



Building a **Materials Data** Infrastructure

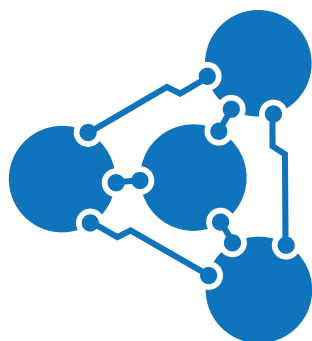
Opening New Pathways to Discovery and
Innovation in Science and Engineering



TMS

A Study Organized by The Minerals, Metals & Materials Society

Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering



Building a **Materials Data** Infrastructure

**Opening New Pathways to Discovery and
Innovation in Science and Engineering**

A Study Organized by The Minerals, Metals & Materials Society
Pittsburgh, PA 15237

On behalf of the National Science Foundation (NSF)

www.tms.org

Notice: *This report was prepared as an account of work sponsored by the National Science Foundation (NSF) of the United States government. Neither NSF, the United States government, nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring NSF, the United States government, or any agency thereof. The views and opinions expressed herein do not necessarily state or reflect those of NSF, the United States government, or any agency thereof. Similarly, although the information enclosed in this report represents the collective compilation of thoughts and input from the volunteer technical experts on the Materials Data Infrastructure teams, this report in no way represents the specific views of any of the individuals who contributed to this report, or any of their employers and affiliated organizations. These individuals, their affiliated organizations, and the publisher make no warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of the information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by these individuals or their employers.*

Cover: Representation of the three core digital components of a materials data infrastructure—repositories, tools, and e-collaboration platforms—along with related elements. Finite element analysis image is from "Solution to the heat equation in a pump casing model using the finite element modelling software Elmer" by A1 and is licensed under CC BY-SA 3.0. Normal direction inverse pole figure image is from A. Salem, et al. "Workflow for integrating mesoscale heterogeneities in materials structure with process simulation of titanium alloys," published in *Integrating Materials and Manufacturing Innovation* in December 2014 and is licensed under CC BY 4.0. Cover design by Bob Demmler, TMS.

Electronic copies of this report are available online at www.tms.org/mdistudy.

Copyright © 2017 by The Minerals, Metals & Materials Society, Pittsburgh, PA 15237.

All rights reserved.

DOI: 10.7449/mdistudy_1

ISBN: 978-0-692-86044-1

The Minerals, Metals & Materials Society (TMS)

*Promoting the global science and engineering professions
concerned with minerals, metals, and materials*

The Minerals, Metals & Materials Society (TMS) is a member-driven, international organization dedicated to the science and engineering professions concerned with minerals, metals and materials. TMS includes more than 13,000 professional and student members from more than 70 countries representing industry, government and academia.

The society's technical focus spans a broad range—from minerals processing and primary metals production to basic research and the advanced applications of materials.

In recent years, TMS has particularly established itself as a leader in advancing integrated computational materials engineering, computational materials science and engineering, and multiscale materials modeling and simulation.

To facilitate global knowledge exchange and networking, TMS organizes meetings; develops continuing education courses; publishes conference proceedings, peer-reviewed journals, textbooks, and technology studies and reports; and presents a variety of web resources accessed through www.tms.org.

TMS also represents materials science and engineering professions in the accreditation of educational programs and in the registration of professional engineers across the United States.

A recognized leader in bridging the gap between materials research and application, TMS leads and enables advancements in a broad spectrum of domestic and global initiatives.

www.tms.org

TMS

The Minerals, Metals & Materials Society

Contents

Acknowledgments xi

Preface xvii

Executive Summary xxi

Introduction and Motivation.....1

Background and Prior Efforts..... 7

Challenges19

Recommendations33

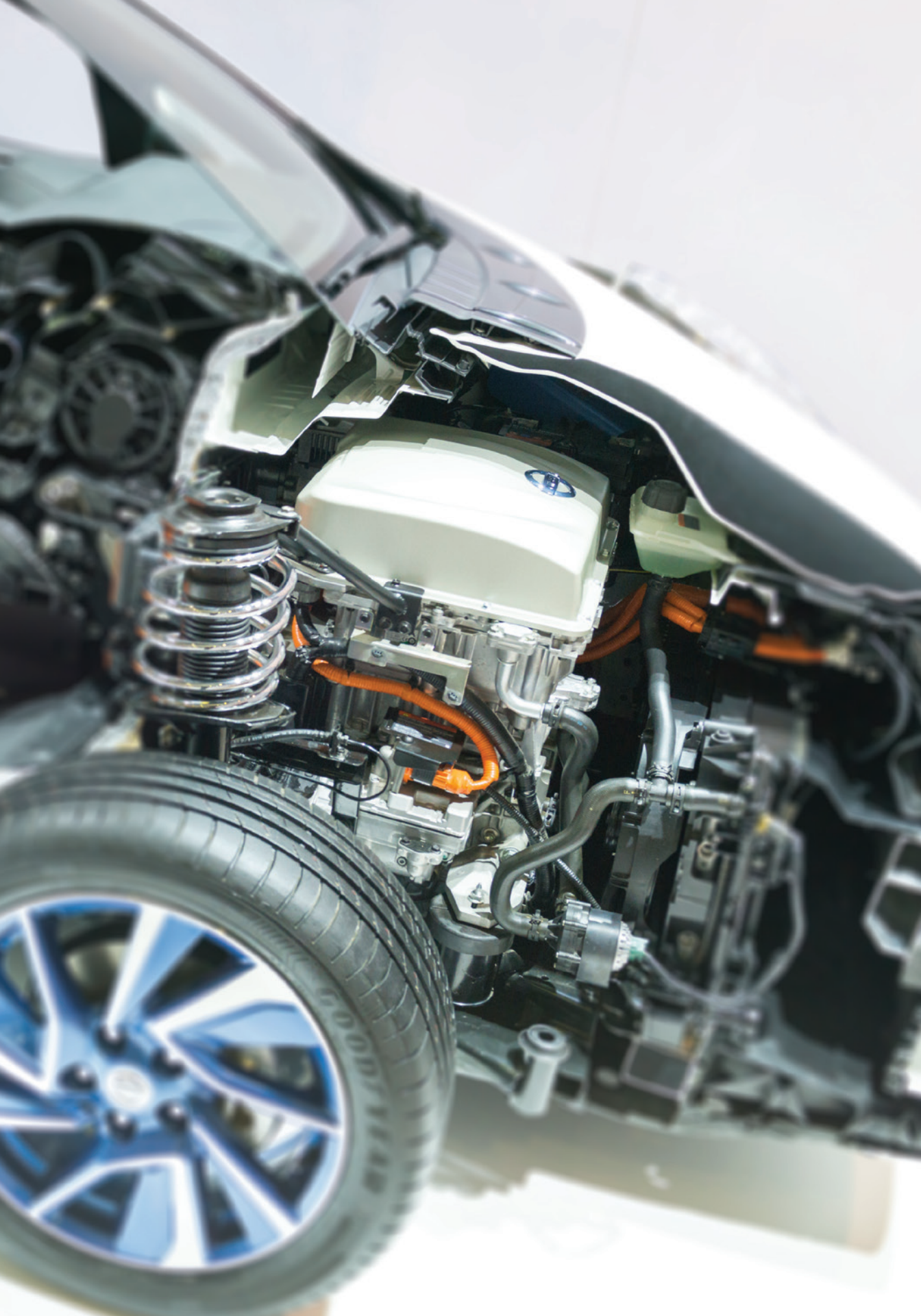
Conclusion.....53

References.....55

Additional Reading 61

Appendix A: Acronyms & Abbreviations..... 63

Appendix B: Summary of Prior Workshops & Event Outputs 65



Acknowledgments

The *Building a Materials Data Infrastructure* final report is a culmination of the efforts of globally recognized technical experts from across academia, industry, and government, who volunteered significant time to this study. Collectively, these experts participated in online meetings and teleconferences, in-person meetings, interviews, writing, editing, and reviewing this document.

Their dedication and involvement was foundational to this effort and we sincerely express our gratitude for their hard work and are confident that it will have a lasting impact on the community. The information enclosed in this report represents the collective compilation of their efforts.

