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

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

Abstract

This paper describes a novel methodology of data documentation in materials characterisation, which has as starting point the creation and usage of any 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 experimental assessment of mechanical properties on a small scale.

The new documentation structure for characterisation data (CHADA) is based on the definition of (i) sample, (ii) method, (iii) raw data and (iv) data analysis as the main component of the metadata associated to any characterisation experiment. In this way, the relevant information can be stored

inside the metadata associated to the experiment. 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.

Other fundamental aspects are reviewed in the paper, including the taxonomy and curation of data, the creation of ontology and classification of characterisation 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.

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 analyzed 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 (DMPs) 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 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 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 knowledge base 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].

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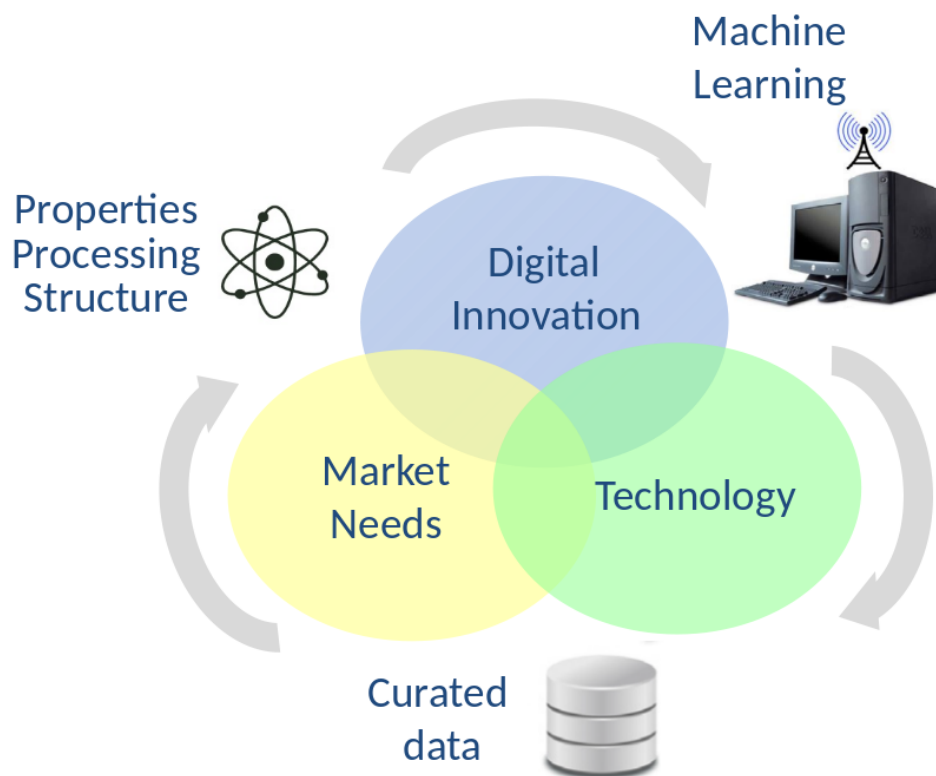
Figure 1: Interrelations between Digital Innovation – Technology – Market Needs [8]

In order to face the latest challenges and opportunity given by the fourth industrial revolution, it seems clear that one cornerstone will be represented by the development of standard procedures for the classification of data, the classification of experimental protocols and the definition of standard taxonomies and ontologies for data representations and linking.

