Innovative Data Management in advanced characterization: implications for materials design

N. Romanos[1], M. Kalogerini[1], E.P. Koumoulos*[1,2,3], A.K. Morozinis[2], M. Sebastiani*[3,4], C. Charitidis*[2,3],

[1]Innovation In Research & Engineering Solutions (IRES), Boulevard Edmond Machtens 79/22 - 1080 Brussels, Belgium

[2]National Technical University of Athens, School of Chemical Engineering, RNANO Lab – "Research Unit of Advanced, Composite, Nano Materials & Nanotechnology", 9 Heroon Polytechneiou Street, GR-15773, Zographos Athens

[3]European Materials Characterisation Council, Belgium, http://characterisation.eu/

[4]Università degli studi Roma Tre, Engineering Department, via della Vasca Navale 79, 00146, Rome

* corresponsing authors: epk@innovation-res.eu; seba@uniroma3.it, charitidis@chemeng.ntua.gr

Abstract

This paper describes a novel methodology of data management in materials characterisation, which has as starting point the creation and usage of Data Management Plan (DMP) for scientific data in the field of materials science and engineering, followed by the development and exploitation of ontologies for the harnessing of data created through experimental techniques. The case study that is discussed here is nanoindentation, a widely used method for the determination and/or modelling of mechanical properties on a small scale.

The same methodology can be applicable to a large number of techniques that produce big amount of raw data, while at the same time it can be invaluable tool for big data analysis and for the creation of an open innovation environment, where data can be accessed freely and efficiently.

 doi:10.20944/preprints201903.0205.v2

Aspects covered include the taxonomy and curation of data, the creation of ontology and classification about characterization techniques, the harnessing of data in open innovation environments via database construction along with the retrieval of information via algorithms. The issues of harmonization and standardization of such novel approaches are also critically discussed.

Finally, the possible implications for nanomaterial design and the potential industrial impact of the new approach are described and a critical outlook is given.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

1. Introduction

Nowadays, the challenge of digital innovation is to connect fast growing and emerging technologies to the market needs and society demands, while it is commonly accepted that innovation is a significant determinant of business competitiveness in markets [1]. The outcome of investment in research and innovation process is shifting, leading stakeholders to get connected into aggregates that have the characteristics of 'living' ecosystems, where information can be exchanged openly. In that way, massive amounts of data can be handled and analysed simultaneously, as a result stakeholders can get clearer and wider insights towards new horizons in productivity growth [2].

As an overall practice, data management is strongly connected with the entire lifecycle of big data implementations, including the primary steps of data creation, growth, variations and final storage. Data Management Plans (DMP) could facilitate the above aspects, as they play a major role in data organization, traceability, accessibility, interoperability, reusability and finally long-term and secure storage [3]. As data can be of different types (numerical, nominal, categorical, Boolean -true or false-, structured), a common vocabulary for every aspect promotes the sharing of information in specific domains. The above practices in combination with the use of ontologies as tools for bridging datasets across domains as well as fast and efficient data extraction, can contribute to simplify big data's representation and promote the development of efficient computing models for advanced materials design.

Experimental material science sees today a unique opportunity for a groundbreaking innovation through materials digitalization, as confirmed by the relevant investment plans that have been recently established in the worldwide leading economies.[4,5]

In this framework, as materials science is a multi-stakeholder field, a Materials "Entity" Initiative for Competitiveness (shortly called as "entity") is needed, in order to will reduce development time providing infrastructure and training to parties for optimal discovering, development, manufacturing and deployment of innovative materials. This initiative could boost production and commercialization of materials in a more expeditious and economical way, increasing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

competitiveness. In this way, the whole cycle from research to manufacturing must operate both faster and at lower cost. Main pillars of the initiative should be data sharing and analysis (e.g. computational capabilities, data management, standards), that will generate a knowledgebase for better leverage and complement of investments.

Barriers until now are briefly described below [4]:

1) The lengthy time frame for materials to move from discovery to market. As much of the design and testing is currently performed through time-consuming experiment and characterization loops. Some of these experiments could potentially be performed virtually, through materials digitalization, with powerful and accurate computational tools.

2) *Several discrete stages are present during the path from conception to market deployment*. A connection among stages is needed, to facilitate continuum processes.

3) *Data transparency, communication and integration.* There is currently no standard method for researchers to share data, as well as predictive algorithms and computational methods.

4) *Recyclability and sustainability of new materials.* Recyclability must become a design parameter during the whole manufacturing cycle in order to deal with sustainability.

To overcome the aforementioned barriers the "entity" must embrace open innovation and act as a data exchange system- Hub (index, search, and compare data). It could help replace lengthy and costly empirical studies with mathematical models and computational simulations, reducing costs and time. Such modification is expected to shorten materials deployment cycle from its current 10-20 years to 2-3 years [6]. To work for the benefit of stakeholders and community, the "entity" requires contribution in three critical areas, namely computational tools, experimental tools and digital data.

Moreover, tools to simplify and promote data discovery, data reuse, and development of advanced materials informatics, is critical to transforming a research-to-market adoption pipeline [7].

 doi:10.20944/preprints201903.0205.v2

Digital data can be transferred efficiently, kept safe and reach worldwide stakeholders at high speeds, while also acquire value and significance far beyond their original purpose. This is one driving force that leads information science practitioners to provide data curation services. As the necessary information coming from computational and experimental data is available in a machine-readable format, data curation is a fundamental practice. Some factors that influence actions to provide data curation services include incentives from the funding bodies and the scholarly publishing entities, while other factors are associated with the research communities themselves, which demand higher transparency in research [8].

Increasingly, in recent years, there has been a strong recognition for the critical need of a third component of data science. This component deals with the online tools designed specifically to seed and nurture cross-disciplinary research collaborations between application domain experts and data scientists. Taking into consideration the vast amount of data that are spread over various libraries it is impossible for any single research group or single organization to assemble all this information. For that reason, e-collaboration platforms are created which can enhance such forms of data sharing by providing the relevant context, discussions, and annotations of the data in ways that add tremendous value to the end user. This can be considered as another strategy towards the acceleration of the rate at which new materials can be designed, manufactured, and deployed.

As a result, there is a need for the development and implementation of data-driven materials design protocols for objective decision support at various stages of materials development [9].

Another effort in similar direction can be considered the Materials Data Facility service that provides intuitive interfaces through which any researcher can access a growing set of advanced capabilities. The focus is on two services, data publication and data discovery, with features to promote open data sharing, self-service data publication and curation, and encourage data reuse, layered with powerful data discovery tools [10].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

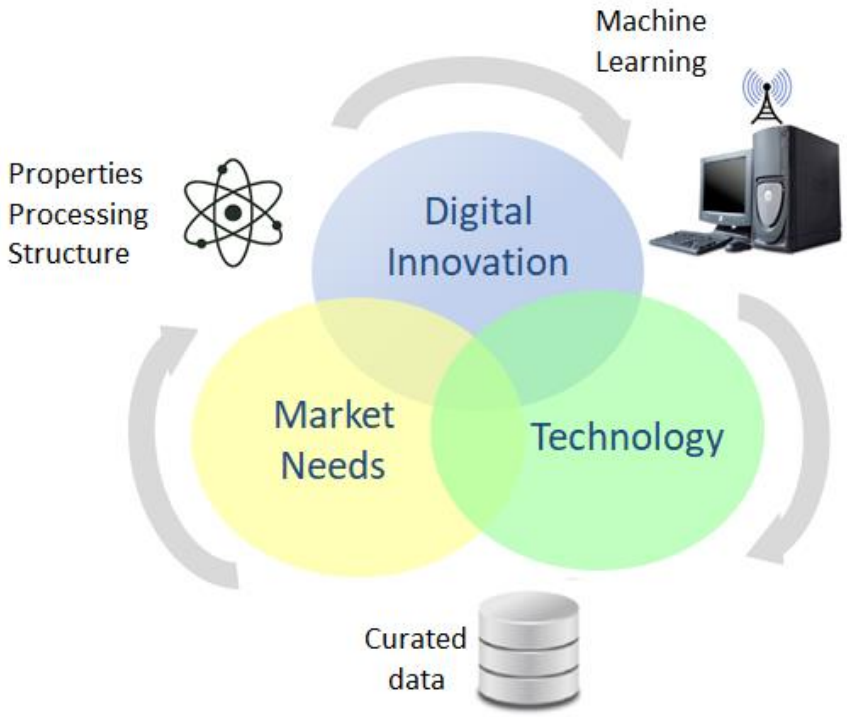


Figure 1: Interrelations between Digital Innovation – Technology – Market Needs [8]

For the above-mentioned reasons, materials science domain is on the verge of adopting data-driven discovery. Under this context, the goal of this paper is to present a novel approach for development of a systematic and holistic approach in order to collect, process and analyse trends in data and big data. In a future work, these trends can be used as a basis for predictions about materials properties, which can have significant impact on new materials' design.