Charles Ward, Study Team Chair
Justin Scott, Project Leader

Study Team Members

- **L. Cate Brinson**, *Jerome B. Cohen Professor of Engineering, Northwestern University*
L. Cate Brinson is currently the Jerome B. Cohen Professor of Engineering in the Mechanical Engineering Department at Northwestern University, with a joint appointment in the Materials Science and Engineering Department. Her research investigations involve characterization of local polymer mechanical behavior under confinement, nanoparticle-reinforced polymers, the phase transformation response of shape-memory alloys, nano- and microscale response of biomaterials, and materials genome informatics research. Brinson has served as an associate editor of the *Journal of Intelligent Material Systems and Structures* and the *Journal of Engineering Materials Technology*, spent two terms on

the National Materials Advisory Board of the National Academies, and has chaired two National Research Council Studies.

- **Giulia Galli**, *Liew Family Professor of Electronic Structure and Simulations, Inst. for Molecular Engineering, University of Chicago; Senior Scientist, Argonne National Laboratory*

Giulia Galli is the Liew Family Professor of Electronic Structure and Simulations at the University of Chicago's (UChicago) Institute for Molecular Engineering. She is also a senior scientist at Argonne National Laboratory (ANL) and is a senior fellow of the UChicago/ANL Computational Institute. Galli's research focuses on the development and use of theoretical and computational tools to understand and predict the properties and behavior of materials (solids, liquids, and nanostructures) from first principles. She is a Fellow of the American Physical Society and the American Association for the Advancement of Science, and has received an excellence award from the U.S. Department of Energy.

- **Surya R. Kalidindi**, *Professor, Woodruff School of Mechanical Engineering, and Lead MGI/ICME Strategist, Institute for Materials, Georgia Institute of Technology*

Over the past two decades, Surya R. Kalidindi's research efforts have made seminal contributions to the fields of crystal plasticity, microstructure design, and materials informatics. In addition to Kalidindi's professorship in the Woodruff School of Mechanical Engineering at the Georgia Institute of Technology (Georgia Tech), he holds joint appointments in the School of Materials Science and Engineering and the School of Computational Science and Engineering at Georgia Tech. He also serves as the lead Materials Genome Initiative (MGI)/Integrated Computational Materials Engineering (ICME) Strategist for Georgia Tech's Institute for Materials. He is a 2015 Fellow of TMS, as well as a Fellow of ASM International, the American Society of Mechanical Engineers, and Alpha Sigma Mu.

- **Apurva Mehta**, *Staff Scientist, Stanford Synchrotron Radiation Lightsource*

With more than 20 years of experience developing x-ray scattering and spectroscopy techniques, Apurva Mehta currently works as a staff scientist at Stanford Synchrotron Radiation Lightsource, a directorate of the SLAC National Accelerator Laboratory. Over the last three years, Mehta has been developing and collaborating with computer and data scientists to explore and adapt computer vision, statistical and machine learning, unsupervised data mining, and dynamic and active data management to materials science and x-ray measurements to bridge the widening gap between data collection and extraction of scientifically relevant information.

- **Bryce Meredig**, *Co-Founder and Chief Science Officer, Citrine Informatics*

Bryce Meredig co-founded Citrine Informatics to build a data-driven software platform for materials R&D and manufacturing, and to help translate the vision of the Materials Genome Initiative into practical industrial reality. As Citrine has grown to serve Global 1000 manufacturing leaders and individuals at 2,000 institutions worldwide, Meredig and the Citrine team have received over 25 invitations for talks, working groups, papers, and book chapters on the topics of materials data infrastructure and materials informatics. During his doctoral studies at Northwestern University, Meredig was awarded a Presidential Fellowship from Northwestern University and a National Defense Science and Engineering Fellowship, administered by the American Society for Engineering Education.

- **Jonathan Petters**, *Data Management Consultant, Virginia, Polytechnic Institute and State University*
Jonathan Petters provides research data management planning, training, and curation support to researchers across Virginia Polytechnic Institute and State University through the University Libraries. Previously he was in a similar role at Johns Hopkins University with the university's Data Management Services group. Petters was an American Association for the Advancement of Science (AAAS) Science and Technology Policy Fellow in the U.S. Department of Energy's Office of Science, where he investigated data management policies and needs within the physical sciences. During his Ph.D. studies in meteorology at the Pennsylvania State University, Petters researched and published on aerosol-cloud-radiation interactions in various cloud systems.
- **Brian Puchala**, *Assistant Research Scientist, PRedictive Integrated Structural Materials Science Center, University of Michigan*
Brian Puchala is an assistant research scientist and member of the PRedictive Integrated Structural Materials Science (PRISMS) Center at the University of Michigan. He is a lead developer of CASM, a first-principles statistical mechanical software package for the study of multicomponent crystalline solids, and is a domain scientist with The Materials Commons, an information repository and collaboration platform for the materials community. Puchala's research interests include development of both novel methods and basic infrastructure for integrated computational materials engineering (ICME), with particular focus on computational modeling of materials thermodynamics and the kinetics of solid-state atomistic processes.
- **Zachary Trautt**, *Material Research Engineer, Materials Measurement Science Division, National Institute of Standards and Technology*
Zachary Trautt is a Materials Research Engineer appointed in both the Office of Data and Informatics and the Materials Measurement Science Division, within the Material Measurement Laboratory at the National Institute of Standards and Technology. His primary duties are focused on improving the discoverability, reusability, and interoperability of materials data and metadata. Prior to this position, Trautt was a research assistant professor in the School of Physics, Astronomy, and Computational Sciences at George Mason University, where he researched the thermodynamics, kinetics, and mechanisms of motion of grain boundaries.
- **Vasisht Venkatesh**, *Materials Modeling Group Lead, Pratt & Whitney, Materials & Process Engineering*
Vasisht Venkatesh leads efforts in the development, implementation, and maturation of computational methods and analytical models across the materials and processes engineering discipline at Pratt & Whitney. Venkatesh also currently leads the U.S. Air Force-funded Foundational Engineering Problem on the ICME of Residual Stress in Ni-base Superalloy Rotors program. His experience is in the areas of microstructure-property relationships, materials characterization and testing, process monitoring, non-destructive evaluation development and application, and static and dynamic testing. Prior to joining Pratt & Whitney, he worked in TIMET's R&D lab, where he led key research activities to develop, validate, and implement numerical modeling tools to optimize various titanium alloy processes for microstructure, texture, and mechanical property enhancement.

- **Charles Ward** (Study Team Lead), *Lead for ICMSE, Materials and Manufacturing Directorate, Air Force Research Laboratory*

In addition to his role leading efforts in integrated computational materials science and engineering (ICMSE) at the Air Force Research Laboratory, Charles Ward is the co-chair of the Materials Genome Initiative Subcommittee under the National Science and Technology Council (NSTC) Committee on Technology. He is also an adjunct faculty member at the University of Dayton and is editor of the TMS journal *Integrating Materials and Manufacturing Innovation*. In a professional career spanning 30 years, Ward's research has focused on the microstructure-property relationships in titanium and titanium aluminide alloys. He is a Fellow of ASM International, and an active member of TMS, where he has served as the chair of the Materials Innovation Committee.