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 analyze 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 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 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 – 5.5.0 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.
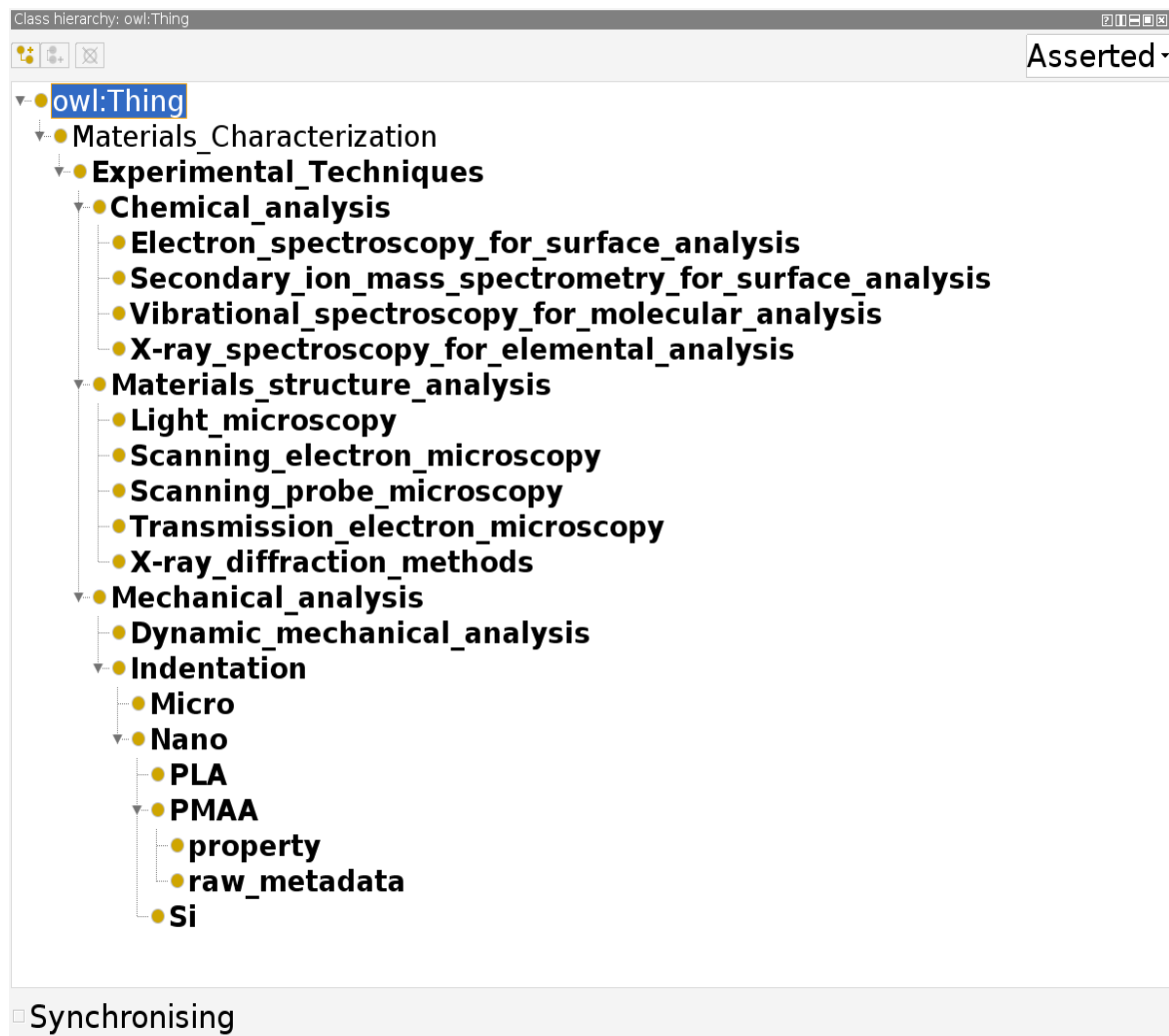
Figure 2: Ontology class hierarchy
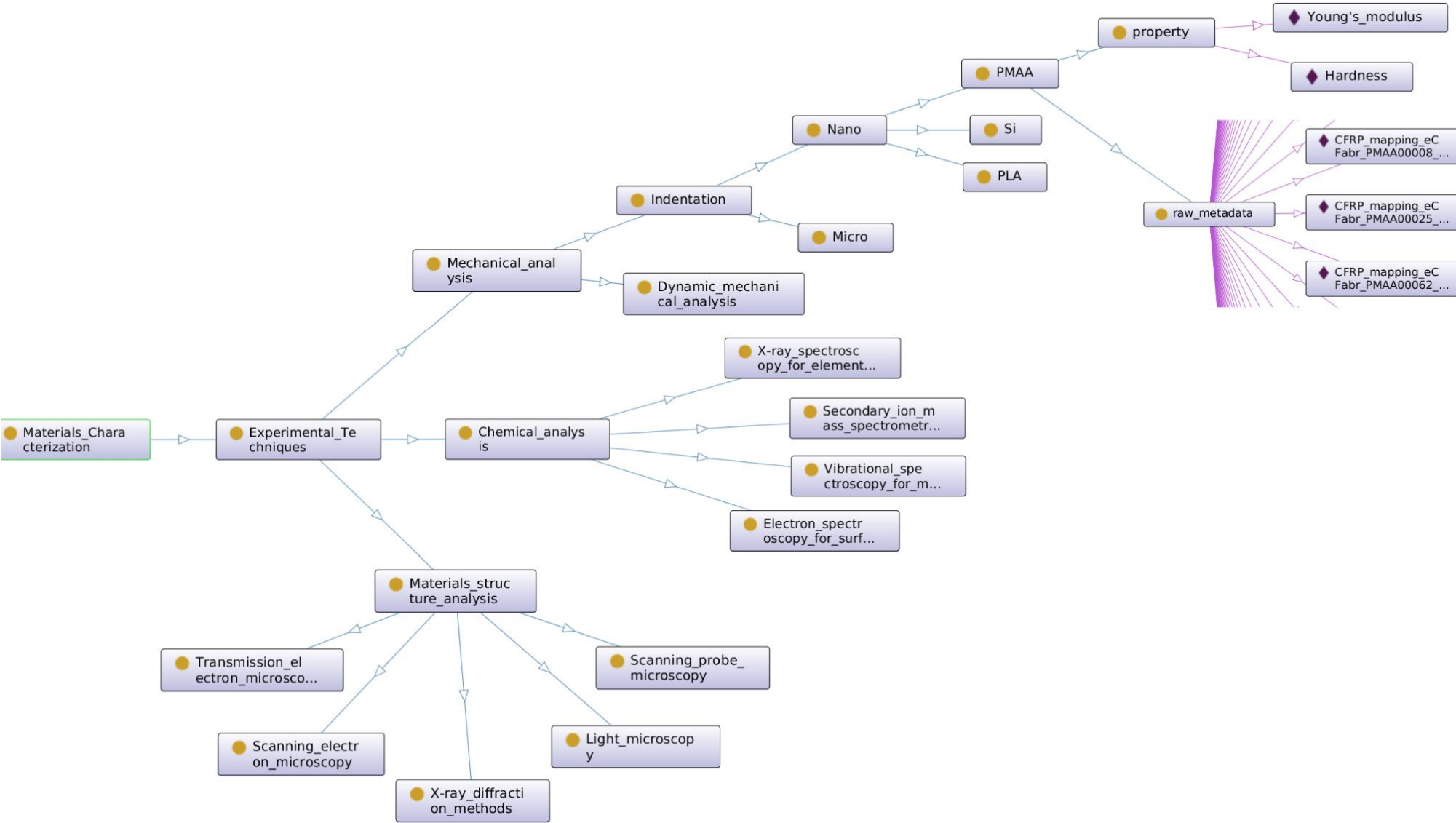
Figure 3: Ontology indicative example diagram for materials characterization methods.

