Analyzing the Current Landscape of Materials Data Infrastructures

Easu Kwesiga Jonan
Kabale University, Kabale Municipality, Kabale Kigali Highway, Kabale, Uganda.
easukj@kab.ac.ug

Correspondence should be addressed to Easu Kwesiga Jonan : easukj@kab.ac.ug

Article Info
Journal of Computational Intelligence in Materials Science (https://anapub.co.ke/journals/jcims/jcims.html)
Doi: https://doi.org/10.53759/832X/JCIMS202402004.
Received 10 October 2023; Revised from 20 January 2024; Accepted 16 February 2024.
Available online 24 February 2024.

Abstract – The development and production of engineering materials receive substantial investments from both the industrial and research sectors. The materials are produced and certified in adherence to a set of product and evaluation criteria that have progressed over numerous decades to fulfill increasingly rigorous specifications. However, the considerable amount of data generated from these endeavors remains predominantly inaccessible. The endeavor to create a digital framework for engineering materials data can be dated back to over thirty years ago. However, the extensive implementation of machine-readable formats that facilitate the regular exchange of engineering materials data is yet to be accomplished. This article provides a comprehensive overview of the development of materials data infrastructures that are designed to gather, store, and disseminate materials data to various stakeholders. Prior to delving into the present state of digital materials infrastructures, we shall first contemplate the early iterations of such infrastructures. Several challenges that must be addressed before the development of a robust materials search engines and discovery tool are also reviewed.

Keywords – Materials Infrastructures, Materials Discovery, Materials Science, Materials Data, Aalto Materials Digitalization Platform, Novel Materials Discovery.

I. INTRODUCTION

The utilization of data as a driving force in scientific research is widely recognized as a novel approach in the domain of materials science. Within this domain, information is considered a valuable commodity, and insights are derived from voluminous or intricate materials datasets that surpass conventional human cognitive capabilities. The primary objective is often to identify novel or enhanced materials or material properties. At present, our focus lies in the establishment of two different material data infrastructures, namely the Aalto Materials Digitalization Platform and the Novel Materials Discovery (NOMAD) laboratory. The utilization of data-driven methods is transforming the approach to modeling, prediction, and control of intricate systems. The present textbook amalgamates the fields of machine learning (ML), engineering mathematics, and mathematical physics to unify the modeling and control of dynamical systems with contemporary techniques in data science. The text elucidates numerous contemporary developments in scientific computing that facilitate the implementation of data-driven techniques on intricate systems, including but not limited to turbulence, climate, the brain, epidemiology, robotics, finance, and autonomy.

The management of digital research data is becoming increasingly significant in all areas of application within the field of engineering sciences. This phenomenon is particularly evident as a result of the increasing volume of data derived from both experimental and simulated sources. The process of deriving insights from data is commonly known as the data-based 4^{th} science paradigm, which falls under the umbrella of data science. The complexity of research and comprehension of novel materials is increasingly evident in the materials science field. In the absence of appropriate analytical techniques, the continuously increasing volume of data will become unmanageable. The research data organized storage and its accompanying metadata is a crucial aspect for conducting effective data analyses. A requisite for effective research data management is the implementation of standardized protocols, facilitated by the availability of suitable data infrastructures such as repositories. These systems have the potential to address challenges in inter-institutional data exchange, facilitate comparison between experimental and theoretical data, and design duplicable workflow effective for the analysis of data, including enhancing similar data storage. In addition, the act of associating the dataset with obstinate identifiers allows for other scholars to make direct citations to them in their research.

Research data repositories are increasingly prevalent, serving as platforms for the storage and dissemination of data, both internally and publicly. As a result, data dissemination (as a freestanding entity or as a supplement to a textual

publication) has become more common and mandatory. Re3data and FairSharing are two available services that can be utilized to locate an appropriate repository. In addition, these services facilitate the discovery of repositories that are specialized in materials science data pertaining to specific subjects. The Materials Project and the NOMAD Repository are two widely recognized instances. Repositories that are indexed are typically hosted in a centralized or institutional manner and are primarily utilized for the dissemination of data. Notwithstanding, certain fundamental systems may be incorporated by users for internal usage within particular study groups. Moreover, this feature facilitates complete management of archived data and intra-data transmissions, in the event that said functionality has not yet been integrated into the repository. Open-source systems are of significant importance due to their ability to provide vendor independence and facilitate modifications to existing functionality or incorporation of supplementary features, often through integrated plug-in systems. Ckan, Dataverse, DSpace, and Invenio are among the systems that fall under this category. It is worth noting that Zenodo is built on the foundation of Invenio. The aforementioned repositories are of a general nature and solely constitute a subset of the extant open-source platforms.

Electronic lab notebooks (ELN) [1] are becoming increasingly popular in experimentally driven research areas as a supplementary system to repositories. Presently, Electronic Lab Notebooks (ELNs) have expanded their capabilities beyond the mere substitution of traditional paper-based laboratory notebooks. ELNs now encompass additional features such as data analysis, which is evident in tools like Galaxy or Jupyter Notebooks. Both of the aforementioned systems prioritize the provision of computational research that is both accessible and reproducible. The demarcation between unstructured and structured data is becoming increasingly indistinct, with the latter being conventionally confined to laboratory information management systems (LIMS) [2]. The majority of current Electronic Laboratory Notebooks (ELNs) are tailored to particular domains and are constrained to research fields such as biology or chemistry. Based on present understanding, there is currently no system that is specifically designed for the field of materials science. Open-source systems, including eLabFTW, SciNote, and Chemotion, are available for Electronic Laboratory Notebooks (ELNs). In contrast to the repositories, the assortment of Electronic Lab Notebooks (ELNs) is comparatively limited. In addition, it should be noted that only the initial two systems are considered to be generic.

It is noteworthy that universal data software and systems are accessible for both repositories and Electronic Laboratory Notebooks (ELNs). These systems and software have the potential to be utilized in the field of materials science as well. The open-source resolutions that have been enumerated hold significant importance, owing to their adaptability to diverse requirements and their potential applicability in a customized setup within individual research cohorts. Nonetheless, both facets can pose a significant obstacle, particularly for less sizable collectives. The absence of adequate resources and a systematic approach to research data management pose significant challenges for groups seeking to make their data available for future use. The absence of a system that can be implemented and utilized in both centralized and decentralized contexts, as well as internally and externally, without significant hindrances is the remaining deficiency.