Expert Contributor Satellite Meeting – St. Charles, IL, July 10, 2016

- Paul Dawson, Cornell University
- Sean Donegan, BlueQuartz Software
- Dorte Juul Jensen, Danish Technological University/Risø National Laboratory
- Emmanuelle Marquis, University of Michigan
- Henning Friis Poulsen, Technical University of Denmark
- David Rowenhorst, Naval Research Laboratory

Expert Contributor Satellite Meeting – Chicago, IL, July 27, 2016

- James Barkley, UI Labs, Digital Manufacturing and Design Innovation Institute
- Laura Bartolo, Northwestern University
- Tia Benson Tolle, Boeing Commercial Airplanes
- Elif Ertekin, University of Illinois Urbana-Champaign
- Ian Foster, University of Chicago
- Dane Morgan, University of Wisconsin
- Peter Voorhees, Northwestern University

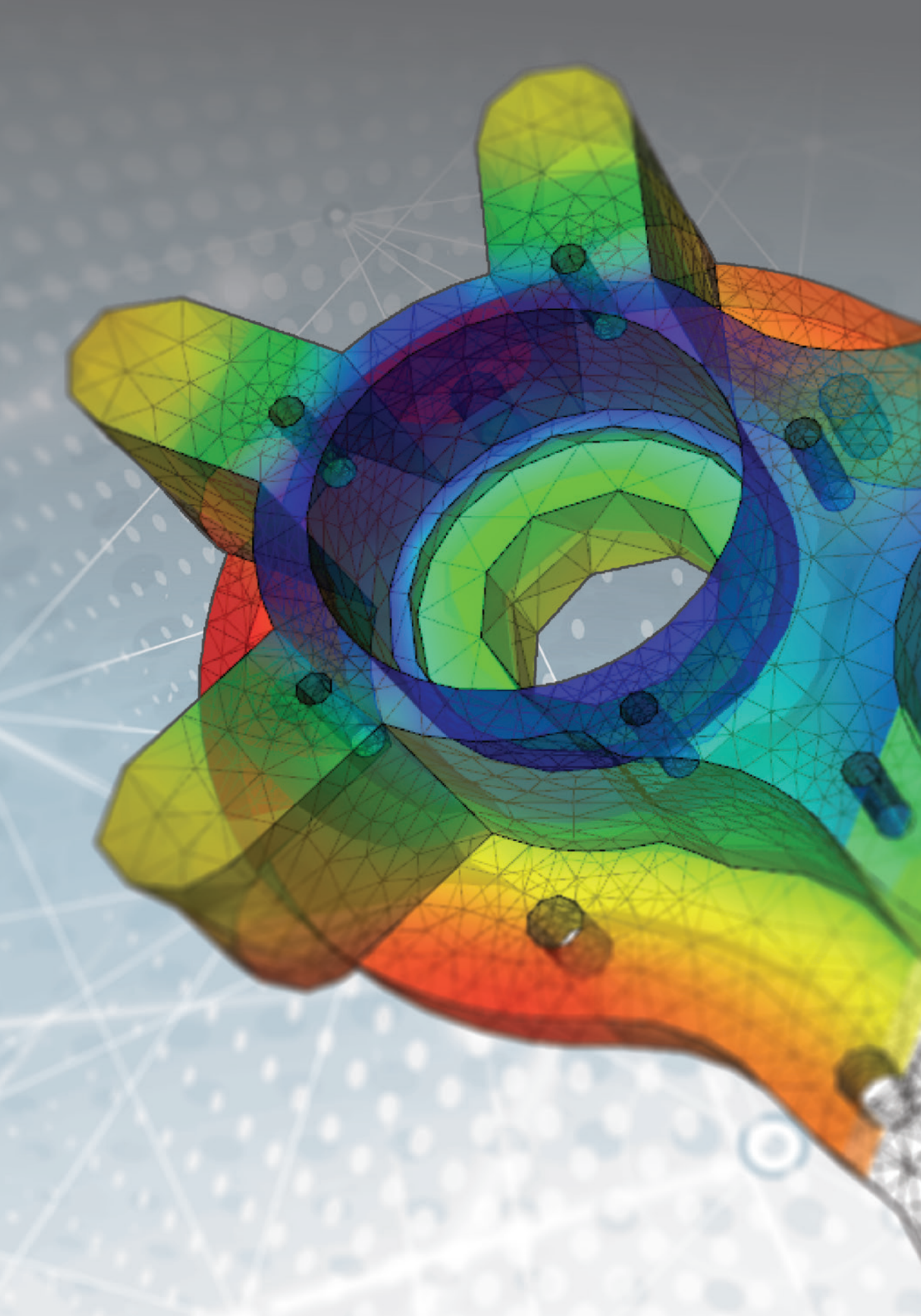
Subject Matter Expert Interviewees

- Parker Antin, University of Arizona (Life Sciences)
- Sky Bristol, United States Geological Survey (Earth-Observing Community)
- Anita de Waard, Elsevier (Publishing)
- Michael Madison, University of Pittsburgh (Legal)
- Nirav Merchant, University of Arizona (Life Sciences)

Final Report Review Team

- Chandler Becker, National Institute of Standards and Technology
- Paul Dawson, Cornell University
- Lan Li, Boise State University
- Dane Morgan, University of Wisconsin
- Joe Rasche, Rolls-Royce
- David Rowenhorst, Naval Research Laboratory
- Taylor Sparks, University of Utah
- Terry Wong, Aerojet Rocketdyne

This report was sponsored by the National Science Foundation under a grant (CMMI-1619577) administered to The Minerals, Metals, & Materials Society (TMS). The principal investigator of this activity was Justin Scott, Technical Project Leader at TMS. The technical experts were convened and their input compiled into report format through the efforts of TMS and Nexight Group, LLC (contracted by TMS) staff. Other important TMS staff contributions to this report include those of George Spanos (Technical Director), Laura Beringer (Technical Specialist), David Rasel (Media Manager), Bob Demmler (Graphic Designer), Maureen Byko (Editor, *JOM*), Shirley Litzinger (Production Editor), Marleen Schrader (Accounting and Human Resources Specialist), Matt Baker (Content Senior Manager), and Lynne Robinson (Communications Manager). Nexight Group, LLC staff members who were heavily involved in this effort include Ross Brindle (Chief Executive Officer), Jared Kusters (Technical Project Manager), and Changwon Suh (Technical Program Manager).



Preface

Who should read this report?

This report contains valuable information for anyone who has an interest in materials data infrastructure issues, particularly those issues associated with storing and sharing materials data. People with such interests are likely to include many in the materials science and engineering (MSE) community, as well as individuals from other related disciplines across academic, industrial, and government sectors. This includes any materials scientist or engineer who produces or uses technical data—activities undertaken by nearly all members of the field—and thus has a stake in where materials data are stored and how they are disseminated.^a Additionally, federal agencies, private enterprises, and other institutions that support and finance acquisition and use of materials data will also find this report useful. Many readers will especially benefit from the challenges and recommendations provided for building a robust materials data infrastructure (MDI). Beyond those experts who can directly contribute to and benefit from the MDI, other groups who would be interested in reading include more peripherally related professionals or students who want to learn more about key issues associated with storing and sharing materials data. Finally, those who might be engaged in guiding the data infrastructure of related science and technology areas in other disciplines outside of MSE may also acquire valuable insights from this report.

a In this context “materials data” primarily refers to data used in the discovery, design, development, and implementation of materials, materials processing, and materials innovations.

More specifically, this report will be of particular value to the following stakeholders:

- (1) materials data producers (scientists and engineers)
- (2) service providers related to materials data (e.g., data platform providers, scientific instrument providers, publishers, librarians, data service managers)
- (3) materials data users/consumers (including users of platforms and tools within the MDI)
- (4) managers and funding officers
- (5) policymakers
- (6) educators
- (7) students

How to navigate this report

Readers are encouraged to navigate this report by first examining the Executive Summary to get a general sense of the different parts of this document and how they might be of most relevance to your expertise, interests, and organization. It is our hope that this report will encourage you to take specific actions related to your skills and interests to support development and utilization of a strong MDI. The Background section provides insight into the current landscape, and the Challenges section will prompt you and your colleagues to think about the challenges most important to you and the challenges to which you may be able to contribute solutions. As you explore the Recommendations section, you can begin to focus on the tactical details that resonate most with your priorities, and you can think in more specific terms about the actions that you and your colleagues might undertake. Then perhaps you can begin to take some concrete steps toward initiating such activities, as discussed below.

What actions could be taken after reading this report?

A primary goal of this study report is to ***stimulate direct action by a wide variety of stakeholders who read this report***. These actions should be focused on contributing to the development and sustainability of a robust MDI, and achieving the benefits from such a MDI. After reading this report, some next steps to take action could include: (1) identifying specific challenge or tactical recommendation areas that you and your colleagues may address, and from which you and your colleagues would gain the most benefit, (2) sketching out a detailed action plan, and (3) taking concrete steps to initiate this activity. These steps would be different depending on your role and domain(s) of interest but may broadly follow the general approach below.

The pathway to action for materials data producers, service providers, materials data users, educators, and students could include some of the following steps: (i) give thought to how these actions fit into your existing programs, as well as to opportunities for obtaining support for new programs; (ii) if new support is required – write and submit a proposal to the appropriate funding body, perhaps using the knowledge in this report to help; (iii) set up any needed collaborations to implement the planned activity; (iv) begin implementation.