3. A new concept for data documentation in materials characterisation: the CHADA

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 structured data documentation 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]. By using a standard and reproducible architecture of the metadata associated to a generic characterisation experiment, the experiment itself can be reproduced at any time with maximum detail and, even more importantly, the quality of the data and its reusability can be maintained during the years. The whole concept, presented here, is to provide a new structured format, which could allow to include in the metadata all the information necessary (and sufficient) to ensure reproducibility of the experiment itself, reusability of the data and certification of its quality.

To this aim, 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 materials' characterization methods. 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 *probed material*, and the information on the surrounding *environment*, which interacts with the probe and generate a detectable (measurable) *signal* (information);

2. **Method**, which represents 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;

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

By using this simplified approach, a generic characterization method can be documented by the following scheme (Figure 4), which can be used for the construction of the metadata structure of any generic material's characterisation process.
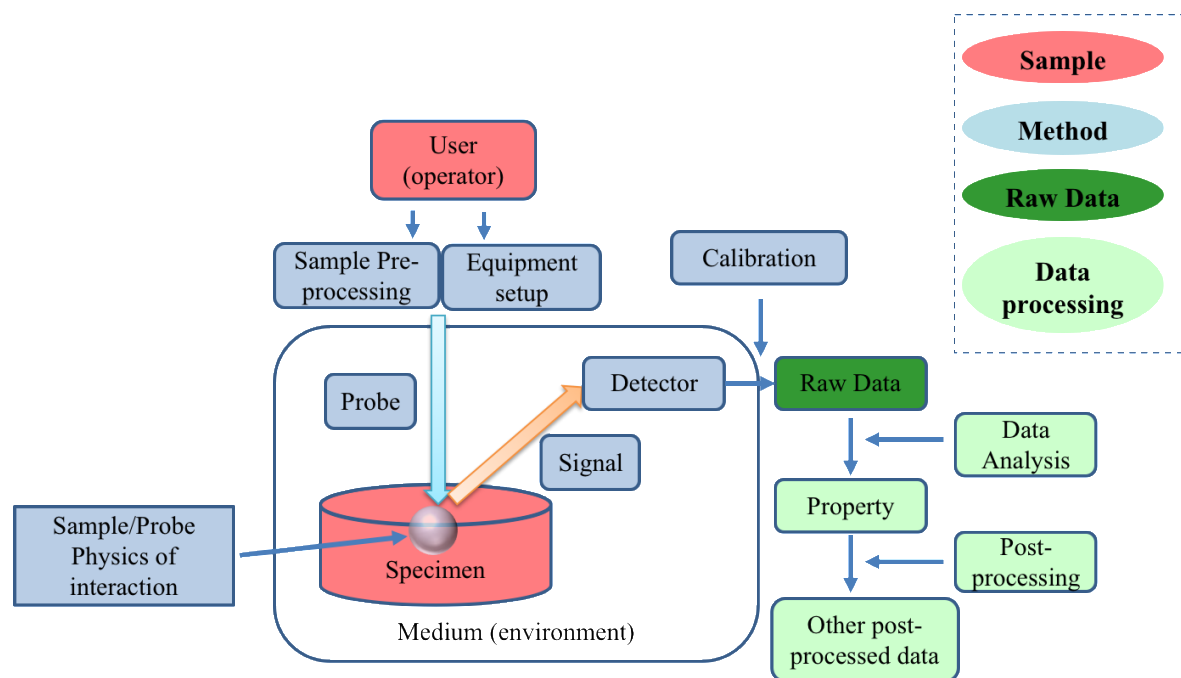
Figure 4. Visual representation of a characterisation experiment with keywords and colors