In this work, a novel possible approach for classification (ontology) of materials characterization methods, based on the centrality of the measurement "probe" as a classification criterion, is presented. Within Metadata and Ontology in materials characterization, a novel concept and structure for data structuring in experimental materials characterization (called CHADA) is introduced, providing also a study case of Data Management Plan (DMP) for advanced (nanoscale) materials characterization and a specific nanoscale characterization method (nanoindentation testing). These can be integrated into an Open Innovation Environment, a digital platform that is being designed within the European Materials Characterisation Council (EMCC). Finally, we give our vision and opinion on how novel

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

approaches for characterisation data handling will bring innovation in the frame of big data towards realization of the "fourth paradigm" in materials science.

2. Classification of materials characterization methods (ontology)

Many definitions of ontologies exist [11, 12], but one that seems to be the best is based on the definition of Gruber [13]: An ontology is a formal, explicit specification of a shared conceptualization, where: *conceptualization* refers to an abstract model of some phenomenon in the world and to the relevant concepts of that phenomenon, *explicit* because the type of concepts used and the constraints on their use are explicitly defined and *formal* for machine readability.

The purpose of ontology is interrelated with its characteristics. First of all, they are comprised by vocabularies, which not only describe terms but also the relationships among the terms. One difference that can be mentioned between taxonomy and ontology is the set of relationships that are developed within the ontology and the fact that due to these relationships questions and queries can be answered. For example, if a material is a member of alloys and these alloys are members of a family with specific range of Young's modulus, then that alloy will share that range as well.

The ontology can be expanded to more characterization methods accordingly. As a result, it will be a domain ontology and not an upper ontology [14]. The concept will be to describe both data and metadata (data about data) of the experimental technique, thus being a subset of the domain knowledge about materials characterization.

For this paper, an ontology was created, having as the main basic classification criterion the physical probe that is used for the measurement. This is assumed to provide the most general framework for classification of experimental characterisation techniques, independently of the complexity of the material being tested. In this way, the class can be separated in three different sibling classes, which correspond to different experimental probing techniques: Mechanical Analysis, Chemical Analysis and Materials Structure Analysis. Our focus was on the Nanoindentation process, which is considered a subcategory of Mechanical Analysis. Furthermore, the material under

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

consideration was the PMAA. For that taxonomy, Figures 2 and 3 show the ontology class hierarchy and ontology diagram. The program Protégé, Versions 4.3.3 – 5.2.0 were used [15].

The characterisation techniques and the way they were categorized follow the analysis from the references [16] and [17], where the first one focuses on microscopic and spectroscopic methods, whereas the second one focuses on mechanical analysis. What it interests the most is the instances of the class raw metadata and that of the class property, where information about the experimental data of Nanoindentation and the results of Young's modulus and Hardness can be found. The same graph can be expanded in order to include more data and/or metadata related to the experimental data, either for different experimental techniques or for the same technique, yet for different materials.

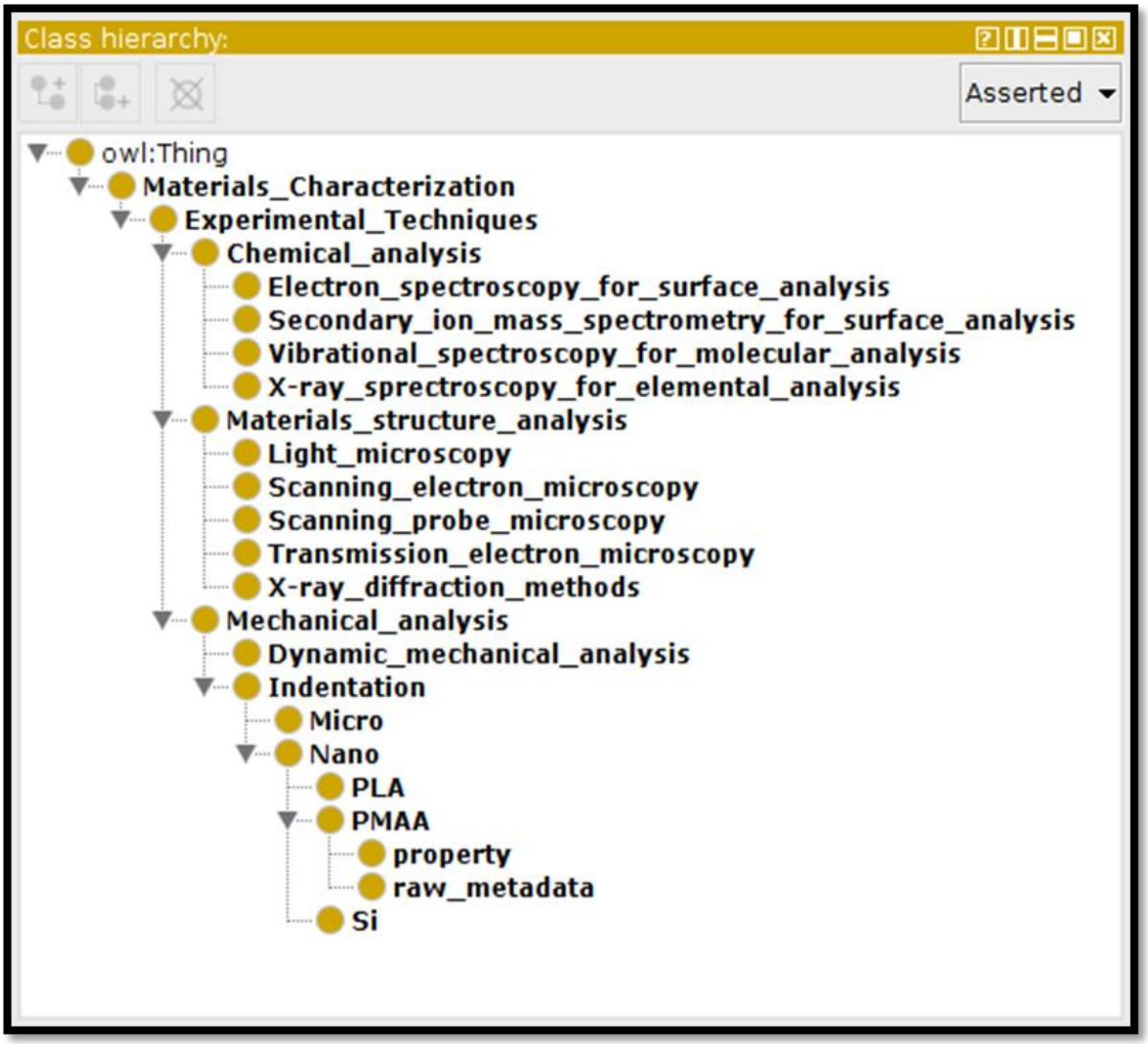


Figure 2: Ontology class hierarchy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

