD) that CIF is not just a syntactic file format but includes semantics, via a dictionary.

The experience and lessons learnt from the Allotrope project were summarised by Geoff Gross in Session 2 as such:

- Provide guidance on modelling granularity/scope and utilize to inform domain division
    - e.g., from Allotrope: material, process, equipment, result

Comment: *MODA provides such high level domain division*

- Define a unified taxonomy/ontology style guide early and strictly follow design guidance
    - e.g., vocabularies for labels, equivalencies, minimum viable concept, tracking definition sources, etc.
    - Pre-coordination of terms
    - Develop examples to test structural model and educate newcomers
- Standardize taxonomy/ontology development process
    - Leverage existing documentation to make best use of SME time
- Avoid overly narrow concepts and definitions
    - Use subclassing where needed with more general classes
- Version control taxonomies as code
    - Identify standard toolset to simplify diffs across versions (e.g., standard version of OWL API)
- Consider concept and relationship lifecycle management early
    - Develop domain working groups for governance

- Establish and publicise release schedule

The experience and lessons learnt from OpenPHACTS were outlined in Session 5 by Colin Batchelor as such

- Use existing ontologies wherever possible. It is easy to create new ontologies, but is not always needed.
- Use generic words, not depending on material
- Identify things by a naming convention with one human readable label and one machine readable identifier.
- Don't express everything as RDF. RDF can get really huge. For example, instead of including a picture (data) in RDF, include only a reference to the picture (data). If you must express data as RDF, don't use XML!
- Think about processes, inputs and outputs
- Use Aristotelian definitions
- Use the tree structure of an ontology to work for you (avoid "not otherwise specified")

Furthermore, in his survey response Colin Batchelor referred to an article by Barry Smith: "On Classifying Material Entities in Basic Formal Ontology", In Interdisciplinary Ontology, Proceedings of the Third Interdisciplinary Ontology Meeting. Keio University Press (2012) and also referred to the Journal of Cheminformatics which regularly publishes the use-cases of the semantic web technologies.

Amit Bhave (CMCL) contributed experience from engineML - CMCL's XML standardisation for IC (Internal Combustion) engine data (automotive sector) used in process informatics and analytics for automated model calibration, surrogate formulations, etc. Lessons: Importance of wider adoption of ontologies and meta data schema across a sector/community.

Finally, the question was raised whether all this hard work is required or whether a technology like google could work here. The issue is that it is unlikely to work for questions that are not often asked. Also, in fact Google themselves (Knowledge graphs) as well as IBM Watson and Wolfram Alpha use semantics and ontologies.

Stakeholder expectations

The following expectations were expressed, grouped by Modellers, Translators, Software Owners and Manufacturing industry.

- Modellers would like to understand other modellers and to become understood by the industrial people (Merja Sippola). Modellers expect to benefit from easier data exchange (Franz Roters) and semantic description of variables, constraints and processes (Piotr Macioł).
- Translators may benefit from taxonomy, but less than most people think. Person-to-person talk is the best (Merja Sippola). They would benefit from semantic description of variables, constraints and processes (Piotr Macioł).
- Software owners can make manuals using the taxonomy and hope to be understood. The risk is that no one understands (Merja Sippola). Fabio Sacconi (Tiberlab) commented: From the point of view of Software Owners, the need is for tools which are able to meet the requirements of industrial customers. This means that Software Owners can afford the effort of adopting a common data model following a metadata schema only if this will eventually be of actual benefit to industry. Since many data formats and "standards" have been presented in the last years, Software Owners may be unwilling to adopt a new standard, since this can require a strong investment in code development and maintenance. Thus, an appropriate effort should be devoted to reaching a wide

agreement among stakeholders about the classification of models first and then the definition of metadata standard.
- Stakeholders such as the Open Crystallographic Databases expect stable, uniform, royalty-free standards for data interchange.
- Manufacturing industry people would like to understand what those modellers are talking about. (Merja Sippola)

Stakeholder co-operation to come to a common approach to metadata schema and ontologies for materials modelling

It is clear from the above lessons learnt that close stakeholder involvement and co-operation is important (see e.g. the points regarding domain working groups for governance).

In the survey, the following were proposed:

- The approach to reach a common schema should be progressive: better start to define a common schema with few elements rather than to wait for full convergence.
- Operate a schema repository and registry. Operate a datatype registry. Operate a vocabulary server.
- Joint research projects with the leaders of the field, e.g. German DFKI (Prof. Hans Uszkoreit), Max Planck Institute for Informatics (Prof. Gerhard Weikum) etc.