In this scheme, we also define the fundamental vocabulary that describes the main elements of a characterisation experiment, which could be further refined for any specific technique.

Since all standardized 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). 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.
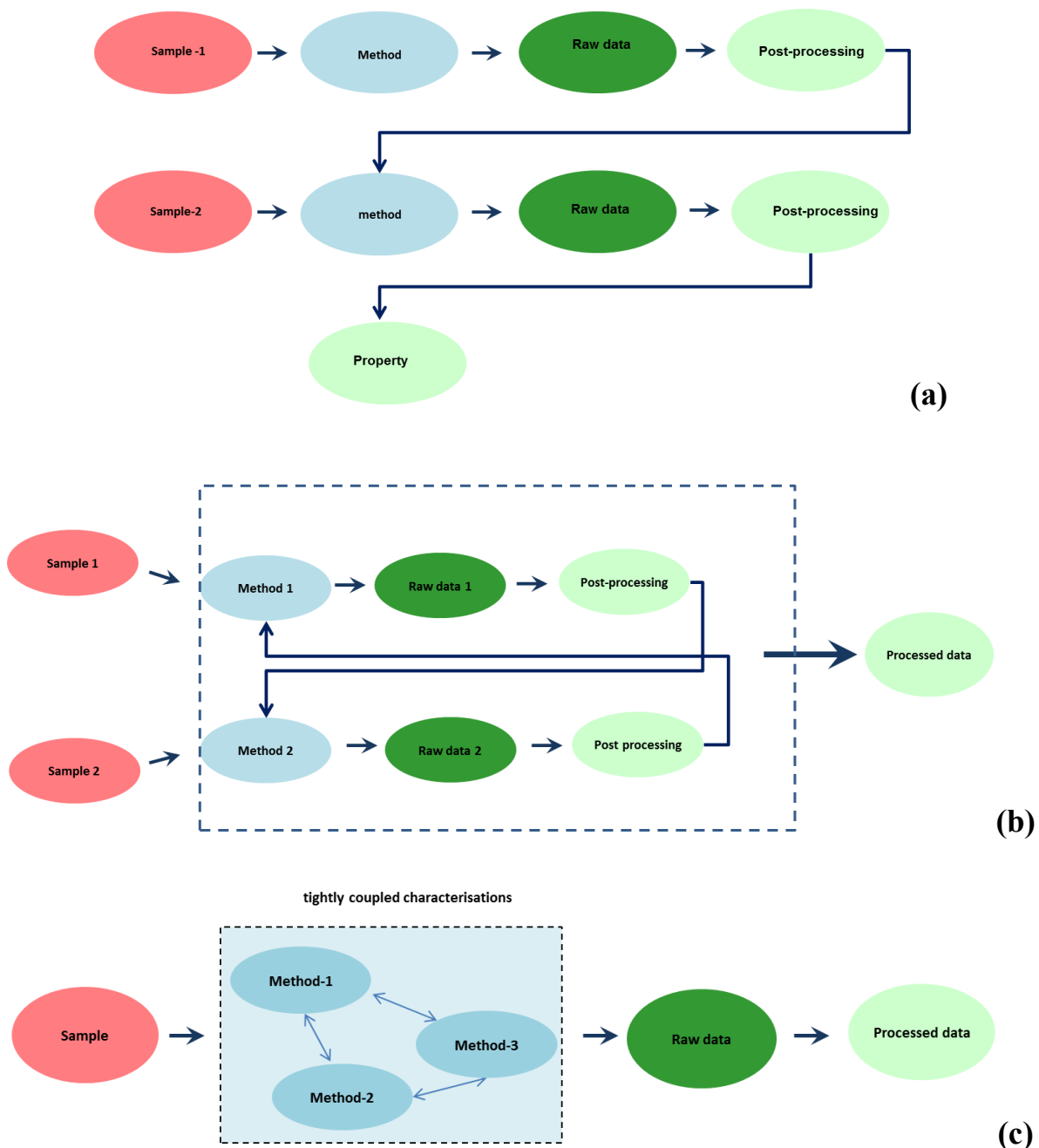
Figure 5. Visual representation of (a) sequence of multiple-samples/actions and multiple data processing steps, and (b) iterative workflow and (c) an example of coupled characterizations