It is imperative that the system provides comprehensive support to researchers throughout the entirety of their research process. This includes facilitating the extraction and generation of raw data, as well as enabling structural exchange, structured storage, and data analysis. Ultimately, the system should enable researchers to publish their results with ease. By integrating the functionalities of the repositories and Electronic Laboratory Notebook (ELN), a virtual study ecosystem is established, which expedites the production of novel ideas by fostering the cooperation among scholars. In the realm of materials sciences, characterized by its interdisciplinary nature, there exists a distinct necessity to construct models that accurately depict the highly diverse workflows of the researchers involved.

After contextualizing Open Science, our subsequent focus is on the development of materials infrastructures, which gather, store, and disseminate materials dataset to relevant parties. Prior to delving into the present state, we shall initially contemplate the early digital materials infrastructures. The subsequent sections of the article have been structured in the following manner: Section II presents an evaluation of materials discovery and infrastructures. In this section, different themes are discussed: autonomous materials discovery and manufacturing systems, development of materials science and infrastructure, and materials infrastructure. Section III presents a detailed evaluation of the present materials infrastructures. Section IV focuses on current challenges facing materials infrastructures such as relevance and adoption, completeness, standardization, acceptance and ecosystems, and longevity and diffusion. Lastly, Section V presents a conclusion as well as directions for future research.

II. MATERIALS DISCOVERY AND INFRASTRUCTURES

Autonomous Materials Discovery And Manufacturing Systems

The utilization of an Autonomous Materials Discovery and Manufacturing (AMDM) System facilitates a repetitive and flexible implementation of dynamic PSP learning associations, in addition to a consecutive exploration of an MDS via a cyclical interplay between the body (hardware) and the brain (software) constituent, with limited human engagement. The software is integrated with two different modules. The initial module, which is the forward mapping module, endeavors to acquire knowledge of the PSP correlation by integrating data from diverse sources. Its objective is to forecast the configuration and the pertinent characteristics of a substance at a designated design and process configuration, which is essentially a MDS point. The data sources would comprise of empirical measurements and observations of materials and production procedure, simulations of various physical procedures, which encompass the PSP associations across various resolutions, in addition to the experiential knowledge and domain proficiency. Multi-modal and multi-fidelity sources of

data are utilized to acquire knowledge on the relationships between the process, structure, and properties (PSP). The second module, which is the inverse mapping module, is designed to provide a prescription for experimental ecosystem and guide the MDS search. This is done by utilizing the PSP relationship model, with the ultimate goal of discovering materials as well as their corresponding recipe of their manufacturing. The software provides guidance on how to optimize predictions of the PSP relationship based on available data, as well as recommendations for experimental settings to obtain additional data and advance materials discovery objectives.

The physical structure comprises a dual-module system. The objective of a synthesis module is to independently carry out cerebral signals (either a process or input recipe) to effective generate different samples of materials and modify the workflows of processing according to previous results or observations, similar to reflex actions. Furthermore, a feature module is designed to investigate the manufactured materials, while also assessing and/or quantifying the resultant features (such as thermal conductivity or materials hardness) or/and performance (such as holistic metrics e.g., the efficiency of energy that relies on the full system of manufactured elements and processes, instead of entirely on the materials).

In the field of materials science, the term "characterization" pertains to the comprehensive and inclusive procedure of examining and quantifying a material's properties and structure. The process in question is deemed fundamental within the realm of materials science, as it serves as a prerequisite for establishing any scientific comprehension of engineering materials. The term "materials science" can mean anything from macroscopic techniques like mechanical testing and thermal analysis to sub-microscopic ones like density calculations and studies of materials' atomic structures. From angstroms, where individual atoms and chemical bonds can be seen, to centimeters, where coarse grain patterns in metals can be seen, the scale of the structures seen in materials characterization spans from these.

Numerous techniques have been utilized for characterizing materials over the centuries, including rudimentary optical microscopy. However, novel methods and approaches are continually being developed. The field of materials science has undergone a significant transformation in the 20th century, primarily due to the introduction of advanced technologies such as the electron microscopy and secondary ion mass spectrometry. These technologies have enabled researchers to visualize and analyze structures and compositions at a much smaller scale than was previously feasible. As a result, there has been a substantial improvement in comprehending the underlying reasons for the distinct properties and behaviors exhibited by various materials. In the past three decades, atomic force microscopy has augmented the uppermost attainable resolution for scrutinizing specific specimens. The implementation of a procedural protocol within the manufacturing module could range from a basic procedure of combining designated quantities of distinct species of chemical in test-tubes to a complex sequence of operations aimed at converting the precursors of input, such as consumables and raw materials, into a finished product. The synthesis module integrates different chambers of synthesis such as reaction beds, test-tubes, and machine tools, including storage systems and material delivery mechanisms for distributing the required raw materials to the chambers for synthesis. Moreover, a system for materials handling is incorporated to facilitate the transportation of materials between and within synthesis chambers.

The feature/characterization module employs distinct testing and measurement phases to evaluate the features and functions of the synthesized component or material. The goal of the testing procedures is to determine the quality of the object being tested, whether it be the strength of tensile force, magnetic permeability, or toughness of a functional alloy that has been synthesized, or the efficiency with which energy may be converted, or the rheology, of a chemical that has been grown in a petri dish or test tube. It is fundamental to note that distinct initiatives that have been introduced over the past few decades have predominantly centered on the materials' autonomous discovery. Various artificial intelligence techniques have been devised for exploring the MDS. Promising preliminary experimental platform realizations have been documented, encompassing digital as well as topological material configurations. The task of devising and potentially uncovering the procedural instructions for producing desired materials in large quantities and transforming them into a final product, commonly referred to as manufacturing process planning, presents notable supplementary complexities.