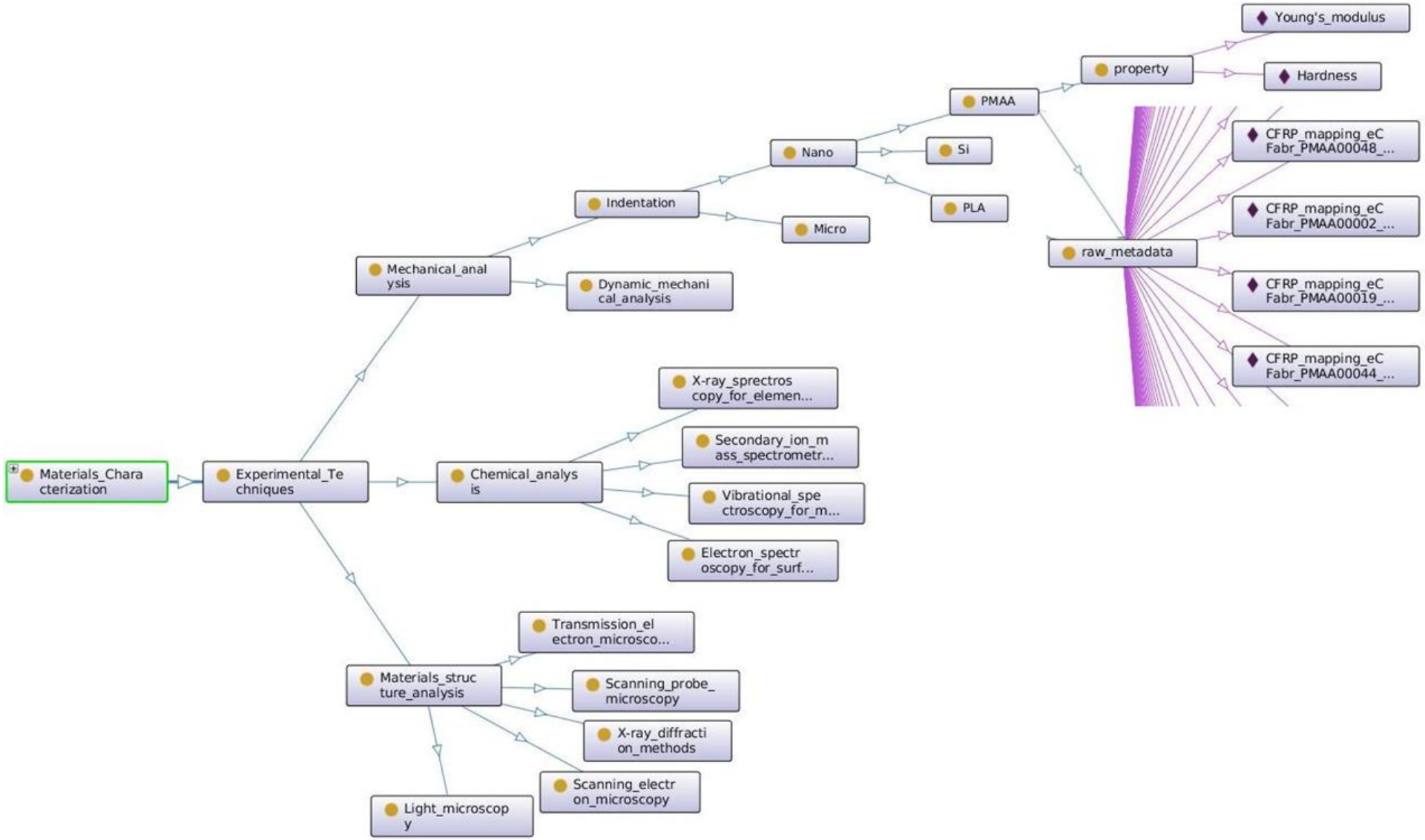


Figure 3: Ontology indicative example diagram for materials characterization methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

3. Metadata and ontology in materials characterisation

Once published, scientific data should remain available on the cloud and be used long after their publication. In that approach, the format and structure that is used for data storage is the critical factor to ensure traceability and reproducibility. To understand completely the data, the readers need the metadata, where information including the instruments used, the experimental protocols, the post-processing and finally the way and time that data were gathered, becomes available [18].

In relation to this, it is clear that storage and sharing of large amount of data requires strictly the definition of a standardized vocabulary and a standardized structure for the metadata.

The concept of this paper is that having a data management plan as a provision step, and utilizing its concepts in a specific experimental domain, this domain can be mapped using an **ontology**, which, in turn, can be utilized for extraction of **accumulated knowledge**, via discoverability and reuse [19].

With an initial data management plan, one can identify the different naming patterns of the files and in turn can homogenize this pattern. As a result, problems that may arise from ontology may be eliminated due to the avoidance of using spaces in-between string names of files. The whole concept, presented here, is to enhance the extraction of data in a fast and efficient way. Concepts like tuples, which are encountered in both databases and ontologies, will not be of primary concern, because this paper focuses on basic aggregation and accumulation of data, which can be thought of as a subdomain of an ontology about materials characterization, which as we are aware of, is in the process of creation. As an example, an n-tuple can be a set of n-objects in a specified order: material, alloy, specific combination of metals.

Here, we present a novel approach for the definition of terminology, classification and metadata for materials characterization methods, where the main purpose is to arrive to a standard structure (that we will call CHADA) for representing materials characterization data.

The first step towards this goal is the definition of the terminology associated to material characterization methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

We propose that only four types of concepts are used for the classifications of the different steps of an entire characterisation workflow (which can be simply called "characterisation"):

1. **Sample (or "user case")**, which represents **volume** of material, and the information on the surrounding **environment**, which interacts with the probe and generate a detectable (measurable) signal (information);
2. **Method**, which represent the **process** (or the sequence of processes) by which the metrological chain is defined; within a single method, the following fundamental elements are identified : user, probe, signal, detector, noise;
3. **Raw data,** Is the set of data that is given directly as output from the metrological chain, usually expressed as a function of time;

   Data processing, which represents any process (or sequence of processes) by which the data are analysed to arrive to the final shape.

By using this simplified approach, a generic characterization method can be presented by the following scheme (Figure 4), which can be used for the construction of the metadata structure of any material characterisation process.

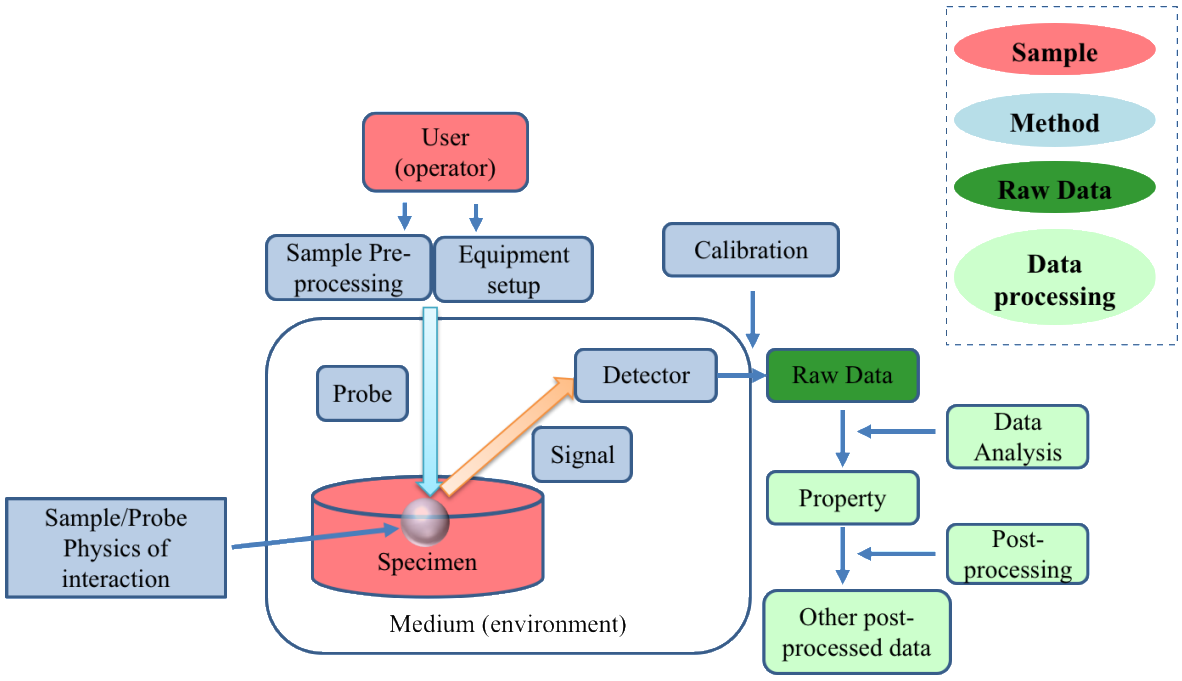


Figure 4. Visual representation of a characterisation experiment with keywords and colors

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Since all standardised characterization methodologies consists, in practice, of a well-defined sequence of items and actions, the same approach can be used to develop a generic workflow program (Figure 5), where a sequence of multiple-samples, multiple actions and multiple data processing steps could happen:

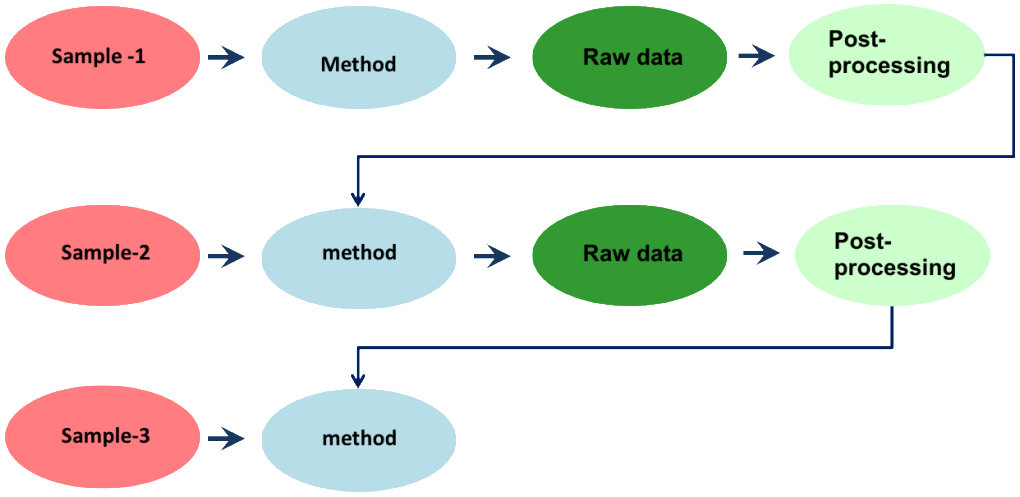


Figure 5. Visual representation of sequence of multiple-samples, multiple actions and multiple data processing steps

This scheme is also based on a similar concept developed within the European Materials Modelling Council (MODA, https://emmc.info/moda/), where the data from a generic model are represented according to the representations of User Case, Model, raw outputs and processed outputs.

To give an example, in the chapter n. 5 the CHADA workflow for nanoindentation testing is presented.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

4. Data Management for materials characterisation

4.1 Workflow management systems

Over the last few years, various workflow management systems have been formed in order to manage the accomplishment of different workflows on complex and heterogeneous computing resources. Scientific workflows are considered an important concept that controls data processing for the calculation of extended and complicated scientific applications. A number of workflow management systems have been developed with the main focus to automate the data management responsibilities and plans as well as the supply of the required resources 20].

4.2 Data Management

Data Management will ensure the accessibility of the created data from other users thus enabling data sharing. In addition, data protection will be enhanced because data will be preserved and curated at specific databases, which will reduce the risk of storage. Data Management is especially useful for increasing the work organization and productivity and enabling the easy retrieval of data, avoiding extra costs since duplication of data can be reduced or even eliminated. Finally, the creation of an open innovation environment the impact of publications can be increased due to the easy access not only to data but also to published work [21,22,23].

In this context lies the concept of making data FAIR, which means Findable, Accessible, Interoperable and Reusable. These principles were established in order to satisfy the need to improve the infrastructure supporting the reuse of scholarly data and should be applied not only to ‘data’ in the conventional sense, but also to algorithms, tools, and workflows that led to that data [3]. The data created can be of specific types and can be curated at specific repositories [24], however, not all datasets or even data types can be captured by, or submitted to these repositories. In response to this, there is the emergence of numerous general-purpose data repositories, at scales ranging from institutional, to open globally-scoped repositories such as Zenodo [25].

Humans and machines often face distinct barriers when attempting to find and process data on the Web, that both can be ineffective when big data are created and stored. On one hand, humans have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

an intuitive sense of 'semantics' (the meaning or intent of a digital object) and are unable to operate at the scope, scale, and speed necessitated by the scale of contemporary scientific data and complexity of e-Science. On the other hand, computational agents can undertake the discovery and integration tasks, to be capable of autonomously and appropriately acting when faced with big and different types of data [3].

4.3 Data Management Plan

A Data Management Plan (DMP) should include information about the handling of data during and after the end of the project. It is responsible for all the data created, collected and processed, providing information about which methodologies and standards will be applied, further information about how data will be shared and how they will be preserved and curated.

Towards the direction of data management both E.U. and U.S.A. have made progress. E.U. with the Horizon 2020 initiative requires a DMP for all projects, which participate under that framework, while The National Science Foundation in the United States now requires an explicit data management plan in all proposals [26]. In the UK the Jisc-funded Digital Curation Centre (DCC), in order to assist UK HEIs in improving their capacity for research data management and sharing, produced DMP online, which is the first tool to assist in the data management planning process [27].

Research Data Management (RDM) offers opportunities and challenges at the interface of library support and researcher needs. Libraries are in a position of balancing the capacity to provide support at the point of need while also implementing training for subject liaison librarians grounded in the practical issues and realities facing researchers and their institutions. The North Carolina State University (NCSU) Libraries has deployed a DMP Review service managed by a committee of librarians. A training ground model is established, which aims to develop needed competencies and support researchers through relevant services and partnerships. Library support for data management is attractive because it offers an avenue for building collaborative networks, integrating library support into the research process, and supporting open access to research data [28].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Along with this initiative, the European Commission released a document providing general guidelines about the creation of DMP [29]. Another attempt is the DMP plan created for the OYSTER project of Horizon 2020 program (www.oyster-project.eu). The steps followed for the creation of a spreadsheet version of DMP were the following:

1) the E.U. DMP guidelines were transferred to a spreadsheet adjusted accordingly for the OYSTER project.

2) Then there was an effort to reduce unnecessary writing from partners by providing lists with options for them to select.

3) In case the lists provided were not exhaustive, the partners had the option to add information – the information added was adjusted accordingly and finally implemented to the new version of DMP, something that is a characteristic of a living document, which is constantly updated**.**

4) The information gathered can be easily handled, since the information provided is specific, in part strictly organized, and not written in free format compared to a document form. In this paper, we present the section of DMP, which is suitable for the experimental technique of nanoindentation of materials, which can be seen in Figure 6.

The entire DMP is dsgined to be fully compliant with the new CHADA schemes that were presented in the chapter 3. In the following figure, the DMP section corresponding to “method” description in the CHADA are shown, as an example.

METHOD				
This section refers to the Methodologies followed				
	Data Origin	Define and describe the origin/source of your data. Data can be gathered from different sources.		
	Observational		Data captured in real time - often not reproducible i.e. sensor readings, images, telemetries, sample data...	
	Experimental		Data from lab equipment, often reproducible, but with high costs - i.e. chromatograms, magnetic fields readings...	
	Simulation		Data generated by computational models where model and metadata are equally important to output data - i.e. climate models, economic models, materials models,...	
	Method	Define and describe the scientific method used for this Dataset.		
	Physics/Chemistry of Interaction			Example: Detection of the surface by the tip (stiffness triggering value based) - Penetration of the tip inside the sample using prescribed load function - Hold of the maximum load (or the load for the prescribed depth - unloading of the tip by steps - tip removal from the sample.
	Discipline			e.g. Characterisation - Nanoindentation

Figure 6: Part of DMP corresponding to “method” description .

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

5. Case study: High-Speed Nanoindentation

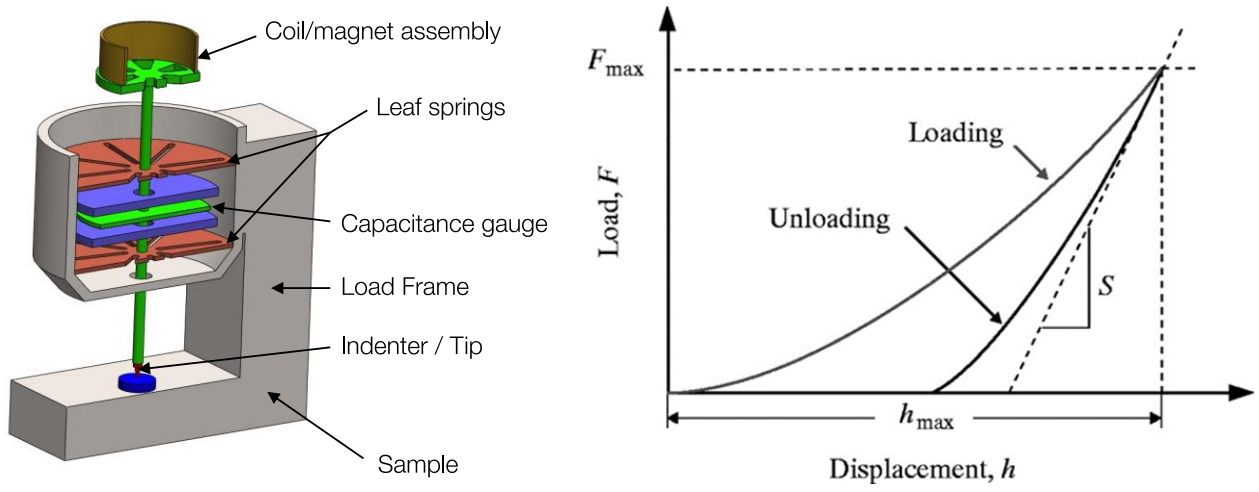
Nanoindentation is a widely used technique for the measurement of hardness and elastic modulus and which has become ubiquitous for mechanical properties at surfaces [30].