4. 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 main features of this method, including hardware description and a typical load-displacement curve, are shown in Figure 6.

(a)   (b)
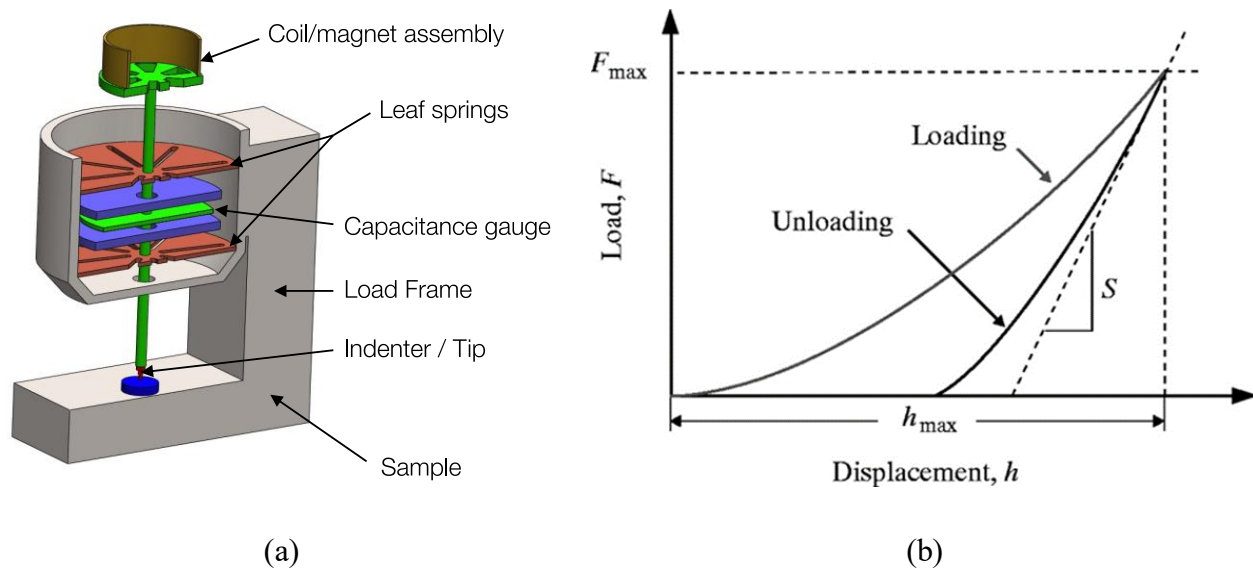
Figure 6. Basic instrumentation and output of nanoindentation technique.

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 analyze the load/unload curves to extract hardness and elastic modulus as main outputs [31].

Recently, high-speed nanoindentation [36-37] 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 7, 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.
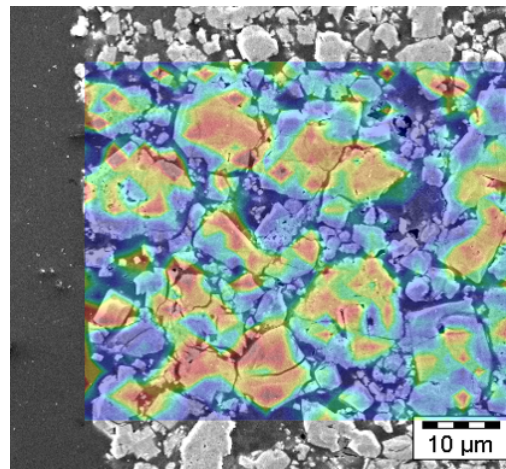
Figure 7. Example of the overlap between the microstructure of a battery composite (SEM image) and the corresponding high-speed nanoindentation map of the elastic modulus.

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 8). In fact, each of the block in the workflow represents a set of information that is stored in the metadata and will be required for reproduction and reuse of the information.

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 analyzed 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.