In the realm of manufacturing, it is essential to take into account the complete processing chain in order to develop effective recipes. Merely outlining the "static" parameter contexts could not be adequate. To accurately forecast the structure, it is imperative to gauge or approximate the development and transformation of the state variables of processes over time, space, and in the entire process chain stages. Furthermore, the discovery of processes frequently arises from innovative and unconventional process improvisations, rather than solely optimizing the process parameters. The process of finding and identifying manufacturing recipes can be likened to a game of chess, where each move is carefully considered based on observations of the board at every step. Similarly, in a multi-stage process chain, indicators and physical perceptions are taken into account to determine the appropriate recipe from a continuum of opportunities rather than finite combinations, in order to achieve a desired functionality.

The field of Advanced Manufacturing and Design for Materials (AMDM) has the potential to create a novel avenue for material-on-demand manufacturing, thanks to the recent advancements in data science/AI, materials genomics, and 3D printing. The management of materials microstructure and composition, rather than solely the morphology and form of a 3D-printed component, could be accomplished at a resolution, which ensures the functionality of components. The prospect of utilizing a novel generation of intelligent machine tools as reactor beds for self-governing, large-scale, and just-in-time production of innovative materials is currently under consideration. This ability presents intriguing prospects for producing components with predetermined functionality through inventive processing techniques, without depending on costly and limited resources (such as rare earth materials), thereby tackling the crucial materials concerns highlighted

by the science national academy. The progress made in AMDM can be analyzed by examining the advancements in data science and computational AI, which enables autonomy as well the accomplishment of materials discovery, including the improvements in hardware systems that enable the execution of adaptable workflows for autonomous experimentation in the areas of synthesis and discovery.

Development of Materials Science and Infrastructures

The selection of a particular material during a specific time period can serve as a significant characteristic. The terms Stone Age, Iron Age, Steel Age, and Bronze Age, are historical designations, albeit somewhat arbitrary. Materials science is a field of engineering and applied science that has its roots in the production of ceramics and its possible derivative, metallurgy. It is considered to be one of the traditional types of applied science and engineering. The field of contemporary materials science has its origins in metallurgy, which in turn can be traced back to the utilization of fire. During the late 19th century, Josiah Willard Gibbs, an American scientist, made a significant advancement in the comprehension of materials. He illustrated that the thermodynamic characteristics associated with atomic system in distinct phases are interconnected with the physical materials characteristics. The Space Race played a crucial role in the development of contemporary materials science, particularly in the comprehension and manipulation of metallic alloys, silica, and carbon materials. These materials were instrumental in the establishment of space motor vehicles, which facilitated space exploration. The materials science field has been instrumental in both propelling and being propelled by the advancement of groundbreaking initiatives such as semi-conductors, biomaterials, plastics, and rubbers.

Biomaterials are of paramount importance in the endeavor to achieve a sustainable society. The utilization of biomass-derived feedstock, such as wood, in biorefineries has the potential to offer renewable distribution of materials, including polymers, solvents, and chemicals that can be further incorporated into commodities that are of a high value. Biological materials also generate alternate routes for the management of wastes through biodegradation approaches and enhance equity within the economy by reducing our dependence on limited natural resources. Our research aims to expedite the advancement of biomaterial technologies through the utilization of machine learning-assisted methodologies for materials processing and modeling. The present endeavor centers on the utilization of Bayesian optimization for experiment planning and outcome prediction, facilitated by our proprietary software BOSS.

Prior to the 1960s, and in certain instances extending for several decades thereafter, a considerable number of contemporary materials science departments were originally established as either metallurgy or ceramics engineering departments, thereby reflecting the predominant focus on metals and ceramics during the 19th and early 20th centuries. The expansion of America's materials science was partly prompted by the agency for advanced research initiatives that provided funding for a sequence of laboratories hosted by universities during the 1960s, with the aim of broadening the local initiative of education and research in materials science. When comparing mechanical engineering to the emerging field of material science, the latter concentrates on the study of materials at a macroscopic level and emphasizes the importance of designing materials based on an understanding of their behavior at the microscopic level. The design of materials has evolved to be based on specific desired properties, owing to the increased understanding of the correlation between atomic and molecular processes and the overall characteristics of materials. The field of materials science has expanded to encompass a wide range of material classes, such as polymers, ceramics, semiconductors, biomaterials, magnetic materials, and nanomaterials. These materials are typically categorized into three distinct groups: ceramics, metals, and polymers. In recent decades, a notable development in materials science has been the increased utilization of computer simulations for the purpose of discovering novel materials, forecasting properties, and comprehending phenomena.

Materials Science

If one holds the belief that materials scientists are underappreciated and their contributions are often overlooked, the following statement may serve as supporting evidence. Given the temporal constraint of 50 years, the aforementioned inventory holds significant pertinence. The present discourse concerns the impact of materials science on contemporary society.

International technology roadmap for semiconductors

This is not novel research finding, but rather a methodology for arranging research objectives and strategizing efforts for research and development (R&D). The International Technology Roadmap for Semiconductors (ITRS) is a noteworthy accomplishment, as evidenced by its historical background, which outlines objectives pertaining to innovation, technology requirements, and metrics for advancement that are universally agreeable within the highly competitive realm of microelectronics. The ITRS, which encompasses science, technology, and economics, appears to be an effective driver of progress in various domains such as materials, characterization, fabrication, and device design. It can be considered a suitable initial option among the alternatives presented. The indispensability of electronics in contemporary society is coupled with the interdependent progress of semiconductor processing and materials science over the past half-century.

Scanning probe microscopes

Gerd Binnig and Heinrich Rohrer's establishment of the STM (scanning tunnelling microscope) was justifiably recognized with the the 1986 physics Nobel price. This microscopy approach is not only noteworthy in its own right, but it also enables the direct investigation of local sample properties with nanometer precision. The advent of the atomic force microscope (AFM) [3] was swiftly succeeded by a newfound ability to access the nanoscale realm, which is widely considered to have facilitated the present-day prevalence of nanotechnology. (For further clarification, please refer to the accompanying explanatory box: "Making sense of the nanoworld".) The innovation significantly enhanced our capabilities at this particular level.