Maximise the benefit to industry

In general terms it was commented that fast access to reliable data and easier interoperability should be a great enabler for industrial processes. The better understanding between communities, the faster is the development!

- As the Functional Mock-up Interface standard shows, industry has a strong interest in interoperability and even started initiatives to achieve it.
- Allotrope.org is such an initiative driven by the pharmaceutical industry which invests in metadata standards out of need.
- Ontologies are already heavily used in industry by information technology companies: e.g. Google, IBM Watson and Wolfram Alpha (references are given in https://blog.tilde.pro/semantic-web-technologies-on-an-example-of-family-trees-7518f3f835a9).
- There are also several smaller nearly-commercial projects, e.g. Chemical Semantics (http://chemicalsemantics.com). The benefit is similar to using the databases, with the difference that the data are self-described.
- A metadata system which can ease data mining and interoperability would help industry in adopting material modelling at the R&D level.
- Easy and license free or cheap access to at least the metadata and some summaries. And the contact information of the person/people/organisation who gave the data.
- Autonomous communication between 'non-human-actors', but that requires common and stable ontologies.
- Provide open standard data formats.
- Industry needs reliable software tools which are able to yield an added value to the product design cycle. This value can be enhanced through interoperability, provided that it satisfies real use case requirements.

Going forward, industry requires clear use case/demonstration for industrial applications. Adoption and usage by industry and academia will come when the models will be described with computer-readable language (automatic searching, translating and 'matchmaking' available). Sustainability and continued

maintenance and upgrade of the ontologies will be required. Working with software owners will be important.

Model interoperability and software integration

This topic was discussed in session 5 and in session 8. The answers to the survey questions shown below tell the story: there is no systematic, coherent approach at the moment. Interoperability is almost completely at the syntactic level. There are clear and strong requests for facilitating interoperability in a more standardised, semantic approach, hence validating the approach pursued by the EMMC.

How do you deal with integration and interoperability at the moment?
What current integration and interoperability implementations do you use?

Replies to these quite similar questions are shown below. The key points are that the current situation is dominated by ad hoc integration and interoperability and that syntactic approaches are used.

- Differently in different problems. **No systematic approach**; each workflow is different: e.g. based on either simple approaches (parameter passing and surrogate models) or transferring, manipulating complex data files. At the moment we often do it in terms of **import export of data and there is no semantic framework in place**.
- For interoperability, development of **XML schemas to define interchange formats** for data and metadata. For integration **convening**.
- In our software (OCTA, http://octa.jp), we are providing **common data format** which can be used by MD and continuum model. E.g., using the results of mean field, initial structure of MD is created. In next version, functions of MD solver dll can be called from other model.
- We started to **connect** nanoHUB tools **with outside resources** like openKIM, Materials Project. We use Web-services and site-specific APIs.
- Our institution is using an **integration platform**. Also we heavily use **open, standardized data formats**, such as vtk or hdf5.
- Current practice **- manual integration** of submodels:
    - on the programming **API** level
    - with manually developed translating **files**; **JSON APIs** translations in C++
- Now developing: C++ based common modelling platform**, enforcing coherence of submodels interfaces on programming level** (metaprogramming, verification by C++ compiler)
- We try to develop **open interfaces** and use s**pecific data transfer** modules.
- Our company has developed a software code which is able to couple internally not only different physical models based on FEM representation, but also these FEM continuous models with discrete models based on an atomistic representation. It is in principle possible to link our software with other FEM and atomistic codes via **appropriate APIs and file exchange**. Beyond internal coupling through TiberCAD capabilities, we have been able to **exchange calculated data through discretization grids** from other simulation software, such as CST or COMSOL.
- **REST APIs**, version control systems and platforms based on them (Github, BitBucket)
- ImageJ Weka segmentation + Simpleware + ABAQUS. For example. No direct connection, but **import and export**.

Which integration and interoperability environments are you familiar with?

It is symptomatic of the incoherence of the field that quite a range of tools and environments are mentioned (Firework, Cordra, Salome, MuPIF, SimPhoNy, AM3, DEEPEN). Python seems to be the main scripting tool.

- Most often it is done through **python** scripting. The python scripting does the pre-processing, launch the simulations, collect and post-process the data and feed them to the next model in the workflow. We also used F**irework** for orchestration and **Mongodb** for data storage, exchange and building of surrogate models.
- **Materials Data Curation System**, **Cordra**
- Source code of any computer program eventually serves for human communication (cf. journal article). Thus, the **software development process** per se presents a universal integration and interoperability environment.
- **Salome**, **Mupif** and **Simphony** platforms
- **Python** scripting.
- Own, **in-house code** (AM3) (Piotr Macioł)
- We have developed an integrated platform for multiscale simulation in the framework of the FP7 **Project Deepen**.