Other groups including managers, policymakers, and funding officers would follow different actions, perhaps including some of the following elements: (i) identify organizationally-appropriate focus areas and funding pathways; (ii) initiate implementation – assemble groups or programs and release new funding calls; (iii) establish or award programs to begin these activities.

TMS will take further action and plans to survey the recipients of this report after it is released to determine what resultant or related actions may have been initiated by readers of the report, or by any colleagues known to the readers. TMS can then assemble these inputs and explore ways to facilitate implementation and potentially organize activities that stimulate further development of a robust, coordinated MDI.



Executive Summary

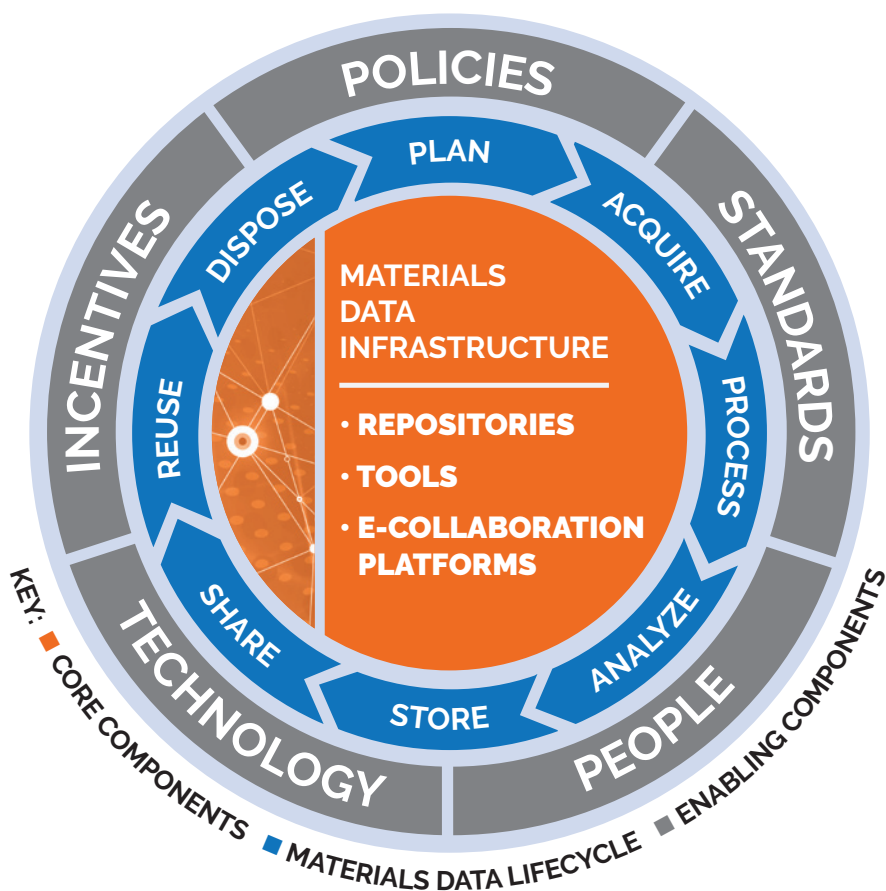
Motivation and Goals

The role of data in accelerating and lowering the cost of materials innovations and technological breakthroughs is increasingly acknowledged in the materials science and engineering (MSE) community. Yet many challenges remain in harnessing the power of an ever-growing amount of materials data, generated both computationally and experimentally. To unlock the great potential of such materials data, there is a strong need to develop a robust, integrated, materials data infrastructure. The experts on this study have defined a “materials data infrastructure (MDI)” as follows:

“The Materials Data Infrastructure (MDI) consists of three core digital components—repositories, tools, and e-collaboration platforms—as well as the technology, policies, incentives, standards, people, and activities necessary to plan, acquire, process, analyze, store, share, reuse, and dispose of materials data.”

This definition is displayed schematically in Figure 1. *Repositories* refer to the hardware, software, standards, and protocols needed to store and make materials data available. *Tools* represent the software that enable data handling and analysis, and bring value and reuse to datasets. Lastly, *e-Collaboration platforms* provide a mechanism for the community to share materials data more actively, and to work together to develop a robust MDI.

The overarching goal of this study is to provide knowledge and guidance, and motivate community action, to help further the development of a robust materials data infrastructure. More specifically, this report aims to: (1) identify major hurdles to long-term storage and sharing of materials data; (2) develop specific recommendations and tactics for overcoming these hurdles; and (3) stimulate near-term, active implementation of activities geared toward the development, use, and sustainability of the materials data infrastructure. Ultimately, it is the intent of this project to serve as a valuable resource as it works to overcome the hurdles of developing a data infrastructure to serve the materials community.



Reproduction of Figure 1. Schematic of the materials data infrastructure including the core components of repositories, tools, and e-collaboration platforms. The inner ring depicts the materials data lifecycle that leverages these components and the outer ring shows the enabling components that contribute to the MDI.

Study Process

To execute the goals mentioned above, an internationally recognized team of 10 experts was convened, drawing from multiple backgrounds across various technical sub-domains, materials data types, and professional sectors (academia, government, and industry). They gathered in online meetings and two professionally facilitated, two-day, in-person meetings, and worked remotely throughout the process for the content development, writing, and editing of the final report. In addition to the work of the study team, two satellite meetings were held with separate groups of experts, to further explore key concepts initially discussed by the lead study team. Additionally, other experts were interviewed concerning issues related to law, publishing, and progress in developing data infrastructures in other scientific disciplines. (See the Acknowledgments section for the names and affiliations of the specific individuals who participated, beyond the lead study team members.)

Who Should Read This Report and What Actions Should Be Taken?

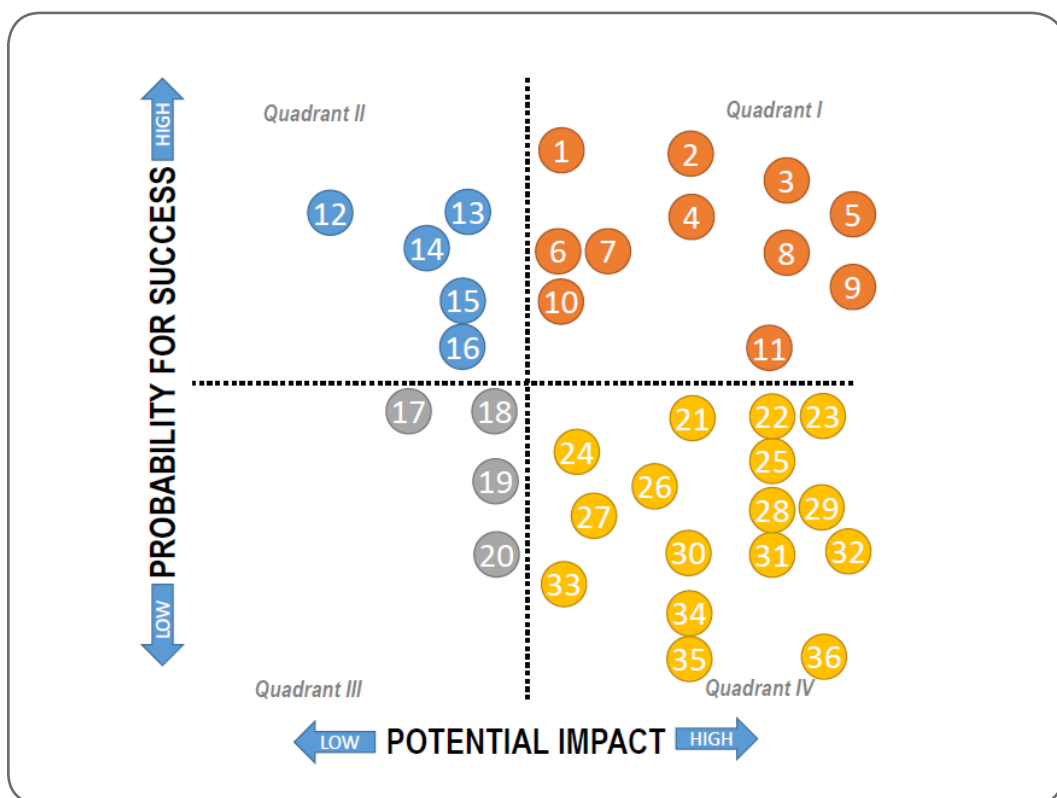
This report contains valuable information for anyone in the Materials Science and Engineering (MSE) community or related disciplines who has an interest in materials data infrastructure issues, including individuals from across academia, industry, and the government. A primary goal of this study is to ***stimulate direct action by a wide variety of stakeholders who read this report***, ultimately for the purposes of contributing to and benefiting from a robust materials data infrastructure (MDI). Although some suggestions of how one might begin taking such actions are provided in the Preface, tactical details for specific recommended actions are provided in the Recommendations section.