Giant magnetoresistive effect

In 2007, the physics Nobel Prize was awarded to Peter Grünberg from Germany and Albert Fert from France for their personal giant magnetoresistance (GMR) discovery effect in late 1980s. Therefore, it is unsurprising to observe this progression on our roster. The Giant Magneto-Resistance (GMR) effect pertains to the significant modification in electrical resistance, which is perceived in a multilayered structure integrated both magnetic and non-magnetic materials upon the modification of an external magnetic field. The utilization of the phenomenon in the hard disk drive read heads has been significantly enhanced through the subsequent research conducted by Cordle et al. [4]. The aforementioned devices possess the capability to extract data that is magnetically stored on a hard disk by means of alterations in electric current. The remarkable GMR sensitivity read heads towards minute magnetic field enables significant reduction in the size of hard disk magnetic bits. The remarkable surge in our capacity to retain information that we presently observe can be attributed to this revelation.

LEDs and Semiconductor lasers

The emergence of light-emitting diodes (LEDs) and semiconductor lasers in 1962 is a remarkable narrative in the field of materials science. Semiconductor devices have become fundamental components in various technological applications such as barcode readers, telecommunications, laser printers, DVD and CD players, and many others. The emergence of solid-state lighting systems is anticipated to have a noteworthy impact on diminishing our energy consumption.

National nanotechnology

Former US President Bill Clinton is attributed with contributing to the fifth advancement in materials science, as listed. The individual in question is the former President of the United States who made public the inception of the National Nanotechnology Initiative (NNI) in the year 2000. The NNI is a comprehensive technology research and nanoscale science program that is federally funded and multi-agency in nature. The National Nanotechnology Initiative (NNI) has had a significant impact. The aforementioned event solidified the significance and potential of a developing discipline, promptly establishing it as the most captivating domain within the entirety of the natural sciences. The initiative facilitated the emergence of nanotechnology as a distinct field, with a clear vision and substantial financial support. Furthermore, it implemented a mechanism for financing cross-disciplinary scientific research in a manner that would compel other nations to strive for parity.

Mihail C. Roco, associated with the NSF (National Science Foundation), was among the participants who contributed to the preliminary launching of the NNI's vision and nationwide organizational endeavors. Roco recollects his involvement with a group comprising of Paul Alivisatos, Stan Williams, James Murday, Evelyn Hu, and Dick Siegel during the period of 1997 to 1999. The group started with a small number of members. Our vision entails a novel industrial revolution facilitated by the methodical management of matter on the nanoscale. The launching of the NNI in the year 2000 was made possible through the formation of a national coalition that included representatives from academia, industry, and various agencies. This coalition served as the foundation for the NNI, which was established with the aforementioned vision in mind. The National Nanotechnology Initiative (NNI) currently encompasses a total of 26 distinct agencies and is believed to possess a budget of approximately $1.5 billion as of 2008. Over the past seven years, it has emerged as the foremost contributor to nanotechnology research globally, having invested a staggering sum of $7 billion. Currently, nanotechnology has become a subject of national research focus for 65 countries, and the research and development activities in the industry sector have surpassed those of governments across the globe. In 2007, the worldwide research and development budget for nano-related endeavors exceeded $12 billion [5].

Roco introduced the NNI to Office of Science and Technology Policy (OSTP) in the White House on March 11, 1999, representing the interagency group. The concern among the populace was that the likelihood of nanotechnology being established as a program of national priority was minimal. It is possible that the subject matter may only appeal to a limited audience of scholars. The NNI was established in 2001 with a budget of $489 million, based on the definition of nanotechnology as a detailed scientific development platform, education, healthcare, and economic development. According to Roco, the National Nanotechnology Initiative (NNI) was developed with a level of thoroughness equivalent to that of a scientific endeavor.

Carbon fiber reinforced plastics

Over the past half-century, advanced composites have experienced a significant surge in popularity. These lightweight yet robust materials have found numerous applications in the aviation and aerospace industries. Contemporary composite materials have impacted a wide range of industries, encompassing transportation, packaging, civil engineering, and sports. These objects are commonly observed in Formula 1 vehicles, protective gear, and the blades of wind turbines. Carbon fiber reinforced plastics, also known as continuous carbon fiber organic-matrix composites, are at the forefront of this development. The present materials amalgamate rigid polymer matrix with carbon fibers, resulting in a composite material that exhibits remarkable toughness and low weight. During the initial years of the 1960s, carbon fibers were created using precursors such as polyacrylonitrile, rayon, and pitch. The exceptional strength and stiffness of the fibers can be attributed to the long, oriented aromatic molecular chains. The utilization of this particular material represents a significant improvement over the prior utilization of shapeless glass fibers in composite materials.

The emergence of composite materials with precise, tailored characteristics has been made possible by the progress in design, modeling, manufacturing, and the evolution of carbon fibers. According to Chen [6], the employment of organic-matrix composite, including their manufacturing approaches allows engineers to tailor the material to suit a particular application, as opposed to relying on a fixed set of material characteristics. The field of manufacturing science has expanded its horizons by enabling the shift of element designs to the design of materials. The remarkable enhancement in efficiency has resulted in the growing utilization of these substances, notwithstanding the expenses and amplified intricacy in designing, molding, and reprocessing. This is exemplified by the widespread implementation of composites in the wings and fuselage of the latest Boeing 787 aircraft.

Materials for Li ion batteries

Recalling the methods of coping prior to the advent of laptops and cellular phones can prove to be a challenging task. The aforementioned revolution owes its feasibility to the shift from the rechargeable battery using the aqueous electrolytes with H+ as the operative ion, to the significantly greater Li ion battery's energy density. The development of innovative electrode materials that meet various requirements was necessary for the advancement of Li-ion batteries. Specifically, the cathode necessitates a framework structure that is lightweight and possesses interstitial spaces to facilitate the reversible insertion and extraction of a significant quantity of Li ions with higher mobility. The development and exploration of new materials required a combination of astute chemical and electrochemical insight, logical evaluation of technical specifications, and significant empirical endeavor. The pioneering contributions of John B. Goodenough at the University of Oxford during the 1980s have been particularly influential in this field. In 1991, Sony developed a battery utilizing a cathode material known as $LiCoO_2$ in conjunction with a carbon anode [7]. This innovation enabled the creation of portable electronic devices that are ubiquitous in modern society. Ongoing efforts are being made to advance the development of cathode materials that are free of the hazardous element Co and possess three-dimensional framework structures, such as $LiFeO_4$, in order to create batteries that are both environmentally safe and possess high-energy density.