Strengths and weaknesses of current approaches

It follows from the above that current approaches lack generality and can be laborious. They are dependent on adhering to particular formats and metadata. They can still be efficient though.

- It is totally problem dependent. There is **no general solution** to the problem.
- Strength: universality, weakness: overcomplexity.
- Laborious. Case by case.
- Own, in-house code for integration (AM3) **requires high programming skills**, significant amount of work. JSON - slow, not suitable for fully coupled models, difficult to ensure validity of files.
- Specific solutions for individual interfaces, can be **very efficient, but they are not general**.
- Speaking of Deepen platform, this implementation is based on APIs which avoid the need to make any change in the code of software tools to be added to the interoperability environment. This allows an **easy extension to new applications**. The interoperability in this platform is based on a syntactic model. This means that it **relies on a particular data model (in our case HDF-based) as common interchange format**. No semantic information is yet included in the implementation.

What are your top three requests for future interoperability developments?

It is called for future developments to be done collaboratively by users and software developers together. The approach should be based on ontologies and semantics in order to avoid restrictions to particular data formats. The development should be open and community based and lead to open interfaces, open data formats and a common language. The outcome should be easy coupling and linking and transferability of codes (i.e. “plug and play”).

- **Transferability of codes**: running the same simulations with different codes (e.g. dl_poly and Gromacs, LAMMPS and dl_meso, openfoam and fluent)
- **Easy coupling and linking**
  - atomistic and mesoscopic models
  - mesoscopic and continuum models
- **Open, community maintained** development
- **No restriction on particular data format**
- The users and developers to do it together.

- **APIs enriched with semantics**. Standardized APIs. **Common data format (best if semantic aware)**
- Open interfaces. Open data formats. Common language where possible.
- Ideally, implemented **APIs should be based on a precisely defined ontology** of the data type to be exchanged. In this way, interoperability could be generalized more quickly to new codes. Also, all the information about each step of the executed simulation workflow could be stored in an efficient way.

What would be the practical next steps in addressing integration and interoperability?

Contributions to this question can be grouped into considerations about the process, about a syntactic approach and about a semantic approach.

Concerning the process the following were recommended:

- **Discussion** within the community and involving all the stakeholders.
- **Disseminate** information about different approaches to interoperability.
- Give it enough **time**, ensure that the process is widely accepted.
- Make sure that **different ways** and projects will be considered, supported and let the market filter out.

A number of contributions were about formats and syntactic approaches:

- **Recommendations of formats** to use. For examples in image based modelling lossless image storage format is essential. One could also think of a FEM-to-FEM program that would only translate from one program format to other and create metadata to explain the differences.
- Agreement on/development of **common (open) data format**.
- Effort in understanding how **simple data integration** suffices.
- Make "**single entity" models more compatible**, transferable and interoperable.

Regarding semantic approaches:

- Agree on a **common standard terminology**
- **A common approach to define a common metadata schema** for material modelling is fundamental. Based on this a first pragmatic approach based on syntactic level could be carried out.
- Deploy a **schema repository and registry**. Deploy a **datatype registry**. Deploy a **vocabulary server.** Note: In the discussion it was also pointed out that the above are important in order that issues such as typos in data entries can be noticed and corrected. Serialisations such as xml schema, json schema can validate records
- Start supporting construction of **ontology-based simulation** (not just data), cf. Semantics platform.

What promising developments are you aware of?

- **NIST** will operate a **schema repository and registry** for the materials science and engineering community.
- **Semantic technologies** (e.g. DublinCore, RDF, ontologies), **social network mash-up** technologies (e.g. OpenGraph, Oembed), **mobile programmable** collaborative environments (e.g. Slack, Telegram)
- **nanoHUB tools can query openKIM** for interatomic potentials, download appropriate ones and use them in simulations.
- **Object-oriented approaches** are the right way, due to their capability to abstract from particular data storage, data formats, or implementations.

- FMI/FMU standards **Functional Mock-up Interface**: https://www.fmi-standard.org/
- Image based modelling is developing fast. But good meshers like HyperMesh should be linked to this development.
- Building a **Materials Data Infrastructure** by TMS.
- Increasing use of **HDF5**.