Background

Many past efforts including projects, publications, and workshops have examined the development of data infrastructures including materials-specific components. The Background section attempts to provide some context and summarize previous work on relevant science and engineering data issues, as well as materials-specific themes that have been discussed in other workshops and activities. This information is conveyed through a summary of key elements from various technical publications, reports, and workshops with more detailed information on specific activities provided in Appendix B.

Challenges

Thirty-six challenges associated with building a robust MDI were identified and are discussed in the Challenges section of this report. Challenges were organized and are presented in Figure 2 in the form of a plot of relative probability of success vs. the potential impact (of overcoming that challenge). This summary figure is reproduced below. It is noted that this plot should not be considered quantitative, but instead provide useful qualitative guidelines for consideration.



Reproduction of Figure 2. A plot of the relative probability for success vs. the potential impact of overcoming the 36 challenges identified by the study team.

Some representative examples of the challenges, which are discussed in more detail in the report, and their corresponding numbers from the plot above include: lack of e-collaboration platforms and tools (#3); insufficient participation of the computer science community in the MDI (#7); insufficient mechanisms for crediting data contributors (#9); poor integration of resources among data infrastructure providers (#15); constraints of government technology export regulations (#20); lack of a clear, unified vision of how the MDI will benefit the community (#23); and need for standardized components and documented workflows to enable data extraction and reuse (#27).

Recommendations

Eight overarching recommendations are provided in the Recommendations section of this report, along with multiple tactics and detailed suggestions for implementation. The overarching recommendations and tactics are summarized in Table II, which is reproduced below.

Table II: Priority Recommendations
Recommendation 1: Strengthen the MDI core in repository, registry, and tool development <ul style="list-style-type: none"> • <u>Tactic #1</u>: Develop and deploy robust repositories • <u>Tactic #2</u>: Develop and deploy registries for MDI repositories and tools • <u>Tactic #3</u>: Develop analytical and visualization tools that enhance the speed and capabilities of materials data use and analyses • <u>Tactic #4</u>: Launch and sustain e-collaboration platforms • <u>Tactic #5</u>: Develop automated data/metadata capture tools for scientific instruments
Recommendation 2: Sustain and grow MDI-dedicated funding programs <ul style="list-style-type: none"> • <u>Tactic #1</u>: Support the MDI community by leveraging and coordinating current federal programs • <u>Tactic #2</u>: Develop and launch new funding programs
Recommendation 3: Create, execute, and monitor incentive mechanisms <ul style="list-style-type: none"> • <u>Tactic #1</u>: Establish incentive mechanisms for materials data sharing • <u>Tactic #2</u>: Execute and monitor the incentive programs to energize and sustain MSE community involvement
Recommendation 4: Develop demonstration projects and cross-disciplinary community efforts that enhance and accelerate adoption of the MDI <ul style="list-style-type: none"> • <u>Tactic #1</u>: Establish materials-data-driven design projects that enable property prediction, to enhance MDI adoption • <u>Tactic #2</u>: Establish materials-data-driven projects for manufacturing, to enhance MDI adoption • <u>Tactic #3</u>: Launch targeted community efforts to help achieve MDI critical mass in the MSE community • <u>Tactic #4</u>: Fill gaps in the MDI materials data domain
Recommendation 5: Establish a MDI ecosystem and business cases <ul style="list-style-type: none"> • <u>Tactic #1</u>: Develop a reference architecture concept for the MDI • <u>Tactic #2</u>: Develop and demonstrate business cases for data storage and sharing • <u>Tactic #3</u>: Explore concept of a materials data “app store”

Recommendation 6: Develop and invest in education and training programs for the MDI workforce (providers and users)

- Tactic #1: Integrate MDI into existing MSE curricula and build new cross-disciplinary curricula
- Tactic #2: Conduct outreach and training programs for professionals

Recommendation 7: Create MDI consortia and working groups:

- Tactic #1: Create a long-term coordinating and advisory body comprising relevant MDI stakeholders
- Tactic #2: Create Community of Practice (CoP) groups around the MDI

Recommendation 8: Define and establish clear policies and guidelines associated with the MDI

- Tactic #1: Establish an interagency (federal) council to foster consistent data preservation and sharing policies
- Tactic #2: Increase emphasis on data management plans for funding support considerations

For each recommendation and tactic there is a discussion in the Recommendations section which includes actions and details such as tasks, timeframes, recommended implementers, estimated costs, and possible sources of financial support to accomplish these activities. The particular tactics suggested in this study should not be viewed as all-inclusive. Readers of this report are strongly encouraged to use this knowledge to identify new tactics and actions geared towards overcoming the challenges and accomplishing the high-level recommendations identified in this study.

Introduction and Motivation

Increasingly, the materials community is acknowledging the role of data in accelerating the pace of scientific and engineering progress across sectors ranging from national security and healthcare to manufacturing, energy, and beyond. Yet many challenges remain in harnessing the power of an ever-increasing amount of data generated—both computationally and experimentally—by materials research and development efforts. Much of the data produced through the course of research and development is often not reused, typically because of the added resources required and the perceived lack of benefits associated with preserving and sharing data. The majority of data is still commonly stored locally, with little to no access by outside groups despite recent requirements for data dissemination and sharing from federal funding agencies. This makes it difficult to truly leverage data in sophisticated analyses, models, or experiments that may benefit from a more open data-sharing environment. Many agree that an open, collaborative approach to materials data is the key to unlocking new research questions and speeding up materials discovery and innovation.¹⁻⁵ Over time, a number of individuals and groups have come to describe how materials data could be stored and accessed through a “materials data infrastructure (MDI),” which the study team defined as follows: The Materials Data Infrastructure (MDI) consists of three core digital components—repositories, tools, and e-collaboration platforms—as well as the technology, policies, incentives, standards, people, and activities necessary to plan, acquire, process, analyze, store, share, reuse, and dispose of materials data.^b

b In developing this definition of the Materials Data Infrastructure, the study team built on information contained in the Office of Management and Budget (OMB) Circular A-16 concerning “Coordination of Geographic Information and Related Spatial Data Activities.”⁶ Readers are referred to <http://dictionary.casrai.org> for definitions of other terms used throughout this report.

One of the early acknowledgements of the importance of building a data infrastructure for materials science and engineering can be found in the 2008 U.S. National Research Council report *Integrated Computational Materials Engineering (ICME): A Transformative Discipline for Improved Competitiveness and National Security*. As part of its conclusions, the report emphasized the need for a cyberinfrastructure to enable the framework of ICME along with “a widely accepted taxonomy, an informatics technology, and materials databases openly accessible to members of the materials research and development, design, and manufacturing communities.”⁷ The Materials Genome Initiative (MGI), launched in June 2011, further recognized that an accessible, extensible, scalable, and sustainable data infrastructure is needed to accelerate materials discovery and development. As described in the initial whitepaper¹ and underscored in the associated strategic plan document,² a means for storing and sharing materials data is essential to the “materials innovation infrastructure” associated with the MGI. In fact, one of four key challenges specified in the MGI strategic plan was “Access to Digital Data.”² In addition to storage, the MGI strategic plan referenced the need for accessibility and discoverability.

“The Materials Data Infrastructure (MDI) consists of three core digital components—repositories, tools, and e-collaboration platforms—as well as the technology, policies, incentives, standards, people, and activities necessary to plan, acquire, process, analyze, store, share, reuse, and dispose of materials data.”