Materials Infrastructures

The advancements in first-principles techniques and computational science have stimulated the interest of materials researchers in exploring novel computer-aided techniques of materials design and discovery. These approaches are expected to be more efficient, rapid, and cost-effective than traditional methods. The Phase diagram approach and database calculations are considered as one of the initial endeavors to utilize materials information in a novel and more effective manner. This approach involved the consolidation of numerous phase diagram calculations into a centralized database, thereby expediting the process of designing and developing new alloys. During the 1990s, the increasing ability to collect, store, and analyze massive amounts of data spurred academics to study the prospects of data science in scientific investigation. The material researchers at the MIT (Massachusetts Institute of Technology) have industrialized tools to forecast the characteristics of materials based on datasets, in light of these groundbreaking concepts.

Simultaneously, a group of researchers from the Denmark Technical University exhibited the capabilities of evolutionary approaches in materials discovery possessing particular characteristics. Additionally, they employed high-throughput screening to identify potential materials with essential parameters, thereby reducing the number of necessary experimentations. The MIT scientists have projected the possibility of constructing a "virtual materials laboratory" utilizing computational tools, wherein novel materials can be formulated and assessed through computer-based calculations. The aforementioned concepts ultimately culminated in the initiation of a meticulously selected repository of information, presently denominated as the Materials Project. The proposed Open Access database aims to utilize computing framework with high throughput to effectively determine the features of all the prevailing inorganic materials. This would facilitate the discovery of suitable materials for future researchers through dynamic investigation and data mining.

In response to the growing popularity of big data and data science, the United States government introduced the Materials Genome Initiative (MGI) in 2011. The aforementioned endeavor placed significant emphasis on the practicality of data informatics in the context of materials exploration and development. The global launch of comparable initiatives advocating for the openness and accessibility of electronically stored information in science has established a trend and ushered in a novel paradigm of materials science, known as data-based material science. An increasing number of Open

Access materials data initiatives have been established globally with the aim of reducing the investment and time needed to effectively support modern 10 to 20 year cycle of "research-development-commercialization" of new materials.

The temporal axis is partitioned into 3 periods, which correspond to the development of the materials infrastructures. The majority of the initial initiatives of materials data were established as databases that provided search capabilities and data hosting, with the aim of promoting the sharing of data among materials scientists within a broader society. The Materials Genome Inception and Initiative marked a significant milestone in the realm of data-centric materials science, as databases underwent a transformation into data centers, which generated basic materials and services of data analysis, as depicted in **Fig 1**. The growing fascination with data mining and artificial intelligence has led to a heightened enthusiasm among materials scientists to incorporate these algorithms into their research endeavors. Consequently, the primary objective of the majority of centers shifted towards devising methodologies for facilitating researchers in exploring, extracting, and interrogating the databases. The aforementioned event represents a significant milestone in data-drive materials science evolution, as data infrastructures are now being utilized as platforms for materials discovery (refer to **Fig 1**). These platforms have explicitly stated their objective of expediting the identification of new materials.

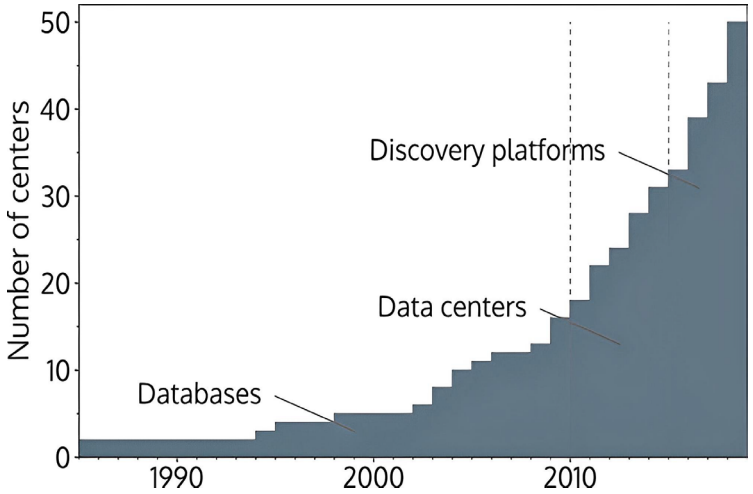

Fig 1. The materials quantity for informatics infrastructures and initiatives plotted as a time function

The classification of data centers, materials discovery systems, and databases, as discussed in the previous paragraphs, is centered on the somewhat vague distinctions provided. The terminology used in this document reflects our understanding of the evolution of materials data infrastructures. It provides a clear classification system for distinguishing between different types of infrastructure. The progress of materials science that is data-driven (materials informatics) is currently being driven by the transfer of knowledge and techniques from data science in the materials science field. The capabilities of machines to detect and analyse different data patterns has sparked novel feedback approaches in the relations between experimentation and the hypothesis. The aforementioned mechanism stimulates the consolidation of computational research, and the human trials-and-errors experiments, thereby leveraging "artificial intuition" to propel scientific inquiry. The tools of data mining can be utilized as an alternative approach to mimic human intuition and suggest potential materials. These materials can subsequently undergo additional refinement through computational and experimental investigations.

The employment of data science and big data is predominant across different scientific domains. The establishment of chemical databases preceded that of materials science databases. In 1965, Chemical Abstracts Service (CAS) established its inaugural database, thereby becoming the principal purveyor of chemical databases. The scrupulous development and management of datasets have played a pivotal role in the progressions achieved in the domain of quantum chemistry. The CERN portal of open data provides accessibility to more than a single petabyte of data that relate to particle physics study done at their facility. In the biology field, various databases are employed for the purpose of storing biological data. ConsensusPathDB is a database utilized for the storage of data related to various types of interactions such as gene regulatory, signalling, signalling, metabolic, genetic, protein-protein, and drug-target interactions in humans. The Data Bank of protein is a database that contains three-dimensional structural dataset of vital biological molecules. Furthermore, it is the responsibility of the International Nucleotide Sequence Database Collaboration to gather and distribute RNA and DNA sequences.