Then, a much more detailed metadata can be built up, if needed, by filling each of the CHADA elements 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.). An example of the resulting compiled metadata is reported in the table 1.

The benefits of using such a novel structure for the metadata associated to nanoindentation experiments can be summarized as follows:

1. First of all, the calibration procedure associated to the experiment is stored in the metadata, thus enabling the user to identify any calibration issue even in the future;

2. Then, all information about the test sequence, environmental conditions and test conditions are stored in the metadata;

3. Finally, all the data analysis steps leading to property evaluation are stored, thus enabling future users to re-evaluate the quality of the data analysis process.

4. The relevance and potential impact of the adopted approach becomes even more relevant for high-speed nanoindentation, where a very large amount of experimental data is created.
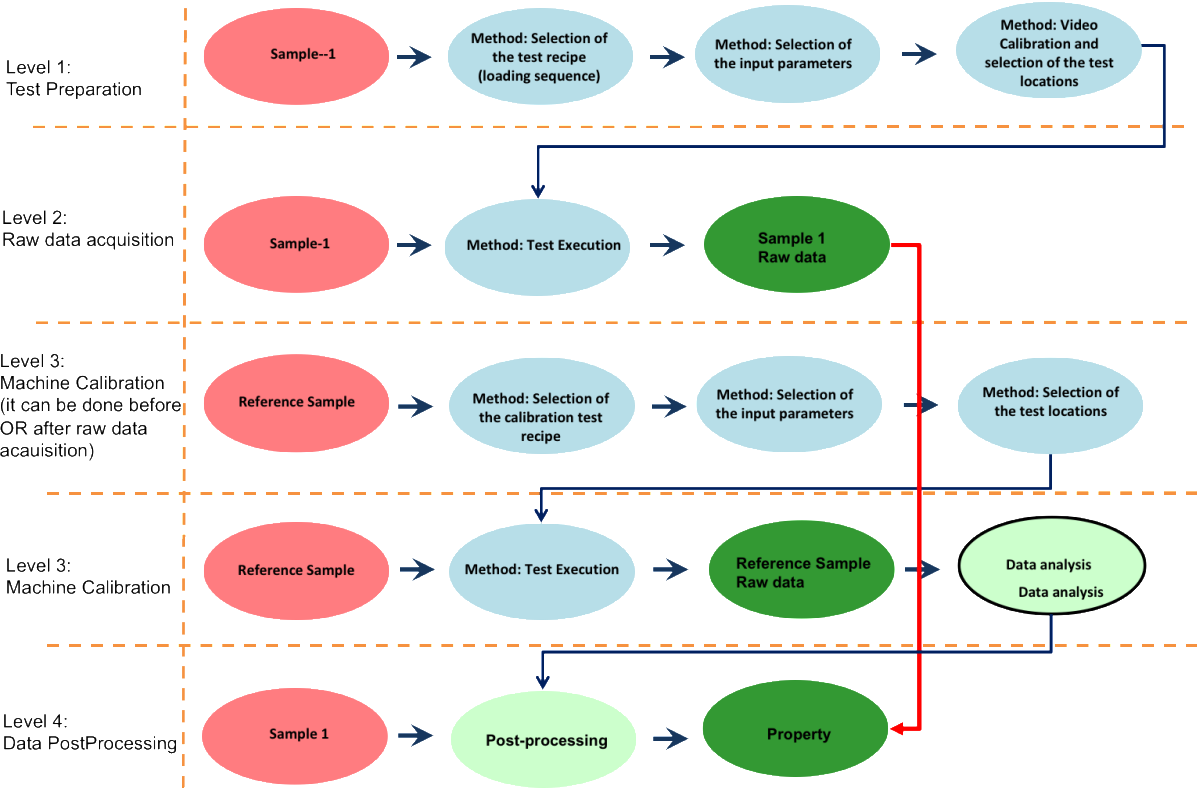
Figure 8. CHADA workflow for nanoindentation

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...).
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.

5. A novel management approach based on CHADA documentation

5.1 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].

5.2 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

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].

5.3 Data Management Plan

A Data Management Plan (DMP) includes 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].

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 9.