Driving forward interoperability including an "Open Simulation Platform" reference design

- It was remarked that "Since many initiatives are already active in several European communities, an effort should be made to take advantage of these activities and to unify the efforts in a single initiative aimed to the implementation of a common strategy."
- On the other hand, a point was made in favour of "natural diversity and decentralization. It is desirable for several reference designs to compete. Here the most important is benevolent environment: education, public and private funding etc."
- The Functional Mock-up Interface standard was mentioned as a successful example of interoperability. "The development was initiated by Daimler AG with the goal to improve the exchange of simulation models between suppliers and OEMs. As of today, development of the standard continues through the participation of 16 companies and research institutes. FMI is supported by over 101 tools and is used by automotive and non-automotive organizations throughout Europe, Asia and North America".
- Open platform, in order to be accepted and widely used must allow incorporation of different models, and even more importantly to support different data formats. In my opinion, this could not be achieved by providing conversion filters between the formats or to one reference format, as the conversion always introduces the additional errors, for example. The possible solution is to establish an abstraction layer on top of different data formats representing the same physical entity, allowing working with all formats using the same, generic interface. This could be realized as software layer between models and data, or using object-oriented approaches.
- Interoperability and the adoption of standards helps to grant the openness of the platform, therefore it should be compliant to the specific community requirements as much as needed and re-use existing standards and components as much as possible.
- Ontologies are not enough for OSP. Low-level tools are necessary to support communication between models and to verify integration (in low-level, programming and high level, semantics of variables, constraints and phenomena). These tools should be consistent with common ontologies.
- Once an ontology is available and full semantic interoperability thus is achieved, the next challenges to be addressed relate to timing and conditions triggering the information exchange and to respective decision making tools being assembled along with the models in flow chart type simulation ecosystems.
- In the end the expectation is for "a set of full turn-key simulation work-flow driven solutions for industrially-relevant practical problems."
- Demonstrators are needed to show that EVERYBODY benefits.

Pragmatic approaches

Session 8 Impulse presentations and discussions focussed on so-called pragmatic (i.e. largely syntactic, file formats and specific metadata keywords based) approaches that provide current and near-term solutions to the above interoperability and integration issues. Impulse presentations were:

- Georg Schmitz: Practical Interoperability - the view of a software provider
- Joerg Neugebauer: How to achieve interoperability? A modeller's perspective

Georg Schmitz pointed discussed file based exchange, providing the possibility to store, exchange, visualise and track data. HDF5 is a format that naturally supports the hierarchical relationships of materials structure. Materials metadata schema (see also publications by Schmitz et al) can straightforwardly be represented in this format.

Joerg Neugebauer discussed challenges in connecting codes such as inconsistence in parameters, lack of error estimates and imbalance in CPU times (i.e. load balancing issues). The Pyiron framework developed by his group is based on creating API standards that work for multiple codes “code agnosticism”). It also utilises the HDF5 format for data friendly and complete output (not ASCII!), to store complex data sets. I addition it uses an SQL database to efficiently browse through projects. Jupyter Notebook is used to run, analyse and document protocols. Run time exchange of data between codes is supported by python bindings, named pipes. It includes automatic handling of (some) runtime errors by creating corrected input files. It also includes automatic setting of code/method specific convergence parameters.

Integration is improved by translating input files for multiple codes and some interoperability provided by model linking schemes.

It was discussed following the presentation whether there are there core parts of e.g. atomistic frameworks that can be sections off and done once and for all e.g. parsers, schedulers, etc.? There seems to be a lot of support/enthusiasm for this - it may be possible to get people together and define some common intermediate.

In the discussions, HDF5 format received further support, e.g. the crystallography community has developed ontologies which are implemented in HDF5 and more widely used codes such as VASP are also moving to HDF5. The reservation that checking of files is more difficult since HDF5 is binary was made but not widely seen as an issue due to the range of tools available.

3. Conclusions and Roadmap Recommendations

Existing approaches to achieve interoperability in practical implementations are on a syntactic level. They are based on file based exchange, typically built on top of some standardized or community accepted formats, such as vtk, hdf5, etc. More elaborated implementations used in different simulation platforms realise the data exchange by means of a so called neutral format. The individual applications have to perform conversion from their native i/o into the neutral format. The advantage of this approach is that only minimal modification of existing tools is necessary, the downside is the need to maintain individual conversion tools and the conversion itself may introduce additional numerical error. The main identified barriers to semantic interoperability are the lack of community accepted standard, missing roadmap and templates illustrating interoperability at the semantic level.

A roadmap towards semantic interoperability needs to be laid out, including short and medium term steps.

Taxonomies and Ontologies

In particular, the need for taxonomies and ontologies should be highlighted in the roadmap. These should build on the existing RoMM terminology and MODA, the CWA and take into account the lessons learnt that were outlined in this document. Ontology development for Materials Modelling should also set the field into the context of wider developments including materials ontologies, characterisation ontologies (including Analytical etc.), and PLM, as promoted by the Industrial Ontology Foundry.