III. OVERVIEW OF CURRENT MATERIALS INFRASTRUCTURES

The graphical representation in **Fig 1** illustrates a noticeable increase in the presence of infrastructures for active materials. A significant number of these infrastructures have progressed into highly developed and reliable services that are frequently utilized in various research procedures. As materials discovery platforms have advanced, their services have expanded, as evidenced by the various features listed in **Table 1**. For an alternative view on the constituents of materials data infrastructures, please refer to [8].

Table 1. Material data infrastructure services	
Data upload	Permits users to effectively upload their data and produce DOIs (Digital Object Identifiers).
Open Access	Access to data is made partially or completely free.
Computational data	Contains data that came from computer simulations. The term "experimental data" refers to information derived through experimentations.
Workflow management tools	Offer solutions of the open-source software for the management of workflow or participate in their development.
Web API	Information may be retrieved remotely using scripting automation. Tools for analyzing data: give resources for doing data analysis, either online or off.

The efficient distribution of stored data is arguably the most crucial service provided by a data platform. Frequently, the data can be obtained via a website that is available for remote access by its users. The adoption barrier for this is minimal as it does not require any supplementary software and permits visual exploration of data via a web browser. Several services that fall under this category are the AFLOWlib, NOMAD Encyclopedia, the Materials Project, and the Materials Cloud.

The utility of a browser-based approach is limited in the context of materials informatics applications, as these necessitate the automated retrieval of substantial amounts of data. In order to enhance accessibility to vast amounts of data, it is customary to provide an API (application programming interface) to users, which allows for automated data retrieval. Frequently, the process involves the establishment of a GraphQL interface and Representational State Transfer (REST) for the data. The aforementioned interfaces facilitate programmatic queries for automated access. An alternative approach, as implemented by OQMD, involves providing users with the option to download the database directly as an offline version. The provision of offline access offers superior adaptability and efficiency, albeit necessitating familiarity with the interaction of the underlying database through either object-relational-mapping (ORM) [9] or Structured Query Language (SQL) [10]. However, it is impractical to perform a complete download for extensive amounts of data.

The escalating volumes of data generated by materials science have given rise to a pragmatic apprehension regarding the enduring preservation of this data. Moreover, funding agencies and other institutions are exerting mounting pressure to guarantee the accurate and secure preservation of data over an extended period. In response to this requirement, certain data infrastructures currently offer solutions of data storage for materials data. Presently, the Springer-Nature has established two repositories of data that are highly recommended for the field of materials science, namely the Materials Cloud and the NOMAD Repository. Both of the aforementioned complimentary services are intended for the purpose of mathematical materials dataset. They possess the capability to receive uploads from any given source and ensure the retention of data for a minimum of a decade following its deposition.

Frequently, the quantity of data in experimental investigations, particularly in imaging, surpasses the computational resources utilized. As an illustration, electron microscopes have the capability to produce a substantial amount of data, reaching up to tens of gigabytes, within a single day of operation. Due to the increased quantity of data, the task of arranging a centralized and cost-free repository for experimental data has become considerably more difficult. The provision of storage spaces is enabled by the laboratory or the host university, as demonstrated by the facility of materials data, a cooperative venture among US-based academic institutions and research facilities. Apart from the storage solution that is specific to materials science, there exist various costless alternatives to store scientific data of a general nature, including but not limited to Dryad, Zenodo, Dataverse and Figshare.

The advent of collaborative notebook-centric and web content ecosystem, such as Jupyter notebook, has led to the emergence of modern online analytics tools provided by data infrastructures. The available online resources encompass a spectrum of educational materials, including rudimentary instructional guides and advanced computational models that enable the forecasting of materials properties and the identification of new materials via artificial intelligence. Online resources have gained significance as a means of distribution and education due to their ability to function independently of local hardware and software resources. Certain platforms also engage in the advancement of Open Source software libraries designed for conducting offline analysis of materials data. Libraries of this nature possess significant potential for reuse by researchers who are engaged in the study of materials data. Furthermore, by means of Open Source distribution and contribution mechanisms, these libraries can continue to be utilized actively even after the completion of individual projects.

Materials data has been acknowledged for its significance by companies specializing in materials informatics. The availability of a significant amount of materials data can greatly benefit companies in the selection of the most suitable materials for their products, thereby eliminating the need for them to incur additional costs in establishing their individual research data infrastructures. A novel model that has surfaced in recent times pertains to vending software environments through a SaaS (Software as a Service) framework. The present model entails the provision of on-demand access to pre-integrated cloud-based system for material informatics by corporations. The utilization of such services can prove to be advantageous for corporations and research facilities due to their ability to cater to the present demand, absence of significant capital investments in hardware, and the lack of necessity for expertise in system management and software configuration.

IV. CURRENT CHALLENGES

After conducting an analysis of the present condition of materials infrastructures in materials science field, we will nowredirect our attention to MUSE analogy. Despite significant progress in data-based materials science, there are still numerous challenges that must be addressed before a vigorous materials discovery tool and search engine can be developed.

Relevance and Adoption

In order to gain acceptance from various stakeholders, including scientific communities, governments, industries, and the public, materials data infrastructures must furnish pertinent data and information. The determination of relevance is contingent upon several factors, including the type, quality, and volume of data. Additionally, it encompasses the completeness and homogeneity of the data. The varying specifications of these terms across different communities pose a challenge in the development of universal and interdisciplinary data infrastructures, which could be extensively adopted. The interdependence between adoption and relevance is closely connected to the ensuing obstacles of comprehensiveness, uniformity, and approval. The concept of relevance encompasses various tools that facilitate the classification, analysis, and correlation of data. Machine learning has emerged as a prominent approach in this context. Given the data-intensive nature of machine learning, it is logical to incorporate machine learning applications into materials data infrastructures. The acceptance of materials infrastructures can be hindered by various challenges associated with one-stop-shop solutions. These challenges may arise due to the diverse range of ML algorithms and diversity of data that are currently available. In order for datasets to be effectively utilized by ML algorithms, it is imperative that the features and properties of the data are both relevant and comprehensive.