The main *information revealed in the data management plan can be categorized as follows:*

- *data summary, which includes the types and formats of files used for the oyster project per partner*
- *fair data, which refers to the future availability of the data created*
- *allocation of resources, which refers to the costs related to the process of making data FAIR (findable, accessible, interoperable, reusable)*

Moreover, the entire DMP is designed to be fully compliant with the new CHADA schemes that were presented in chapter 3. For that purpose, the same color scheme is used. More specifically, the DMP was separated and enhanced as follows: six main categories were created, namely General, Sample, Method, Raw data, Data analysis and Fair data. In the first category a general description of the data are gathered, such as purpose and quantity of data along with data security and storage

policies. This information corresponds to the introduction section and provides the general guidelines about the sections that will follow. The next section refers to the samples that are intended for experimental processes, excluding simulation data. The information required, initially concerns the description of the materials used and then, more specific input such as sample's dimension and quantity is requested. At the same time it is stated whether materials are used for calibration of equipment or not, indicating also that equipment. The next category is divided into two subcategories. The first one refers to the data origin (observational, experimental etc.) and the second one to the methodologies followed to produce the data, either these are experimental or simulation ones. In the following figure, the DMP section corresponding to "method" description in the CHADA are shown, as an example.

SAMPLE (not for simulation Data)						
This section refers to the samples used						
Materials used		Sample dimensions (1 inch diam. - 5mm thick). Surface flat and polished. Sample embedding on sample-holder (hot glue or acrylic glue). Optical sample surface alignment with reference sample (SiO2) surface.		State if the described Material is used for callibration of equipment (Yes/No)	Instrument, which is callibrated with the material	Standard CSM tests on reference sample
	Description of Material	User case (sample specifications)	Quantity	Callibration (yes/no) - Instrument-Standards used		

Figure 9: Part of DMP corresponding to "method" description (DMP template available in Zenodo https://doi.org/10.5281/zenodo.2635768).

The fourth section is dedicated to the description of the raw data which can be of type 'text', 'numeric' or 'audiovisual'. Moreover, information about simulated data are collected, such as the model and computer code used. It should be noted that special attention is given to whether the data are re-used or not. The fifth section is devoted to the data coming from the analysis of experimental or simulated

observations. At the same section it is also mentioned the method, description, purpose and software used for the data analysis. Finally, the information is provided on how data are made FAIR.

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, 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).
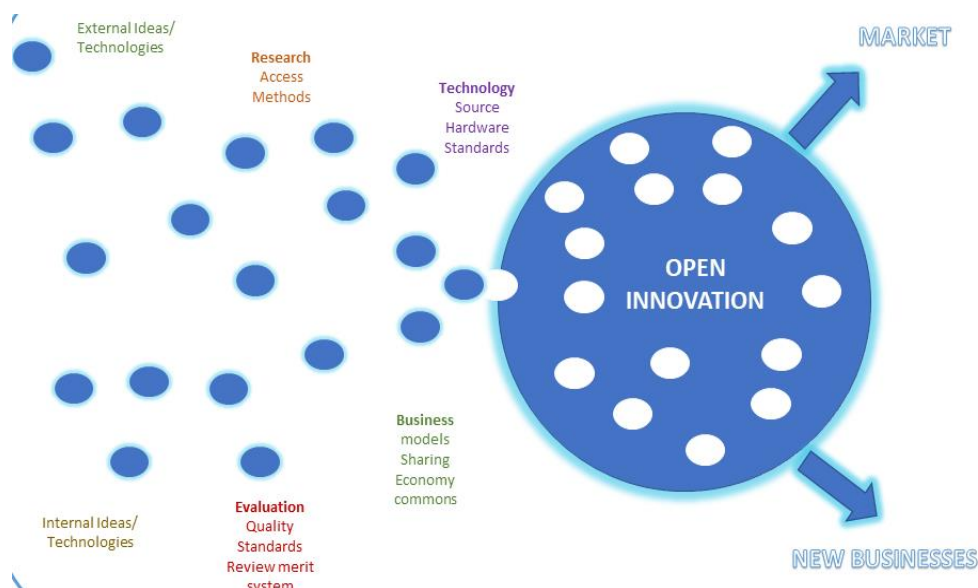
Figure 10: Schematic figure of Open Innovation Environment function

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 10) 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 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. 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 the issue of data structuring/handling/storing.

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.

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, multi-physics 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.

After reviewing the most recent approaches of data science in materials characterisation, including related taxonomies and ontologies, we presented a possible novel approach for the structure of a generic characterisation metadata, which we called CHADA. The new data documentation structure is then applied to the case of nanoindentation testing and the potential impact is discussed.