The method that was first introduced in 1992 [31] has widely been adopted and used in the characterization of mechanical behavior, in particular hardness and elastic modulus, of materials at small scales [32].

The experimental procedure always involve the realization, in parallel with the main sample, of a series of calibration experiments on a fused Quartz reference sample to quantify the frame stiffness and the area function of the adopted indenter. Then, the Oliver-Pharr method is usually adopted to analyse the loading and unloading curves to extract hardness and elastic modulus as the main outputs of the method [31].

The procedure has become a primary technique for determining the mechanical properties of thin films and small structural features. Films with characteristic dimensions of the order of 1 μm are now routinely measured, and with good technique, the method can be used to characterize, at least in a comparative sense, the properties of films as thin as a few nanometers.

The main features of this method, including hardware description and a typical load-displacement curve, are shown in Figure 7:



(a) (b)

Figure 7. Basic instrumentation and output of nanoindentation technique.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Recently, high-speed nanoindentation is emerging, as testing real time and in-line / online is employed to obtain massive data sets (i.e., big data) on the load-depth response of materials, samples and (also intermediate) products towards quality assurance and rapid manufacturing characterisation of test specimens.

An example is reported in Figure 8, where a highly-heterogeneous Li-ion battery cathode composite, where some thousands of nanoindentation experiments were realized over a relevant area. In this way, original and richer information can be gained on the nanomechanical property distributions, as a function of the state-of-charge of the battery.

This is a clear example on how large amount of data (acquired at higher acquisition speed with high-throughput characterisation methods) can be extremely relevant to gain further insights into the process-structure-property correlations of highly heterogeneous materials.

Advanced statistical analysis of such data can, therefore, enable to develop novel design rules for the production of innovative materials with improved performance and enhanced lifetime.

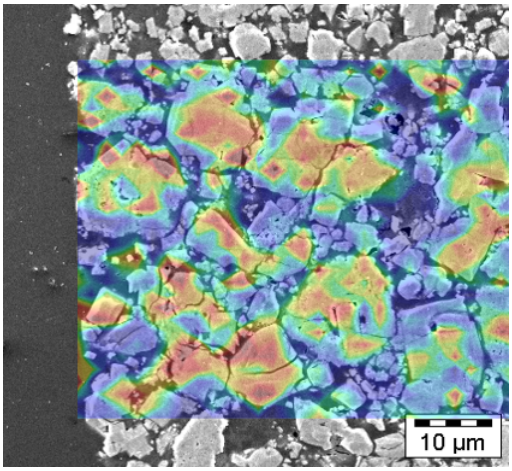


Figure 8. Example of the overlap between the microstructure of a battery composite (SEM image) and the corresponding high-speed nanoindentation map of the elastic modulus.

By using the concepts and basic structure of CHADA described in the previous chapter, the following scheme can be developed for this specific technique (Figure 9).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Using this approach, each of the block in the workflow represents a set of information that is stored in the metadata, and can be retrieved at any time to ensure traceability of the data.

In this way, and in addition to the traditional approach of storing only the calibrated load-displacement curve, the metadata will contain all the information on the sample, user, environmental conditions, calibration procedure and related data, raw data, analysis process and finally the analysed data.

The adopted classification by only four main classes (user case, method, raw data, post-processing) ensures that any characterisation technique can be represented by a simple sequence of standardized elements.

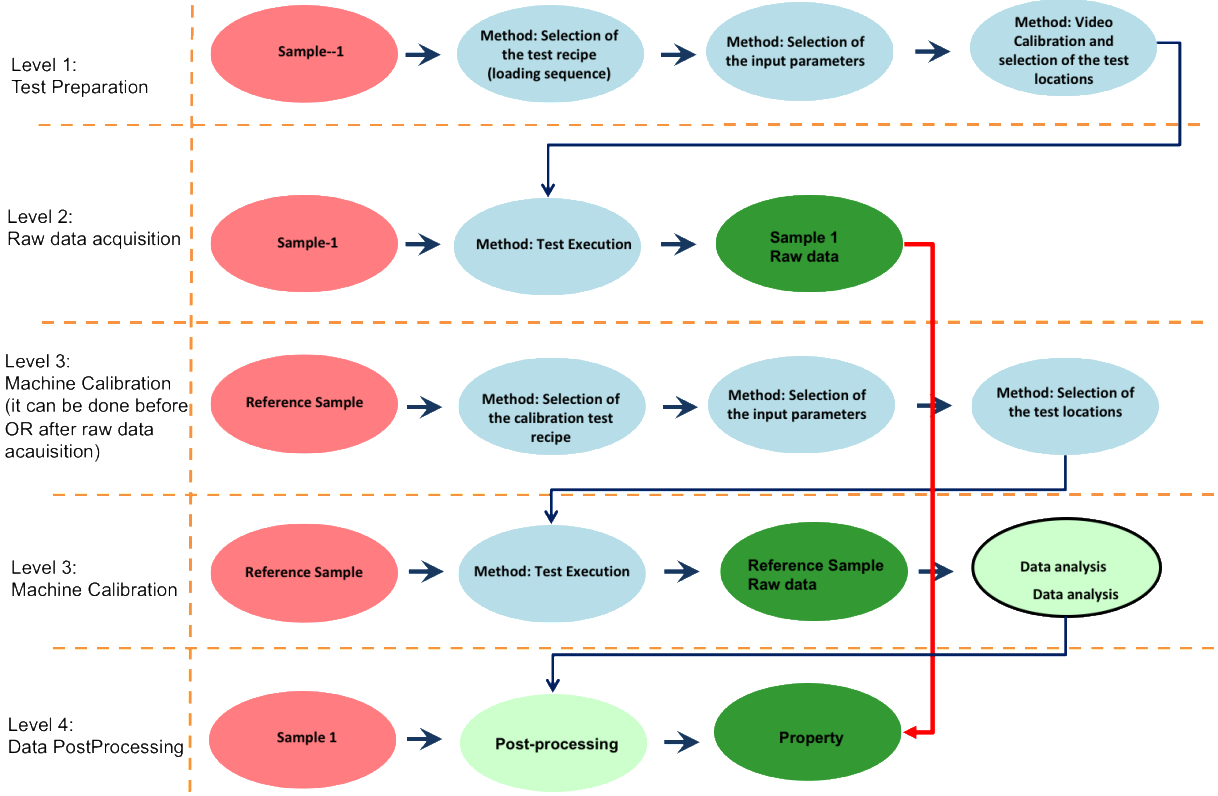


Figure 9. CHADA workflow for nanoindentation

Then, a much more detailed metadata can be built up, if needed, by filling each of the CHADA element with internal information and attributes (e.g. a full description of the adopted calibration recipe, or a full description of the sample preparation procedure, etc. etc.).

An example of compiled metadata, again from nanoindentation, is reported in the following table:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Table 1. Compiled metadata for nanoindentation, according to the new CHADA classifications

Keyword	Description
User (operator)	Human Operator (different levels of automation are available)
User case (sample specifications)	Sample dimensions (example: 1-inch diam. – 5 mm thick.). Surface flat and polished. Sample embedding on sample-holder (hot glue or acrylic glue). Optical sample surface alignment with reference sample (SiO2) surface.
Specimen	Bulk material, coatings, heterogeneous materials, and biomaterial.
Medium (environment)	Equipment Box: Air, Temperature, Pressure, Humidity, Noise, Vibrations (Acoustic or mechanical).
Sample/Probe Physics of interaction	Detection of the surface by the tip (stiffness triggering value based) – Penetration of the tip inside the sample using a prescribed load function – Hold of the maximum load (or the load for the prescribed depth – unloading of the tip by steps – tip removal from the sample.
Equipment setup	Optical alignment of the sample. Method selection and Input parameters for the test (Sample Poisson's Ratio, Prescribed Depth or Load, number of tests, locations of the tests, Engage options).
Calibration	Standard CSM tests on reference sample.
Probe	Selected Tip for the test (Berkovich, Cube Corner, Flat Punch..).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Detector	Electronic controllers and capacitive gauges.
Signal	Electrical current in a coil -> Force (Load)
Raw Data	Raw channels plots (Raw Load, Raw displacement, Dynamic Stiffness,…).
Data Analysis	Check of the surface detection, check of the Load vs depth quadratic curve trend, check of the slope of the unloading curve, removal of the not relevant tests.
Raw Data Analysis	Application of the Oliver-Pharr method (or other data analysis methods)
Post-processing	Raw data calibration using tests on reference sample, check of the results (see data analysis), Selection of the load (or depth) range to evaluate the mechanical properties, Graphs or histograms of interest.
Properties (elaborated data)	Elastic Modulus, Hardness, Yield Stress, Residual Stress, Creep parameters, Fracture toughness, mechanical maps, etc.