Contemporary society is predominantly reliant on technology for its means of communication, economic activities, and progressively for its daily operations. The swift advancement of information technology has provided a multitude of novel resources for post-secondary education. The advent of novel technology has facilitated the provision of educational opportunities to individuals irrespective of their location and time constraints. The higher education institutions have exhibited an unusually swift reaction to this emerging technology. Insufficient allocation of resources towards technology-based education in tertiary institutions could potentially impede their capacity to remain competitive in emerging or evolving markets. The utilization of technologies such as the internet and its associated technologies can enhance an trainer's capacity to enable learners to make essential connections to contexts, content and the society in a more efficient and scalable manner, thereby leading to a more impactful learning experience.

The NCATA (National Council for Accreditation of Teacher Education), the sanctioning body responsible for evaluating teacher groundwork programs in America, is taking proactive measures to ensure that new teachers possess the necessary technological competencies to effectively integrate technology into their teaching practices. Specifically, NCATE is enhancing its standards for the year 2000 to emphasize technology and to be more performance-based. The demand for electronic tools that offer adaptability and contextualized learning is on the rise. Higher education institutions ought to confront the challenge posed by emerging technologies. In the United Kingdom, the Joint Information Systems Committee (JISC) is engaged in the investigation, experimentation, and comprehension of diverse technological tools and their potential applications in higher education.

Completeness

Completeness [11] refers to the state or quality of being entirely whole or flawless, without any omissions or deficiencies. Attaining ideal completeness in practice may prove to be challenging. However, contemporary data infrastructures are confronted with a tangible and severe completeness predicament, as they predominantly comprise computational data with a dearth of experimental data. Due to the digital nature of computational data, database platforms were swiftly embraced by computational scientists. Additionally, it can be observed that computational data exhibits a higher degree of homogeneity and is comparatively simpler to curate in contrast to experimental data. Enabling a smooth comparison between experimental and computational data is a crucial measure towards verifying theoretical forecasts and advancing materials development and materials discovery endeavors.

Materials, which exhibit potential still, necessitate additional assessment, screening, and experimentation. The task of establishing synergies between experimental and computational database remains a vital concern for data-driven materials science development. There exists a worldwide rivalry for the acquisition of skilled individuals among elite students, scholars, and educators. It is imperative for academic institutions to engage in global competition by excelling in both teaching and research. Higher education institutions must uphold elevated levels of research to attain global recognition and remain competitive with other institutions through superior quality and standards of research. In order to enhance their research capacity, higher education institutions can establish high-quality partnerships with other higher education systems around the globe. In addition, Higher Education Institutions are discovering that engaging in international and local partnerships with other Higher Education Institutions, industries, communities, and government entities is imperative in order to capitalize on the advantages presented by globalization.

Standardization

The implementation of a new framework or technology on a large scale necessitates the establishment of some degree of standardization. Effective participation of stakeholders in technology development necessitates a shared language. In the data-based materials science realm, metadata serves as the linguistic equivalent. Metadata establishes connections, or the syntax, among individual data elements, or lexical units. The task of creating a uniform set of metadata for materials science that is comprehensive, informative, and flexible presents a significant hurdle. The establishment of a materials ontology, which refers to a systematic categorization of materials, would serve as the fundamental basis for materials science dataset and the amalgamation of diversified materials science terms into a common language. A plethora of materials can be observed in our surroundings, ranging from architectural structures to spacecrafts. Materials are broadly categorized into three groups based on their chemistry and atomic structure (**Table 2**).

Table 2. Grouping of materials

Metals	Metal is a substance that typically exists in a solid state and is composed of a single or multiple metalling compounds, such as chromium, aluminium, gold, copper, iron, nickel or titanium. Additionally, it may contain nonmetallic elements, such as nitrogen, oxygen, or carbon, in small quantities.
Ceramics	Ceramics are classified as solid materials that consist of inorganic element of metalloid, non-metal, and metal atoms, which are majorly bound together through covalent and ionic bonds. Brick, earthenware, and porcelain are frequently cited as typical instances. Uranium dioxide is a refractory ceramic compound commonly utilized as a nuclear fuel within the nuclear industry.
Polymers	Polymers are macromolecules that consist of nonmetallic elements, hydrogen, and carbon. Polymers encompass a broad spectrum of materials, including both synthetic plastics, such as natural biopolymers, and polystyrene, such as DNA and proteins, which play a crucial role in biological structure and function. Polyethylene (PE), polystyrene (PS), silicone rubber, nylon, and polycarbonate (PC) are among the commonly known and recognizable polymers.

Acceptance and Ecosystems

The efficacy of materials data infrastructures is contingent upon their acceptance as a valuable instrument by diverse stakeholders. In addition to possessing relevance and comprehensiveness, data infrastructures must exhibit user-friendliness to achieve broad adoption. The aspect of user friendliness encompasses the seamless uploading and downloading of data. The ease of uploading data also enhances completeness by minimizing obstacles to data sharing. The attainment of extensive acknowledgement necessitates the establishment of reliance in the archived information, which can solely be accomplished by data curation. Data curation alludes to systematic control and data oversight in its cycle, integrating the initiate storage and creation, as well as its eventual archiving for future reference or deletion when it is no longer relevant. The acceptance of infrastructure varies among stakeholders. As expounded in the preceding section, the present infrastructures are primarily constructed and utilized by scholars in the academic realm.

The systematic study of industry interest and participation has not been conducted, with anecdotal evidence serving as the primary basis for understanding this phenomenon. Certain materials companies utilize reference databases such as IBM and ASM International to enhance their value proposition, whereas others opt to engage intermediaries for such services. In addition, it appears that the industry is currently in the process of investigating the possibilities and potential advantages of materials informatics, yet without complete involvement with the academic community. The lack of alignment between research and development efforts in academia and industry poses a challenge for the latter's engagement in this particular context. The materials gap presents a challenge to achieving broad industry adoption, as the extant materials data platforms do not provide the requisite data that the industry demands. Ecosystems that enable the facilitation of interactions among academic, governmental, corporate, and public stakeholders have the potential to serve as a solution.