Finally, we discuss how an innovative approach for characterisation data structuring/handling/storing/sharing can have a relevant impact on the development of improved nano-materials and nano-enabled products, by introducing the concept of Open Innovation Environment.
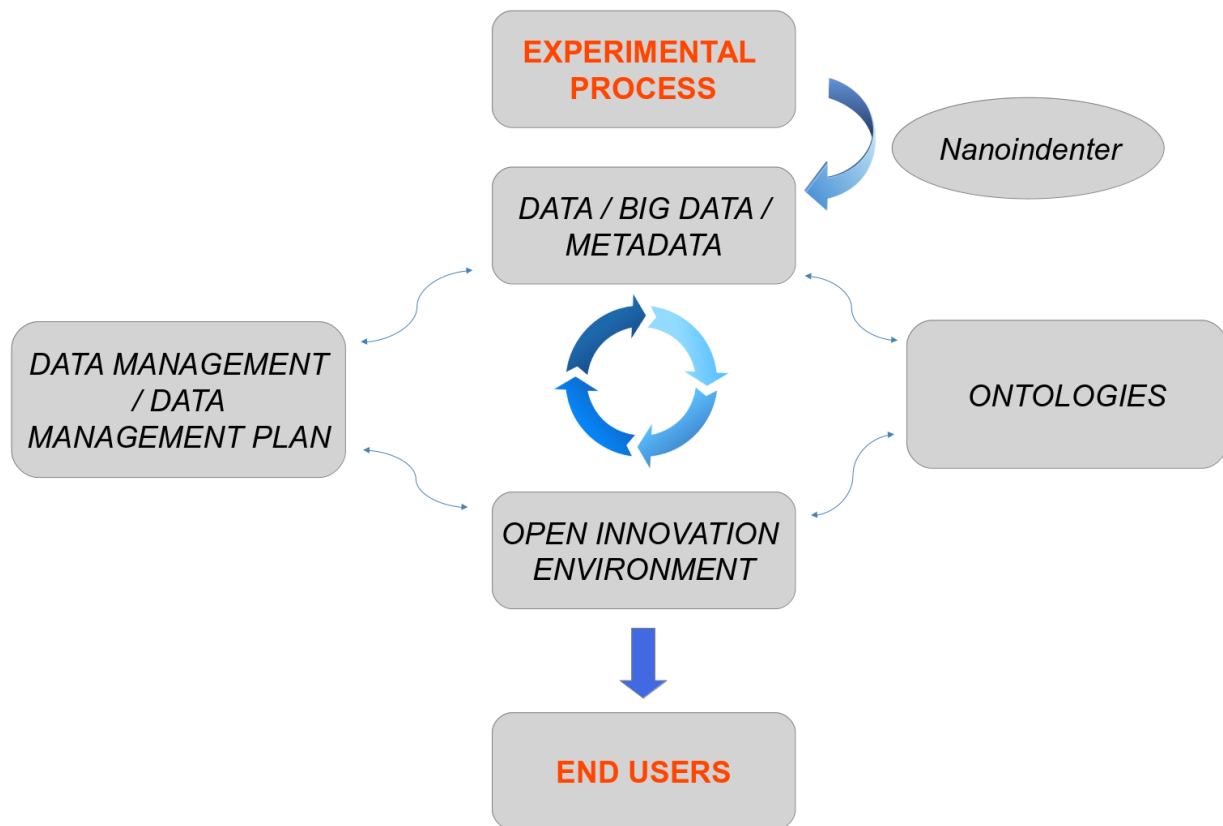
Figure 11: Interaction among Data, Ontologies, DMP and OIE – specific example for nanoindentation testing as main adopted characterization method

In order to make this action more effective, the application of such innovations to materials characterisation is deeply underpinned with a series of required international cooperative actions:

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

(2) The wider sharing of new concepts and structures for documentation of characterisation data (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.

Acknowledgements

All authors equally contributed to this work.

The paper received funding from the European Union, within the large collaborative projects OYSTER, Grant Agreement No. 760827, www.oyster-project.eu/ and for DMP partially from SMARTFAN, Grant Agreement No. 760779, http://www.smartfan-project.eu/. The entire oyster project team should also be acknowledged for the fruit fruitful 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 Research 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.

[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

[35] Weber Chr., Husung St., Cascini G., Cantamessa M., Marjanovic D., Rotini F., The 20th International Conference on Engineering Design (ICED15) – Design For Life, 2015, Milano, Italy

[36] G. Bolelli, M.G. Righi, M.Z. Mughal, R. Moscatelli, O. Ligabue, N. Antolotti, M. Sebastiani, L. Lusvarghi, E. Bemporad, Damage progression in thermal barrier coating systems during thermal cycling: A nano-mechanical assessment, Materials & Design. 166 (2019) 107615. doi:10.1016/j.matdes.2019.107615.

[37] M. Sebastiani, R. Moscatelli, F. Ridi, P. Baglioni, F. Carassiti, High-resolution high-speed nanoindentation mapping of cement pastes: Unravelling the effect of microstructure on the mechanical properties of hydrated phases, Materials & Design 97 (2016) 372-380.