A clear definition of terminology used to describe interoperability needs to be outlined (i.e. interoperability of models, integration of software tools, syntactic file formats for interoperability, what is meant by semantic interoperability etc.).

Short term steps towards interoperability

The short term need for the research and industrial community to have interoperability should be acknowledged. Short term actions to improve interoperability and enhance productivity could be included in the context of a roadmap towards establishing community accepted standards. This could include an overview and guide to interoperability in materials modelling.

4. Annex: Discussion Notes

Discussion Notes Session 2: Vocabulary and taxonomy for improved community integration, communication and interoperability

Introduction

Due to the complexity of materials and the wide range of applications, the materials modelling field consists of a number of communities. With the focus on a specific application domain (along with specific models and domain specific software codes) each community has evolved a domain related terminology. However, applications to industrial problems in advanced materials and nanotechnology require a strong interdisciplinary approach between these fields and communities. There is hence a need to establish a common terminology (definition of concepts and vocabulary) in materials modelling which will lead to greatly simplified and much more efficient communication. A standardised terminology could improve future exchanges among experts in the entire area of materials modelling and also facilitate the exchange with industrial end-users and experimentalists and lowering the barrier to utilising materials modelling. The common language is hoped to foster dialogue and mutual understanding between industrial end-users, software developers, scientists and theoreticians.

Objectives of common vocabulary, taxonomy and metadata

- Greatly simplified and much more efficient communication across the field of materials modelling.
- Much improved communication of materials modelling across other science and engineering disciplines.
- Much improved communication of materials modelling to industry.
- Easier, more efficient use and re-use of data from and information about materials modelling.
- Interoperability, including between models and databases.
- Support the process of translating industrial problems into problems that can be simulated with materials models.
- Assist workflow development where several models can interoperate.

Background information and documents

- Review of Materials Modelling[3] (RoMM): Now in its 6th edition it puts forward a vocabulary, classification and metadata for materials modelling, developed and tested in the context of 130 FP7 and H2020 project involving about 2000 modellers.
- The RoMM includes a Metadata schema, called the Materials Modelling Metadata (MODA) which organises the information about user cases and their simulation so that even complex model workflows can be conveyed more easily and key data about the user case, models, solving and post processing (together called simulation) can be captured.

A key point of the classification is that it leads to a relatively small number of distinct classes in which the overwhelming number of materials models can be placed. This is replacing the current situation of opacity of materials models that make the field hard to access for outsiders.

Ongoing activities

- A CEN standardisation workshop[4] is currently underway with the aim of reaching a wide stakeholder agreement on the terminology and classification of models and thus be a taxonomy.

Discussion points and questions

- What is the experience and lessons-learnt from other communities in establishing a shared vocabulary, classification and metadata?
- How can the benefit to industry be maximised? I.e. How can common terminology and MODA be used to lower the barriers to benefiting from materials modelling in industry?
- What are the expectations from stakeholders (Modellers, Translators, Software Owners, Manufacturing industry)?
- What improvements are needed/recommended for the terminology and MODA?
- How can the common terminology and metadata support publications about materials modelling? A metadata schema would help to communicate, disseminate, store, retrieve and mine data about materials modelling. Should/could CEN/CWA become a guidelines for use of terminology in publications? Should/could MODA be introduced as Supplementary Information in publications?
- Should there be a MODA repository?
- How can classification and MODA be built upon to drive forward interoperability?

Discussion Notes Session 5: Ontologies and metadata schema and their implementations

Introduction and Background Information

All stakeholders of materials modelling face barriers regarding access to, and use of, information about materials modelling, utilisation of the wide range of modelling tools and methods, and last but not least interoperability of models and codes.

Metadata can be defined as “data describing the context, content and structure of records and their management through time”[5]. They provide information that allows for categorization, classification and

3 https://emmc.info/version-6-of-the-romm-is-now-available/

4 https://emmc.info/cen-workshop-to-establish-common-terminology-in-materials-modelling/

5 ISO 15489-1 s 3.12

structuring of data[6]. A well-established example is the Crystallographic Information File (CIF)[7] which provides metadata for atomistic structures and properties.