Longevity and Diffusion

The growing recognition of open and data-driven scientific practices has led to a rise in both local and global funding for Open Science advancement. Consequently, new platforms for materials data are emerging as a result of this trend [12]. The consideration of longevity and diffusion of innovations and new technologies is infrequently taken into account by funding agencies, and there is no assurance of fiscal support for sustainable long-term operation. The primary architects of the first phase of digitalized materials data infrastructures were majorly material data scientists with a primary research emphasis on fundamental science. The preservation and functionality of infrastructure over an extended period of time is frequently not given primary consideration by the majority of scientists. Consequently, the digital infrastructures are at risk of transforming into digital ruins due to the proliferation of Open Science.

V. CONCLUSION AND FUTURE RESEARCH

After contextualizing Open Science, we proceed to examine the development of infrastructures, which gather, store, and furnish materials data to relevant parties. Prior to discussing the present state, we will engage in a reflection of the infrastructures of digital materials during their early stages. In addition, we examine various obstacles that must be addressed prior to the development of a robust materials discovery tools and search engines. The utilization of data-driven

approaches is widely recognized as a novel paradigm in the materials science field. In this domain, information has emerged as a valuable asset, and insights are derived from materials datasets that exceed the capacity of conventional human cognition. The primary objective is often to identify novel or enhanced materials or materials phenomena. Our current endeavors include the establishment of the Aalto Materials Digitalization Platform and the NOMAD laboratory.

The digitization of material research is an emerging trend in the field. Nonetheless, the implementation of this process will necessitate substantial exertions and modifications. However, initial patterns have already become discernible and advancements have been made in numerous domains. The varying classifications of materials are presently situated at distinct stages, which can be attributed to the fundamentally diverse approaches employed in their synthesis, production, analysis, and simulation. Typically, advancement to the next tier necessitates significant advancement beyond the individual sub-domains. The integration of diverse fields, including but not limited to synthesis, characterization, data management, machine learning, and digital fabrication, is both feasible and necessary. The establishment of a comprehensive knowledge chain in engineering would facilitate the realization of a completely automated, digital, and robot-assisted approach to material research, thereby revolutionizing the methods by which novel materials are discovered, developed, and processed. At present, the existing literature solely outlines individual and disconnected methodologies.

Consequently, expedited and enhanced reproducibility of material development will be achievable. In addition, the frequency of duplications in experimental procedures will be notably decreased, thereby minimizing the need for human resources to engage in extensive synthesis efforts. This will allow for greater emphasis on the development of targeted experimental designs aimed at investigating the properties of novel materials. In addition, conducting a greater quantity of experiments and establishing a centralized repository for the necessary data will facilitate a greater number of innovations, potentially yielding materials possessing properties that were previously inconceivable. The attainment of digital, automated, and robot-aided materials research poses numerous challenges. Primarily, it is imperative to advance the progression of current technologies, such as robots and machine-learning. The AI programs must be adapted to better suit the field of materials science and the diverse demands of various material classifications.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

There are no competing interests.

References

[1]. M. Rubacha, A. K. Rattan, and S. C. Hosselet, “A review of electronic laboratory notebooks available in the market today,” J. Lab. Autom., vol. 16, no. 1, pp. 90–98, 2011.

[2]. P. J. Prasad and G. L. Bodhe, “Trends in laboratory information management system,” Chemometr. Intell. Lab. Syst., vol. 118, pp. 187–192, 2012.

[3]. A. Jafari and A. Sadeghi, “A new insight into the mechanical properties of nanobiofibers and vibrational behavior of atomic force microscope beam considering them as the samples,” J. Mech. Behav. Biomed. Mater., vol. 142, p. 105842, 2023.

[4]. M. Cordle et al., “Impact of radius and skew angle on areal density in heat assisted magnetic recording hard disk drives,” AIP Adv., vol. 8, no. 5, p. 056507, 2018.

[5]. C. Pazzanese, “Clinton reflects on foreign policy triumphs and challenges,” Harvard Gazette, 08-Apr-2021. [Online]. Available: https://news.harvard.edu/gazette/story/2021/04/clinton-reflects-on-foreign-policy-triumphs-and-challenges/. [Accessed: 30-Apr-2023].

[6]. T. Chen, “A tailored non-linear fluctuation smoothing rule for semiconductor manufacturing factory scheduling,” Proc Inst Mech Eng Part I J Syst Control Eng, vol. 223, no. 2, pp. 149–160, 2009.

[7]. M. Drapa, “University of Chicago alum John B. Goodenough shares Nobel Prize for invention of lithium-ion battery,” University of Chicago, 09-Oct-2019. [Online]. Available: https://news.uchicago.edu/story/john-b-goodenough-shares-nobel-prize-invention-lithium-ion-battery. [Accessed: 30-Apr-2023].

[8]. A. Antón, M. Torrellas, V. Raya, and J. I. Montero, “Modelling the amount of materials to improve inventory datasets of greenhouse infrastructures,” Int. J. Life Cycle Assess., vol. 19, no. 1, pp. 29–41, 2014.

[9]. P. Michail and K. Christos, “Object relational mapping vs. Event-sourcing: Systematic review,” in Electronic Government and the Information Systems Perspective, Cham: Springer International Publishing, 2022, pp. 18–31.

[10]. P. Tuli and J. P. Patra, “Symbol question conversion in structured query language using fuzzy with deep attention based rain LSTM,” Multimed. Tools Appl., vol. 81, no. 22, pp. 32323–32349, 2022.

[11]. J. L. Christiansen, B. D. Clarke, C. J. Burke, J. M. Jenkins, and the Kepler Completeness Working Group, “The Kepler completeness study: A pipeline throughput experiment,” Proc. Int. Astron. Union, vol. 8, no. S293, pp. 88–93, 2012.

[12]. J. Wang, M. Chao, G. Lin, C. Gao, and D. Liu, “Research on the construction of information service platforms for electricity market large data: Research on the construction of information service platforms for electricity market large data,” in Advances in Energy, Environment and Materials Science, CRC Press, 2016, pp. 9–13.