6. Open Innovation Environment as Materials' Ecosystem

According to the innovation system's theory, ideas, devices or processes are the carriers of innovation, because of multiple sets of relationships among participants in a system, such as companies, universities and research institutes [33]. The efficient combination of professional experience and the exchange of technology and knowledge among stakeholders is a dominant decisive factor of whether the methods tend to be innovative [34]. The challenge of innovation drives technology towards market's needs more quickly. Innovation necessitates not only experimentation,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

coming from a wide variety of technologies, but also access to a wide spectrum of possible service providers and users, even from the early stages of its development. The challenge of bringing innovative partners together is to exceed the potential of each partner coming from singular sector of firms. As a result, the scientific community and European policymakers are interested to establish, maintain or strengthen experimentation facilities and platforms as fundamental means and tools to support broadband innovation [35]; a schematic figure of Open Innovation Environment function is provided below (Figure 10).

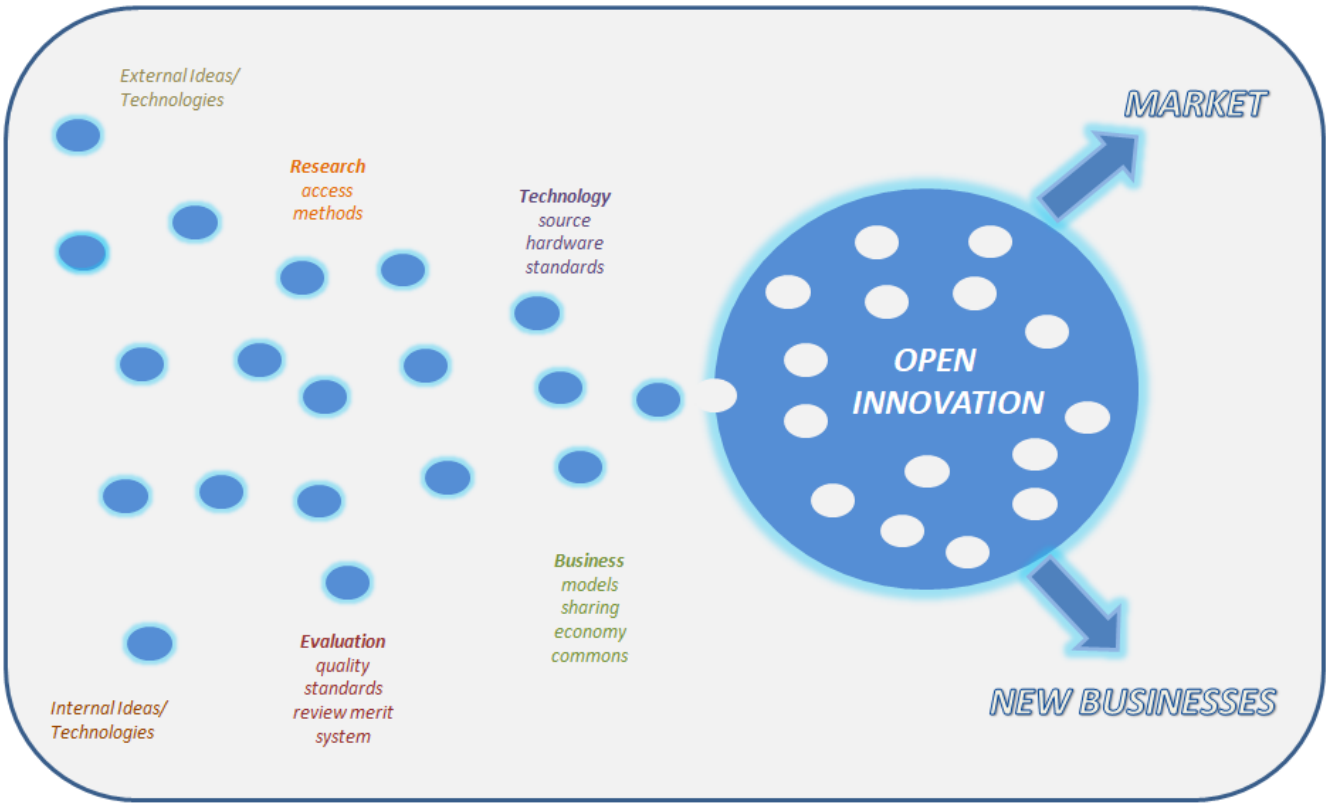


Figure 10: a schematic figure of Open Innovation Environment function is provided below

Based on the foregoing, it becomes evident that in today's rapidly growing business ecosystem, SME's (Small Medium Enterprise), large enterprises, and institutes need to merge their expertise and knowledge with others. An Open Innovative System (Figure 8) appears to be a candidate approach for this, as it supports the innovation capabilities of its members. In order to support the objective to establish a transnational Open Innovation Environment (OIE), for specific characterisation methods, some steps for innovative developed technologies are followed, such as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

data and meta-data management, the use of ontologies and advanced analysis, sharing, interpretation of experiments and proposed models. The OIE stores information about data, including their corresponding metadata. The handling of metadata is of crucial importance in order to facilitate the access to the real data. The effective management of data and metadata is one of the most important activities of data scientists, within a governance practice, enabling data management policy and access to information.

7. Big data towards realization of the "fourth paradigm" in materials science

In recent years, the scale of data generated and shared by academia, industry, businesses and public administrations has vastly increased. Handling and mining big data opens new horizons in productivity growth and consumer impetus, indicating that the era of big data has already been started, a phenomenon also referred to as the Data Deluge [2]. The fact that big data has entered into every area of today's industry and business functions and has become an important factor in production becomes obvious if we consider that every day the world produces around 2.5 billion gigabytes of data [36]. Gantz and Reinsel [37] assert that by 2020, over 40 trillion gigabytes of complex and heterogeneous data will be generated. Big data analytics is the process of researching into massive amounts of data and revealing hidden patterns and possible correlations [38].

With the coming of "big data" era, numerous efforts have been made, in the field of materials science, in order (a) to develop new methods to overcome the deficiencies of these two common methods and (b) to collect large datasets of materials properties in order to provide a powerful impetus to accelerate materials discovery and design.

The term big data implies escalation relevant increase in the amount of data, but it also results in a qualitative conversion regarding the way that we store and analyze such data. The exact definition about ''big data' is not universally accepted, as there is still a lot of confusion about what it actually means, while its size is only one of several dimensions of big data. The concept is constantly evolving and reconsidered, but it remains the driving force behind many forms of digital transformation,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

artificial intelligence, data science and the Internet of Things [38]. Big data can be regarded as a connection and integration of the physical world, the human society, and cyberspace.[40] They can be classified into two categories; the first one includes data which is usually obtained from scientific experiments, simulations, algorithms or mathematical modeling techniques and observations, and the second one includes data from human society coming from domains such as social networks, health, economics etc..

In order to identify a common framework to describe big data, Laney suggested that Volume, Variety, and Velocity (or the Three V's – Figure 9) are the three dimensions of challenges in data management [41]. We describe the Three V's below (Figure 11).