A **metadata schema** can be defined as "a logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality (obligation level) of values." [8]

An **ontology** is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain.[9] Ontologies aim to define which entities, provided with their associated semantics, are necessary for knowledge representation in a given context.[10] Ontologies and related information technology provide an opportunity to share a common understanding of the structure of information within a specific domain, the possibility to reuse domain knowledge, to make domain assumptions explicit and to analyse domain knowledge.[11]

One of the most advanced attempts at establishing a metadata schema is the Chemical Mark-up Language (CML) development in the field of chemistry, electronic and atomistic modelling[12,13]. CML provides support for most chemistry, especially molecules, compounds, reactions, spectra, crystals and computational chemistry. In particular, the effort extended to dictionaries and semantic web tools.

The electronic structure and atomistic modelling community has been advancing metadata and schema through a number of initiatives (some involving experimentalists) including those of (T)COD[14], NOMAD[15], ETSF[16], or the Pauling File[17]. To enable reproducibility, metadata should include provenance which can also be used to generate metadata for any schema. Automated provenance tracking and metadata exporters[18] are included in AiiDA[19].

More widely, there are metadata and schema initiatives in the materials science field, see e.g. the Materials Information File (MIF) schema by Citrine Informatics[20] and the publications by Ashino[21] and Schmitz et al**Fehler! Textmarke nicht definiert.**.

6 G.J.Schmitz et. al.: Towards a metadata scheme for the description of materials – the description of microstructures Science and Technology of Advanced Materials 17:1(2016) 410-430

7 http://www.dcc.ac.uk/resources/metadata-standards/cif-crystallographic-information-framework

8 Building a metadata schema – where to start. ISO/TC 46/SC11N800R1

9 https://en.wikipedia.org/wiki/Ontology_(information_science)

10 Thomas R. Gruber (1993). Toward principles for the design of ontologies used for knowledge sharing. Originally in N. Guarino and R. Poli, (Eds.), International Workshop on Formal Ontology, Padova, Italy. Revised August 1993. Published in International Journal of Human-Computer Studies, Volume 43 , Issue 5-6 Nov./Dec. 1995, Pages: 907-928, special issue on the role of formal ontology in the information technology.

11 David Lamas, Metadata and Ontologies, 2011; https://www.slideshare.net/davidlamas/metadata-and-ontologies

12 http://www.xml-cml.org/

13 http://www1.gly.bris.ac.uk/~walker/CMLComp/

14 Gražulis, S., Merkys, A., Vaitkus, A., Bail, A. L., Chateigner, D., Vilčiauskas, L., Cottenier, S., Björkman, T. & Murray-Rust, P. (2014). Acta Cryst. A, 70, C1736., http://crystallography.net/tcod/

15 https://metainfo.nomad-coe.eu/nomadmetainfo_public/info.html

16 http://www.etsf.eu/fileformats

17 http://paulingfile.com/, http://developer.mpds.io/#JSON-schemata

18 A. Merkys, N. Mounet, A. Cepellotti, N. Marzari, S. Grazulis, G. Pizzi (to be submitted)

19 G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, Comp. Mat. Sci. 111, 218-230 (2016), http://www.aiida.net/

20 http://citrineinformatics.github.io/mif-documentation/#a-name-core-a

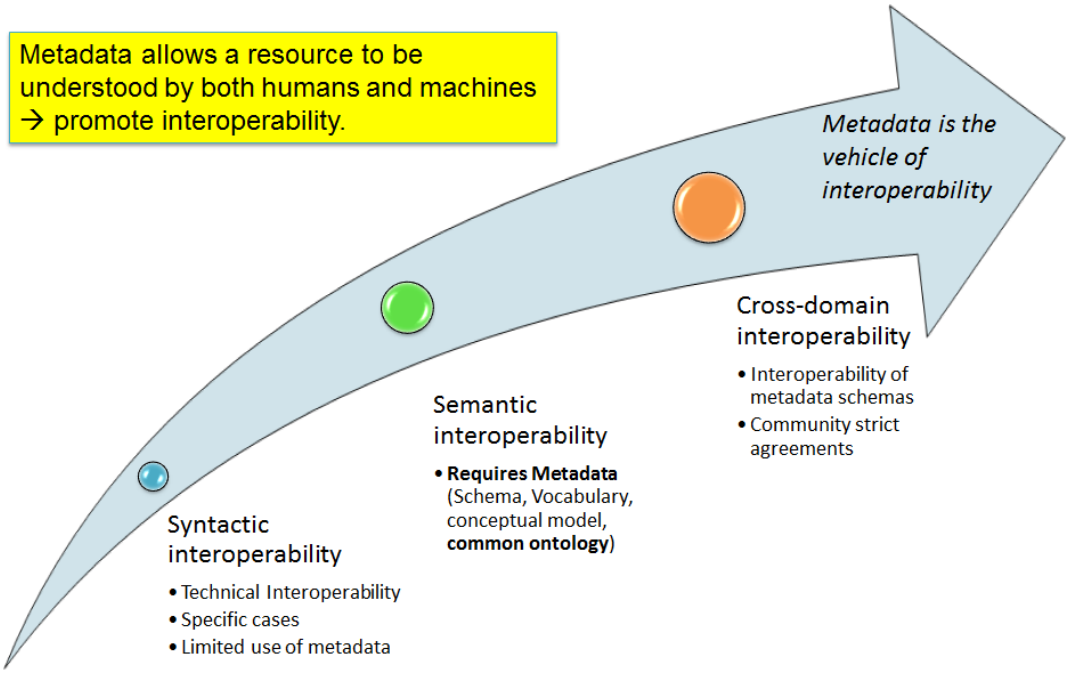
Efforts to advance interoperability include the OPTiMaDe initiative[22] (to define a query/retrieval format to make crystallographic databases interoperable over REST) and the European projects MaX[23] (to develop turn-key solutions to compute physical properties using automated workflows for a variety of codes) and E-CAM[24] (to develop code independent libraries).