Concurrently, the global community of materials scientists and engineers has also been attempting to address the need for a data infrastructure. Examples include the EU Framework Programme 7, which funded a project, Opportunities for Data Exchange, which intends to enable e-science through the collection of emerging best practices in sharing, reusing, preserving, and citing data.³ In addition, it also funded the development of the Integrated Computational Materials Engineering Expert Group (ICMEg), which as part of its charter aims to “facilitate the exchange of data between different tools.”⁴ Another example of a global community undertaking is the Future of Research Communications and e-Scholarship (FORCE11) effort in improving infrastructure standards. Members of FORCE11 have created a set of guidelines to make data sharing and storage accessible to multiple scientific communities, with an emphasis on the ability of computers to automatically find and utilize data.⁸ (See the Background and Prior Efforts section for additional detail on this activity.)

A global survey conducted jointly by the Materials Research Society (MRS) and The Minerals, Metals & Materials Society (TMS) in 2013 identified some of the key motivations for data sharing in support of open research. As part of this activity, MRS and TMS polled nearly 700 materials scientists and engineers and found that the top three reasons for sharing data in open-access formats were: increased visibility of research/work (72% of respondents); the opportunity to receive feedback (67% of respondents); and the opportunity for others to analyze the data, potentially making other discoveries as a result (54% of respondents).⁵

In the United States, federal agencies have been tasked with improving public access to federally funded research results. These requirements were specifically detailed in a February 2013 memorandum to federal agencies from John Holdren, Director of the White House Office of Science and Technology Policy (OSTP) titled “Increasing Access to the Results of Federally Funded Scientific Research.”^{9,10} Agencies including the Department of Energy (DOE), National Science Foundation (NSF), Department of Defense (DoD), National Aeronautics and Space Administration (NASA) and National Institute of Standards and Technology (NIST) have released plans that address the memorandum’s key objectives.^{11–14}

The compilation of activities, trends, and policies above point to a growing recognition of the benefits that a materials data infrastructure could provide the materials community. More specifically, some of the potential benefits of such an infrastructure could include:

- **Accessibility** – Data would be easily discoverable and more accessible to a wider audience.
- **Citability** – Data attribution would be easily enabled to encourage participation.
- **Collaboration** – New and existing collaborations would be facilitated across technical domains and geographic regions.
- **Flexibility** – The infrastructure would be flexible and extensible to accommodate changing needs.
- **Reliability** – Provenance details including pedigree and history of data would be provided; versioning and updating should be seamless.
- **Repurposing** – Computational and experimental data would have a greater impact because they would be usable by many. Disparate datasets could also potentially lead to new discoveries through improved pattern identification.
- **Sustainability** – Long-term archiving of datasets would be made easier.
- **Security** – Sharing and access would be easily controlled to prevent error and fraud while protecting intellectual property.
- **Tools** – Tools for standard analyses as well as complex data analyses would be available.
- **Workflow** – Large datasets with multiple terabytes would be easily transferred, and data could be automatically uploaded as part of a “digital lab notebook.”

Though this is by no means a complete list, it begins to illustrate the value of developing a materials data infrastructure and how it could have a substantial impact on future materials research.

Members of the study team identified three core digital components of the materials data infrastructure, as shown in Figure 1. **Repositories** are the first pillar of the infrastructure and consist of the hardware, software, standards, and protocols needed to store and make materials data available. **Tools** are the second pillar and are defined as software that enables data handling, service, and analysis. Tools help bring value to a dataset by performing tasks such as pattern identification and other data analyses. Lastly, **e-Collaboration platforms** are the third pillar and are a mechanism for a team, laboratory, or community to actively work together with shared datasets. Note that this collaboration does not necessarily have to include the original data provider, and it goes beyond simply sharing data.

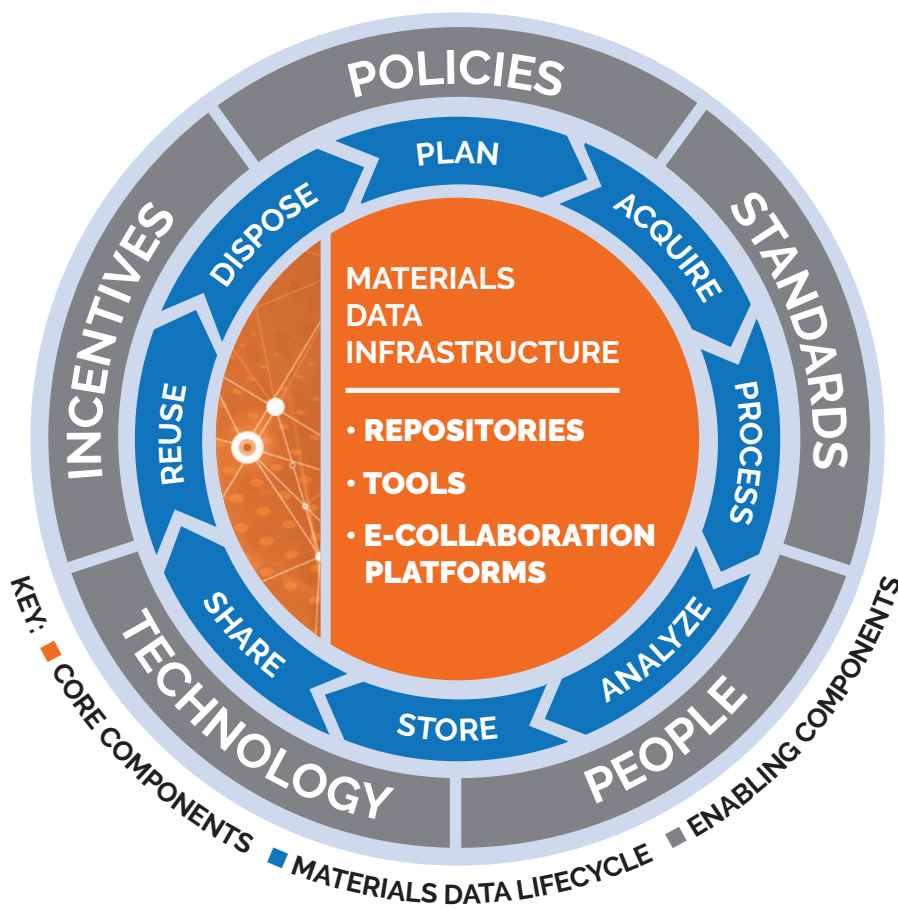


Figure 1. Schematic of the materials data infrastructure including the core components of repositories, tools, and e-collaboration platforms. The inner ring depicts the materials data lifecycle that leverages these components and the outer ring shows the enabling components that contribute to the MDI.

Project Goals and Process

To further the development of a robust materials data infrastructure and help guide the community, this project aimed to accomplish the following: (1) identify major hurdles to long-term storage and sharing of materials data; (2) develop specific recommendations for overcoming the hurdles that are identified; and (3) begin promoting implementation of recommended actions towards the development of a materials data infrastructure. Ultimately, it is the intent of this project to serve as an authoritative resource for the community as it works to overcome the major hurdles of developing a materials data infrastructure.

To execute this charge, an internationally recognized expert study team was formed, drawing from multiple backgrounds across various material classes, technical sub-domains, and work sectors. Expert backgrounds included individuals in a small business start-up, a large-scale industrial corporation, federal laboratories, and academia. As evidenced in the Acknowledgments section of this report, the study team thus represented a wide variety of key stakeholder groups. In addition, multiple other experts participated through satellite meetings, interviews, and an independent review process, to contribute key additional insights.

Throughout the course of this project, the study team was convened via teleconferences and assigned homework to help prepare them for the in-person meetings. Two-day, in-person working meetings were held in June and November of 2016, guided by professional facilitators from Nexight Group. The facilitators worked in close coordination with TMS staff to develop the working agendas, which included sessions on defining key terms, articulating benefits of the MDI, identifying major challenges, and making specific recommendations for implementing a robust MDI.

In addition to convening the study team, two satellite meetings were held with different groups of experts to further explore key concepts initially discussed by the study team. The first of these meetings was held on July 10, 2016 in St. Charles, IL and focused on perspectives from the 3D materials science community, as well as developing tactics to implement recommendations of interest to this group. The second meeting was held on July 27, 2016 in Evanston, IL and was intended to help articulate current methods of storing and sharing data, identifying key benefits of the materials data infrastructure, and methods of ensuring long-term sustainability of data and associated tools and platforms.

Several experts were also interviewed on issues including law, publishing, and progress in developing data infrastructures in other scientific disciplines. Candidates for these interviews and topics of interest were generated by the study team. The interviews were conducted by TMS staff.