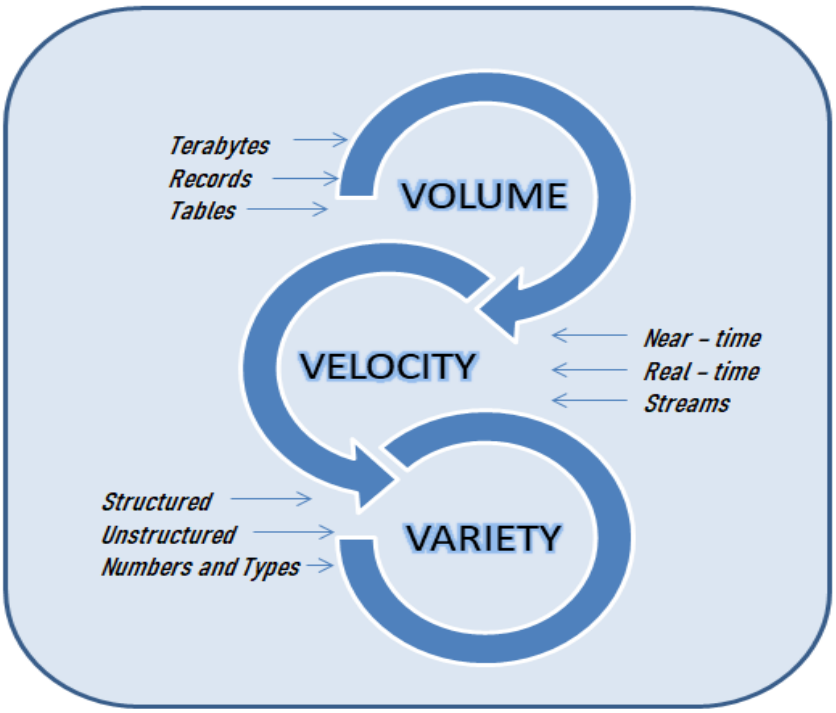


Figure 11*:* The Three V's

Volume: Defines the size of data and is relative and varies by factors, such as time and the type of data.

Velocity: This factor deals with the frequency that data are generated, for example, every millisecond, second, minute, hour, day, week, month or year. We can identify three main categories regarding data procession: occasional, frequent, and real-time [42].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Variety: With increasing volume and velocity comes increasing variety. Variety relates to the types of data, for example photos, videos and audio recordings, text documents books, email messages, presentations, e.t.c..

Data scientists always describe "big data" as having at least the three dimensions, analyzed above: volume, velocity, and variety. In addition, there are two more Vs completing the list, which include variability and value [43].

Additional Vs

Variability: Big data velocity is not consistent and has periodic peaks or troughs. In that way, variability refers to the variation in the data flow rates.

Value: It is of crucial importance to Value is introduced as a defining attribute of big data, to ensure that the outcomes that are generated are based on accurate data and can lead to measurable enhancements.

The scale and complexity of big data necessitate a change in computing paradigm, regarding data's structure procession. Along with the growth of big data, an evolution of databases has turned them into a "non-relational" form. Nowadays, big data typically contains data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a reasonable time.

In order to fully take advantage of the potential of big data, the challenges concern the characteristics of big data, the existing analytical methods and proposed models or the limitations coming from current data processing systems [44]. Many studies concerning big data challenges have been focused on the difficulties of understanding the notion of big data, on deciding the kind of data that will be generated and collected, on privacy issues and ethical considerations as far as the mining process is concerned [45].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

7.1 Challenges on data complexity

The complexity of big data that include complex types and structures makes its understanding, mining tasks, representation and computation more and more challenging. One of the major challenges is to understand, establish and then predict the relationships between data complexity and computational complexity, in the frame of domain-oriented processing methods. It is of high importance to design and evaluate computational models in order to quantitatively define data's complexity, in an efficient way and to develop the principles for dealing with big data into a solid foundation. As a result, in order to simplify big data's representation and facilitate the design of complex computing models and algorithms, it is needed to establish models of data sharing via multi modal interrelationships [45].

7.2 Challenges on computational complexity

In order to lever large amounts of data, new approaches will be needed, while customary computational methods do not successfully support processing, analysis and mining techniques. These approaches and techniques should be based on:

- an independent distribution of the data produced
- new procedures for creating statistical rates
- reviewing the reliability and complexity of the data
- reviewing of data algorithms by proposing new theoretical bases
- Supporting value-based applications in specific areas.

The challenge is to address the computational complexity of big data, with focus not only on developing novel algorithms for continuous computing but also on the development of a large data-driven computing framework with the aim to optimize communication, storage and sharing. Hence, there is a need for further exploration and expansion of modeling methods for categorization and reduction of large data, satisfying the demand on data's high value and velocity [46].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

7.3 Challenges on communication

Machine learning techniques should be developed with a focus on the communication costs, while sometimes these costs are a major concern compared to the processing cost of big data. The goal is to minimize that communication cost while satisfying the additional data storage requirements..

7.4 Challenges on system complexity

Main request is the design, implementation, testing, and optimization of big data processing systems and computing frameworks with a high data recovery throughput, low energy consumption, and highly practical computing. These demands pose new challenges to the configuration of system structures, computing frameworks, and processing systems, while their possible solutions will pose an essential basis for developing hardware and software system constructions with energy optimized and efficient allocated storage. A fundamental research has to be conducted regarding the correlation between complexity and computability of big data demands. Also, there is a need for a measurement of the variety of energy efficiency factors such as system performance and processing capabilities at the same time [47].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

8. Conclusion and Future direction

Materials discovery lies at the heart of human progress and milestones of human progress are related with materials: new materials with high and unprecedented functions and properties, along with understanding their relationship with chemical constitution. There are continued efforts to deploy a minimum amount of materials for a given function, which leads concentrating on nanostructured materials. In addition, there is an increasing effort of reducing costs, risks of experiments and the ability to create better materials for specific purposes in shorter amount of time. The first steps towards these directions were the experimental techniques along with theory. Then, simulation advancements reduced the number of unnecessary experiments and costs, yet, at the same time, the over increasing creation of data was posing an issue.

Materials data management eases the efficient mining and potential for further processing of large materials data sets, resulting in the extraction and identification of high-value materials knowledge, towards design and manufacturing. This is accomplished by using linkages of process-structure-property (PSP) information, with the main focus of data transformations to be in the forward direction (process → structure → properties). As therefore high-value information requires to be linked with the manufacturing and product design routes, the main challenge is, starting from a proper data management plan, to design and build the needed databases stems (tackling challenging issues such as rich internal materials structures that span multiple length scales).

Data Management foresight in materials' advanced characterisation mitigates the inherent risk largely, not only by making decisions more concrete (e.g. in design and manufacturing), but also by capturing failures and successes; information from this is then useful and processable to and from other disciplines. For an effective mitigation plan based on data management is strongly based on the availability of data and the use of data-driven protocols, as the uncertainty associated with the information and knowledge used in making decisions (in materials development workflows) is then quantifiable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

Despite the difficulties appearing due to highly localization (in terms of specialization) distributed in terms of organizations and/or geography, data management and data science build upon cross-disciplinary expertise (e.g. multimodal measurements, multiphysics simulations and materials phenomena descriptors) and provide the essential tools to ignite and boost such collaborations.

Standards, terminology, digitization and automation are few of the requirements to reach process scalability (digital workflow recording, based on standardization and automation). Overall, in order to achieve the desired acceleration of materials development of proper design and at an affordable cost, data management foresight in materials' advanced characterisation is the first and crucial step to begin identification of best practices and implementation.

Within this complex framework, we have shown in this paper how data management, materials informatics and digitalization for advanced materials characterization can be a Key Enabling Technology for introducing groundbreaking innovations in the manufacturing industry.

Yet, the prime factors contributing are the time required to switch the materials research practice to new paradigms, expertise on big data analytics and development of machine learning algorithms [48]. Information industry is directly related to big data and big data is a strong impetus to the next generation of IT industry. An emerging discipline, which is data science, employs various techniques and theories from many fields, including signal processing, probability theory, machine learning, statistical learning, computer programming, data engineering, pattern recognition, visualization, uncertainty modeling, data warehousing, and high performance computing.