For model interoperability, metadata schema also need to cover linking and coupling. Initial work has been carried out in the SimPhoNy project[25].

The Review of Materials Modelling (RoMM) includes a metadata schema, called the Materials Modelling Metadata (MODA) which organises the information about the user case and the simulation. The four elements of the MODA (Use Case, Governing Equations, Solver, Pre-/Post Processing) can be regarded as representing 'core' ontologies. For the Governing Equations, the RoMM provides the elements for an ontology based around the four model entities.

Ongoing activities

The aim is to build on the MODA and develop a widely agreed Materials Modelling Ontology (MMO) with a view to achieving cross-domain interoperability. The ongoing activities work is based on this vision of the chain of development:



In the long term each domain implementation should only need to map to a widely agreed Materials Modelling Ontology (MMO) in order to achieve interoperability with other implementations. The challenge is to make MMO generic enough to express all relevant metadata.

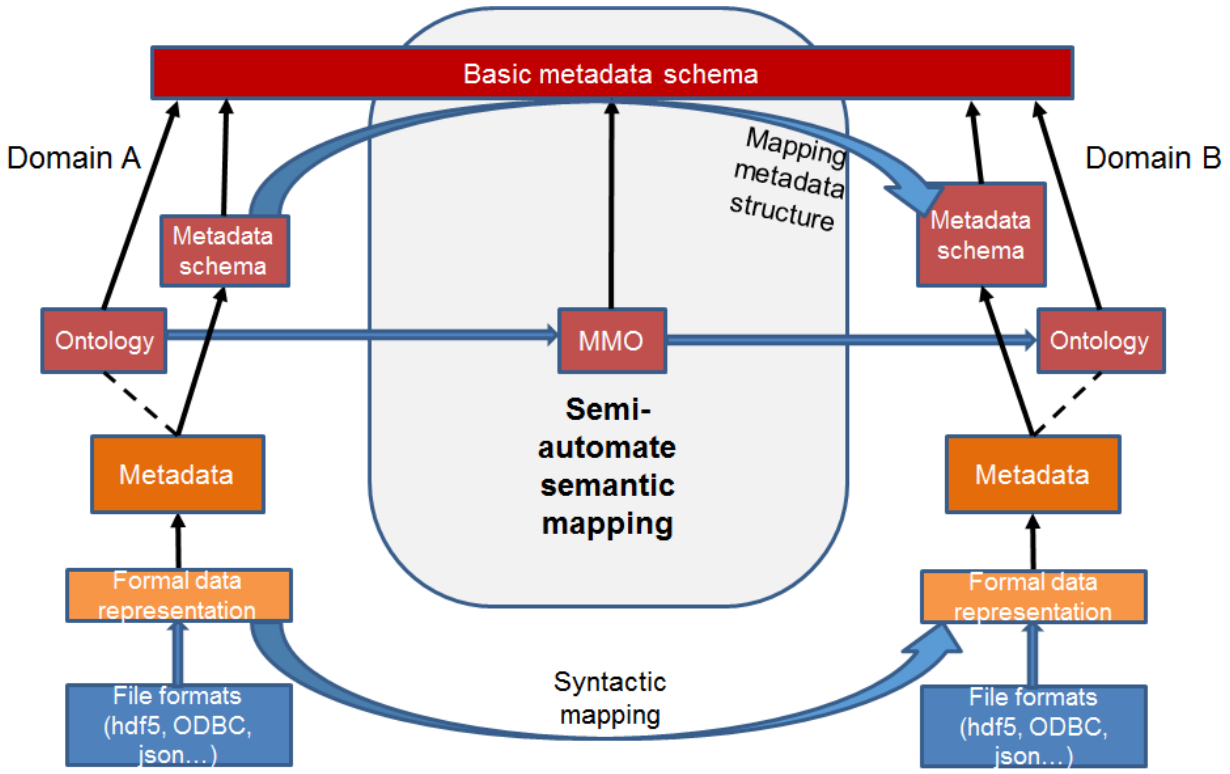
21 T. Ashino: Materials ontology: an infrastructure for exchanging materials information and knowledge. Data Sci J. 2010;9:54–61.