Outputs were captured from each of the aforementioned activities and synthesized in this final report. In addition to editing by the study team, it has also been reviewed by a separate, independent group of experts as listed in the Acknowledgments. While every effort was made to include a wide array of backgrounds from within and outside of the materials community, it is important to note that many of the specific examples and case studies provided in this report are intended to be illustrative of ongoing activities. Other examples are given in the additional reading list, references, and appendices provided in this report.



Background and Prior Efforts

This section aims to provide the reader with a general background on some key groups, concepts and issues associated with science and engineering data, and to summarize many of the materials-specific themes that have been discussed to date in other workshops and activities.

Data in Science and Engineering

The concept of centralized data infrastructures has been discussed at length in a variety of scientific disciplines. (See, for instance, the sidebar on “Lessons Learned from Other Disciplines” on page 8 for examples of two communities that have greatly advanced their data infrastructures.) One forum where discussions have helped shape many fields is within the Future of Research Communications and e-Scholarship (FORCE11) group. FORCE11 is an open community of researchers, publishers, librarians, and funders that desires to bring about change to scholarly communication through effective information technology use. In its vision document, FORCE11 describes key data challenges that are shared by multiple scientific communities.¹⁵ These challenges include verification and validation of data, intellectual property restrictions on publishing and data sharing, and a lack of data-sharing incentives, to name a few. Professor Barend Mons of Leiden University, the Netherlands, is an internationally recognized thought leader in sharing scientific data who conceived the FAIR data principles (Findable, Accessible, Interoperable, Reusable) to promote good data stewardship in science.⁸ These guidelines have been crafted by leading researchers in various disciplines involved with data science to enable a variety of technological innovations, collaborations, and behaviors to advance the practice of data stewardship.^{15,16} The intent is that researchers will employ FAIR principles so that eventually there is a standard protocol for sharing and storing data in various scientific and engineering communities.

Similarly, the Research Data Alliance (RDA) is an international organization that works across disciplines and is dedicated to identifying challenges and forming action plans to overcome hurdles associated with building and sustaining digital infrastructures. The benefits to building a data infrastructure are stated in the RDA mission: “building the social and technical infrastructure to enable the open sharing of data.”¹⁷ The RDA has created working groups across and within research disciplines to tackle the variety of topics related to the creation of data infrastructures and their officially endorsed outputs are available online.¹⁸ The output topics have a broad range; some examples include creation of a data description registry, repository audit and certification features, and workflows for research data publishing. Of note is the RDA Interest Group on Materials Data, Infrastructure and Interoperability.¹⁹

Lessons Learned from Other Disciplines

Multiple disciplines within science and engineering have worked to develop data infrastructures that serve their communities. Two examples are presented below and provide some key insights into how other groups have addressed common issues.

Geosciences and the Earth-observing Community

The U.S. Group on Earth Observations (GEO) is an interagency council chartered as a subcommittee under the National Science and Technology Council (NSTC). One of the group’s goals is to improve Earth system data management and interoperability, which has led to multiple activities and a final report that details the “Common Framework for Earth-Observation Data.”²⁰ Contained within the document are recommendations and guiding principles for data managers to help improve the discoverability, accessibility, and usability of Earth-observation data.

As described by Sky Bristol, a member of the U.S. GEO Data Management Working Group that created the Common Framework document, there were various lessons learned through creating the framework and establishing a unified data infrastructure.²¹ Some examples include:

- *The establishment of “Societal Benefit Areas”²² (i.e., a set of the primary environmental fields of interest) were a key driver for the community. Not only did they provide a means of categorization, but they also enabled community members to better identify how data infrastructure could contribute at local, regional, national, and international scales.*
- *Champions from relevant federal agencies were instrumental in making sure that work continued on developing a common framework of interest to the community.*
- *It was essential to receive community input through conferences, professional societies, and workshops as well as existing groups that had already been formed at agencies such as NASA, the National Oceanic and Atmospheric Administration, and the United States Geological Survey.*
- *The group had to be persistent as the infrastructure took extensive effort over a long period; it was helpful to focus attention on early adopters when the process was just starting out.*
- *Ontology development can be painstaking; while useful, it should not though hold up the process of developing a data infrastructure. The Research Data Alliance (RDA) has been an important component of discussions relating to ontology development.*
- *Another motivating factor is the potential for decision analysis tools that could be enabled by a robust data infrastructure.*

Plant Sciences and Life Sciences

CyVerse²³ is an NSF-funded project that originally aimed to serve the U.S. plant sciences community. It was conceived as iPlant²⁴ in 2008, but has evolved into its current mission to design, deploy, and expand a national cyberinfrastructure for life sciences research. In addition to data storage, CyVerse includes access to a web-based analytical platform, cloud infrastructure for using remote computers, security services, and education and training resources.

Through a discussion with CyVerse staff members Parker Antin (principal investigator (PI) and Project Director, University of Arizona) and Nirav Merchant (Co-PI, University of Arizona), various lessons were shared that broadly apply to the design, development, and deployment of scientific data infrastructures:²⁵

- *While the plant sciences community had specific needs, it was important for the infrastructure to be flexible and not overly specific in its early stages. Common standards were thus useful but purposefully not overly prescriptive.*
- *Onboarding the community was a delicate exercise as some members have in the past invested time in unsuccessful data infrastructure development projects, with no return on that investment. Thus, it was important to do things incrementally and show value to the user's workflow over time. To ease the transition, CyVerse intentionally makes it easy for users to bring their data to the platform or remove it from the platform. This also helps the community avoid getting the impression that their data is simply being leveraged to help CyVerse retain funding.*
- *It was important to work with existing groups that already had piloted projects and were willing to extend their work. For example, iPlant leveraged iRODS (federated storage) and Condor (job scheduling), which already had their own user communities. That meant it was not necessary to create the entire data infrastructure from the ground up.*
- *Good cyberinfrastructure is data agnostic. CyVerse is thus built in a way that it could support materials (or other) data with minimal domain-specific modification.*
- *Metadata was previously viewed as a chore by the community when it was used as “a stick” and was a requirement for upload, but was not well enforced. Later, it was transitioned into “a carrot” as members of the community saw how they could benefit and improve access when they uploaded metadata. This shifted the culture and ultimately improved community response to uploading metadata.*

Ontologies are useful but a lightweight, simplified approach is recommended at first. Over time, refinement of these ontologies will lead to improvements in the search and discoverability of data.

Data Lifecycle

To create a robust data infrastructure that enables useful data to exist beyond the typical timescale of a research project, it is important to understand the various stages of a data lifecycle as this will drive requirements. While many models and concepts exist (see Ref. 26) the U.S. Geological Service's (USGS's) Science Data Lifecycle Model particularly resonated with the study team.²⁷ The team slightly modified and expanded the USGS model to consist of the following key stages, using many of the same definitions of the stages as the USGS for consistency.^c

c The interested reader is invited to explore the USGS Data Management website at <http://www.usgs.gov/datamanagement/> for a detailed description of data management practices.

- **Plan:** Prior to starting a research effort, a Data Management Plan (DMP) should be developed to ensure the effort's data is available at the right time to the people who need the data, that they are able to find the data in a form that is useable to them, and that the data is available for use by others after the program ends. A good DMP provides a strategy for implementing the FAIR principles across the remaining steps of the data lifecycle. More complex efforts and greater participant involvement will only increase the importance of having a good DMP for the overall success of the effort.
- **Acquire:** Data needed in a research effort can be acquired by many means, including observation, instrumentation, computation, extraction from existing open sources, or purchase.
- **Process:** Processing covers any set of structured activities resulting in the alteration or integration of data. Process components support validation, transformation, subsetting, summarizing, integration, and derivation, among activities. Data processing can result in data ready for analysis, or generate output such as graphs and summary reports. Documenting the steps for how data are processed is essential for reproducibility and improves transparency.
- **Analyze:** The analyze stage represents activities associated with the exploration and assessment of data, where hypotheses are tested, discoveries are made, and conclusions are drawn. Data analysis may also be less targeted and deal with interpretation of data to better understand content, context, and quality. In this stage of the data lifecycle, conclusions or new datasets are generated and methods are documented. Analytical activities include statistical analysis, spatial analysis, modeling, visualization, image analysis, and interpretation.
- **Store:** Storing data involves actions and procedures to keep data for some period of time and/or to set data aside for future use, and includes data archiving and/or data submission to a data repository. A primary goal in this stage should be to preserve well-organized and documented datasets that support research interpretations that can be re-used by others.
- **Share:** The ability to prepare, release, and share quality data with others is an important part of the lifecycle process. The data should be format and platform agnostic, with an understanding that transfer may occur via automated or non-automated mechanisms. Controls must be in place to protect proprietary or export controlled data as well as the integrity of the data itself. Data sharing also requires complete metadata to be useful to those who are receiving the data.
- **Reuse:** Data that have been subjected to good stewardship along the data lifecycle to this point are best positioned to add to scientific and engineering progress, either by the data producer or consumer. Reuse of materials data is a primary objective of a developing a robust MDI. The knowledge gained from reuse is ideally contributed back to the community.
- **Dispose:** It is extraordinarily difficult at this stage of the MDI's maturity to provide suitable guidelines for the useful lifetime of materials data. The community is not currently limited by storage space, but that may change as more materials data is shared publicly. Factors to consider in determining data's "shelf life" include the difficulty in obtaining the data and the richness of the data.