Traditional data analysis and mining tasks, such as retrieval, topic discovery, semantic analysis, and sentiment analysis, become extremely difficult when using big data. At present, we do not have a good understanding on addressing the complexity of big data [49]. A step towards the handling of such big amount of information can be the creation of ontologies for specific domains.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2019 doi:10.20944/preprints201903.0205.v2

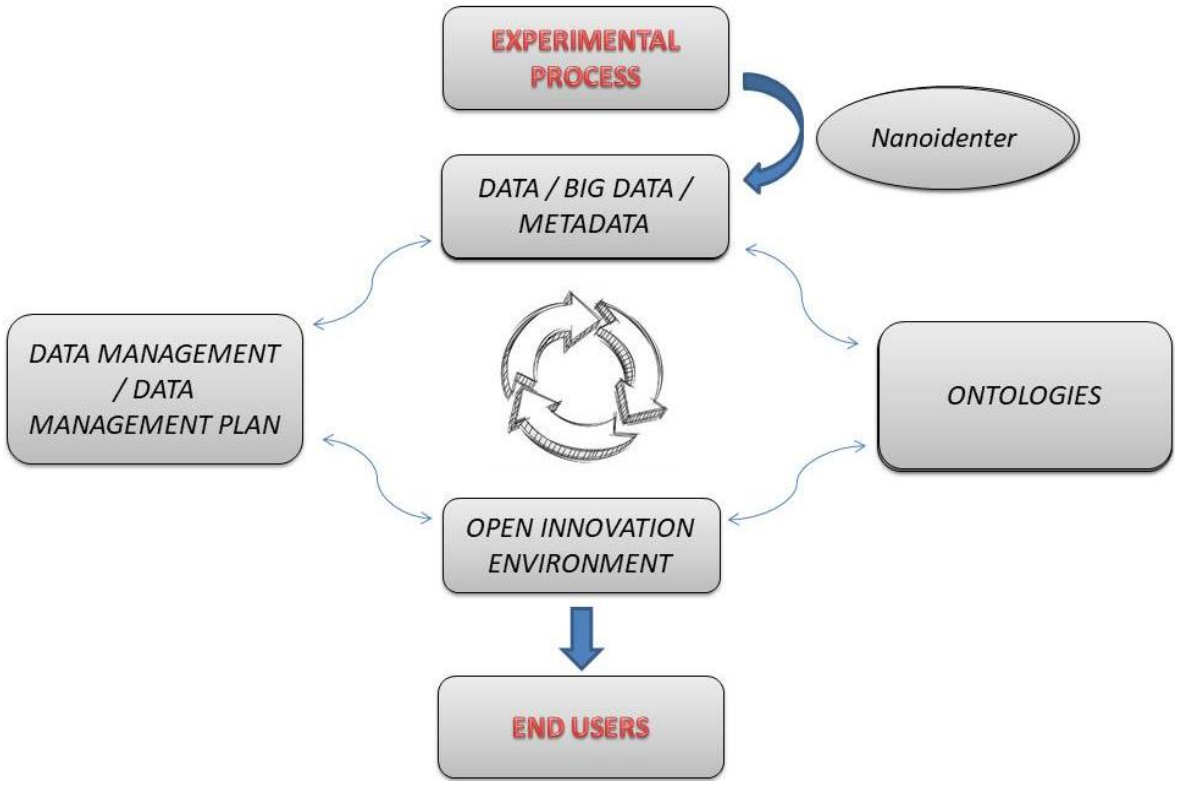


Figure 12: Interaction among Data, Ontologies, DMP and OIE – specific example for nanoindentation testing as main adopted characterization method

In this paper, we have shown hot the application of such innovations to materials characterisation is deeply underpinned with a series of required international cooperative actions, (schematic representation Figure 12) namely:

(1) A coordinated action for development of an ontology and a classification for materials characterization methods;

(2) A novel concept and structure for data structuring in experimental materials characterization (called CHADA);

(3) Advanced and standardized Data Management Plans (DMP) for each method and project; and

(4) Open Innovation Environment (OIE) platforms to behave and the engine for the integration and harmonization of all relevant aspects into one single-entry-point for driving innovation of large industries and SME's.

Coordination and Support actions in Europe (as well as worldwide) are needed to pursue such goals in the most efficient and effective way.

 doi:10.20944/preprints201903.0205.v2

Acknowledgements

All authors equally contributed to this work.

The paper (all authors) received funding from the European Union, within the large collaborative project OYSTER, Grant Agreement No. 760827, www.oyster-project.eu . The entire oyster project team should also be acknowledged for the fruit friutiful discussions and support.

The authors sincerely acknowledge Anne DEBAAS and Jorge COSTA DANTAS FARIA (both from European Commission, DG Research and Innovation, Unit D3 - Advanced Materials and Nanotechnologies), for the very useful and fruitful discussion during development of the novel CHADA structure and DMP development.

References

[1] Böhmer A.I., Lindemann Ud., OPEN INNOVATION ECOSYSTEM: TOWARDS COLLABORATIVE INNOVATION, 2015, ICED15

[2] Muhammad M.K. et al, Critical analysis of Big Data challenges and analytical methods, 2017, J. of Bus. Res., 70, 263-286

[3] Wilkinson M.D., Dumontier M., Aalbersberg I.J. et al, The FAIR Guiding Principles for scientific data management and stewardship, 2016, Sci Data, 3, 160018

[4] Executive office of the president of the USA, Materials Genome Initiative for Global Competitiveness, 2011

[5] Hai-Qing Lin,Boosting computational capabilities, 2016, Nature Materials, 15, 693–694

[6] National Research Council, Integrated Computational Materials Engineering.2008, The National Academies Press.

[7] Hill J., Mannodi-Kanakkithodi A., Ramprasad R., Meredig B.,Materials Data Infrastructure and Materials Informatics, 2018, Computational Materials System Design, D. Shin, J. Saal (eds.), 193-225

[8] Biernacki JJ, Bullard JW, Sant G, et al, Cements in the 21st century: challenges, perspectives, and opportunities,2017, *J Am Ceram Soc.*, 100, 2746-2773

[9] Kalidindi SR, De Graef M, Materials Data Science: Current Status and Future Outlook,2015, Annu. Rev. Mater. Res.,45, 171–93

[10] Blaiszik, B., Chard, K., Pruyne, J. et al, The Materials Data Facility: Data Services to Advance Materials Science Research,2016, JOM, 68, 2045-2052.

[11] J. Hendler, Agents and the Semantic Web,2001, IEEE Intelligent Systems, 16(2), 30–37,

[12] W. Swartout and A. Tate., Ontologies,1999, IEEE Intelligent Systems, 14(1),18–19

[13] T. R. Gruber, A Translation Approach to Portable Ontologies, 1993, Knowledge Acquisition, 5(2), 199–220

[14] B. Chandrasekaran, J.R. Josephson, and V.R. Benjamins, Ontologies: What are they? why do we need them?, 1999, IEEE Intelligent Systems and Their Applications- Special Issue on Ontologies, 14(1), 20–26

[15] Musen M.A., The Protégé project: A look back and a look forward, 2015, AI Matters., 1(4), 4-12

[16] Y. Leng, Materials Characterization, 2008, ed. Wiley, Singapore

[17] Zhi-Qiang Feng et al. Handbook of Nanophysics, 2010, ed. K Sattler

[18] Johnston L.R., Curating Reasearch Data, 2017, 1, Chicago

[19] Pryor, G. (Ed.), Managing research data, 2012, London: Facet.

[20] da Silva R.F., Filgueira R., Pietri L., Jiang M., Sakellariou R., Deelman E., A characterization of workflow management systems for extreme-scale applications,2017, Future Gener. Comput. Syst., 75, 228-238

[21] Guidelines for effective DMP, 2012, Michigan

[22] E.C. Guidelines in data management in Horizon 2020, 2016

[23] Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, 2017

[24] H.M. Berman, K. Henrick, H. Nakamura,Announcing the worldwide Protein Data Bank, 2003, Nature Structural Biology, 10 (12), 980.

[25] www.zenodo.org

[26] Reichman, O. J, Jones, M. B, & Schildhauer, M. P. , Challenges and Opportunities of Open Data in Ecology. 2011, Science, 331(6018), 703-705.

[27] Davidson, J., Jones, S., Molloy, L. and Kejser, U.B., Emerging good practice in managing research data and research information in UK Universities, 2014., Procedia Computer Science, 33, 215-222.

 doi:10.20944/preprints201903.0205.v2

[28] Davis, H. M., & Cross, W. M.Using a Data Management Plan Review Service as a Training Ground for Librarians, 2015, Journal of Librarianship and Scholarly Communication, 3(2), eP1243.

[29] https://emmc.info/emmc-info-data-management-plan-template-dataset-description/

[30] C.A. Schuh, Nanoindentation studies of materials, 2006, Mater Today, 9 (5), 32-40

[31] Oliver, W. C., and Pharr, G. M., J., An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments, 1992, Mater. Res. , 7, 1564-1583

[32] Oliver, W. C., and Pharr, G. M., Measurement of hardness and elastic modulus by instrumented indentation: Advances in understanding and refinements to methodology, 2004, J. Mater. Res.,19, 3-20

[33] Adner R., Match Your Innovation Strategy to Your Innovation Ecosystem, 2006, Harvard Business Review

[34] Kirner E., Kinkel S. Jaeger A., Research Policy Innovation paths and the innovation performance of low-technology firms—An empirical analysis of German industry, 2009, Research Policy, 38, 447-458