22 http://www.optimade.org/, https://www.lorentzcenter.nl/lc/web/2016/826/info.php3?wsid=826&venue=Snellius

23 http://max-centre.eu/

24 https://www.e-cam2020.eu/

25 https://github.com/simphony



Objectives of the session on metadata schema and ontologies

- Discuss the technicalities of how to achieve cross-domain interoperability
- Discuss how to reach co-operation across all players and come to a joint agreement on metadata schema and ontology development.

Discussion points and questions

- What are the experiences and lessons learnt from other communities in establishing, using and maintaining a metadata schema and ontology?
- What co-operation can be installed to come to a common approach to metadata schema and ontologies for materials modelling?
- How can the benefit to industry be maximised?
- What are the expectations from stakeholders (Modellers, Translators, Software Owners, or Manufacturing industry)?
- How can metadata schema and ontologies be introduced and used in repositories?
- What is the best path towards an “Open Simulation Platform”, a reference design that supports wide interoperability?

Discussion Notes Session 8: Pragmatic approaches to interoperability: implementations, realisations and scenarios of practical relevance

Introduction and background information

The issue of interoperability refers to "the ability of computer systems or software to exchange and make use of information"[26]. In materials modelling there is often the need to link models in order to address typical industrial problems. Efficient integration and exchange of data is also becoming more and important in the light of high throughput computation, enabled by modern hardware.

Interoperability can be achieved by different means and on different levels, from syntactic (data format) definitions that enable 'import/export' via integrated platforms that operate on the basis of a certain 'data model' to semantic level interoperability. The long-term goal of achieving cross-domain interoperability based on a semantic framework (see also Session 5) will take some time to achieve. There are, however, a number of pragmatic approaches to interoperability under way. These may be based on limited metadata sets, on particular file formats such as HDF5 and/or limited to certain domains.

It is useful to distinguish between different levels of integration and interoperability, in particular regarding coupling and linking of models.

- Certain codes integrate two models linked or coupled internally in the code.
- There are a number of codes in use that can be considered as standalone, i.e. they are not integrated into any platform.
- A number of 'software packages' (typically proprietary) provide some level of integration within one platform. The integration may be limited to a common visualisation/graphical user interface. Typically there is also further interoperability, e.g. in atomistic modelling packages, different codes can operate on the same atomistic structure. As a result, these platforms support linking of certain models. However, linking is often limited to models of the same entity type (e.g. different atomistic models).
- Open API integration and workflow environments are designed for easier and more flexible integration of codes, and for developing, managing and executing workflows. Examples of existing (open and proprietary) integration and interoperability environments include ASE[27], AIIDA[28], MuPIF[29], DREAM.3D[30], KNIME[31], Pipeline Pilot[32], AixViPMaP[33], Salome[34] and Simphony[35]. The same considerations and limitations regarding linking as above apply.

Objectives

- Establish an overview of current pragmatic interoperability approaches.
- Discuss the objectives, scope, advantages and limitations of existing implementations and realisations.
- Discuss needs and requirements for future developments.

26 https://en.oxforddictionaries.com/definition/interoperability

27 https://wiki.fysik.dtu.dk/ase/

28 http://www.aiida.net/

29 https://sourceforge.net/projects/mupif/

30 https://immijournal.springeropen.com/articles/10.1186/2193-9772-3-5

31 https://www.knime.org/

32 http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/

33 http://www.iehk.rwth-aachen.de/index.php?id=597&L=2

34 http://www.salome-platform.org/

35 https://github.com/simphony

Discussion points and questions

- How do you deal with integration and interoperability at the moment?
- What current integration and interoperability implementations do you use?
- Which integration and interoperability environments are you familiar with?
- What are the strengths and weaknesses of current implementations?
- What are your top three requests for future interoperability developments?
- What would be the practical next steps in addressing integration and interoperability?
- What promising developments are you aware of?