Whereas the data lifecycle describes the general, essential stages in data management, a data workflow is an application of the data lifecycle through the repeatable sequence of actions that are used for a specific experiment or computation. Examples of materials-specific workflows are described in the "Workflow" sidebar on page 12.

Data in the Materials Community

Much of the recently published research and many of the workshops specifically addressing materials data infrastructure are associated with the 2011 Materials Genome Initiative (MGI), first announced by U.S. President Barack Obama in 2011.¹ An overarching goal of the MGI is to greatly reduce the time and cost of developing new, advanced materials innovations. Within the MGI vision statement, there are three main thrusts: digital data, computational tools, and experimental tools. These three thrusts comprise what is termed the “materials innovation infrastructure.” Access to digital materials data, from both computational and experimental methods, is intended to facilitate collaboration between various scientific communities. In 2014, the MGI strategic plan was developed, in which facilitated access to materials data is one of the four overarching goals; more detailed objectives and milestones within that goal are outlined in the strategic plan document.²

Nearly a decade prior to the announcement of the MGI, physical metallurgist Robert W. Cahn described the history of materials data infrastructure in his book titled *The Coming of Materials Science*.³¹ Cahn described the very first materials science databases (or infrastructures) as detailed scientific volumes that included chemical properties tables and phase diagrams, developed by scientists as early as the nineteenth century. In the early 1900s, crystallographic structures were being discovered at a rate frequent enough to warrant specialized journals. The information captured in these journals later evolved into the early crystallographic databases of the 1960s, such as the Cambridge Crystallographic Database. As noted by Cahn, scientists and engineers have been cataloging scientific and technical data, ranging from chemical compositions, crystal structures and synthesis recipes to physical properties, for more than 200 years. However, without a set of standardized protocols for capturing, describing, storing, discovering, retrieving, and sharing this data digitally, valuable information becomes harder to find and becomes effectively lost. In order to fulfill the visions detailed in more recent documents like the MGI strategic plan, the materials community must work together to better manage the vast quantity of data created as a result of significant investment over the past half century in performing innovative research and development. Without a well-designed data infrastructure in place, there is no effective and sustainable mechanism to store, share, discover, and reuse the vast amount of data generated by the large and still growing community of materials scientists and engineers. An early attempt at aspects of data acquisition and storage for materials scientists and engineers occurred with the D3D program supported by the Office of Naval Research. For example, Boyce et al. describe the design of software to share and archive materials data in their 2009 paper, which was an output from the D3D program.³²

Choosing the type of digital architecture for a materials data infrastructure and the requirements for implementing it is not straightforward and many research groups have proposed various frameworks and elements that should be captured. For example, Dima et al., in their 2016 paper,³³ outlined two crucial requirements for such an infrastructure in order to accelerate materials discoveries. These include an interoperable platform that supports modular community standards and a decentralized infrastructure that enables the identification and sharing of materials resources.³³ To achieve these requirements, Dima and colleagues recommended a Python-based framework that uses extensible markup language (XML) documents to store both data and metadata.³³

Workflow: Data Infrastructure Considerations

Using the MDI necessarily results in changes to the scientific workflow. In particular, some additional planning and preparation must be done to determine how data will be captured. In return, the modern MDI can make it significantly easier to analyze, interpret, share, publicize, and reuse data. Exemplars are provided here to illustrate two different use cases: (i) atomistic simulations of grain boundaries, and (ii) parameterization of structure-property relationships for a metal alloy, with emphasis on microstructural data.

The study of grain boundaries, or material interfaces more broadly, is intrinsically inclined to produce a great number of unique data points due to the degrees of freedom contained therein. Grain boundary properties vary with five geometric degrees of freedom, as well as temperature, composition/impurity content, etc. Within atomistic modeling of grain boundaries, many parameters such as interatomic potential or simulation setup/thermalizing procedure can significantly affect the results. Therefore, researchers in this domain may particularly benefit from adopting the MDI and are among some of early adopters.

As a second example, we consider the parameterization of structure-property relationships for a metal alloy, and in particular the capture and handling of microstructural data. Some historically useful measures of microstructure include composition, phase fraction, precipitate size and shape, grain size and orientation distribution, and 2D or 3D imaging. While useful in many cases, these are fairly low order representations of microstructure. In general, a representation of structure can be obtained by classifying structural features and then calculating spatial correlations among those features. This representation can then be used to fit structure-property relationships, validate computations, and infer material parameters, for example.

The first step in a workflow that makes use of the modern MDI involves planning how to represent the work that is being done by breaking the project into a discrete set of experimental and computational process types. The processes should include any important processing steps, sample preparation steps, experimental techniques or computations, and data analysis steps. Then, for each process type, the researcher must plan how to capture data. This includes deciding what data will be captured, where it will be stored, and what metadata must be captured. Generally this only needs to be done once per process type and is encapsulated in some software that integrates a particular type of data, experimental instrument, or computational software with a particular MDI provider, so that the data then can be reused by other users.

For the example of atomistic modeling of grain boundaries, this may involve writing a script which captures the code version and execution environment, parses input files for simulation input parameters, and uploads this data along with generated results files directly to the MDI platform as jobs are run. Historically, data and metadata associated with grain boundary property simulations were made available primarily via plain text files as supplementary material in a unique format developed by the author.^{28–30} Early adopters within this domain are now incorporating materials data repositories based on self-describing file formats (e.g. XML, JSON, HDF5, etc.) in the early stages of their research. To enable better interoperability of research data and metadata, researchers within this domain convened a workshop to begin to develop community standards for the self-describing file formats. Next, researchers are incorporating local (internal/offline) materials data repositories, which are designed to enable storage, access, and precise queries of the self-describing file formats and any other linked files (e.g., simulation input/output files). This adoption promises to help streamline management of both data and simulations.

For the example of microstructural data, the workflow might be broken into three steps: (1) processing and sample preparation, (2) scanning electron microscopy (SEM) and transmission electron microscopy (TEM) experimental measurement, and (3) data analysis. For the first step, the data capture should include basic material, processing, and sample preparation parameters (i.e., metadata) associated with each sample. The plan could be as simple as a standard spreadsheet template, or it could be a standardized corporate or laboratory electronic notebook or web form associated with a MDI provider. For the second step, data capture might be performed by a script that automates upload of SEM and TEM images, along with instrument settings and information such as sample identity and orientation, to the MDI provider. For the third step, it might be sufficient to store a set of image analysis scripts in a git repository. The set of input parameters used for thresholding, classification, calculating correlations, etc. would reside with the MDI provider, along with the git commit command of the script used to do the analysis.

Once these data capture preparations have been made, experiments and calculations can be performed, and data acquired can be uploaded to the chosen e-collaboration platform. In some cases it may be desirable that all data captured is automatically stored within the e-collaboration platform as soon as it is generated; in other cases it may be that some data processing, curation, or analysis is done first, depending on the amount of data, time, and cost for re-acquisition, etc.

Similar to the requirements pointed out by Dima et al., Seshadri and Sparks noted that a materials data infrastructure needs to be open-access for indexing and searching and must have the capability to automatically integrate data from primary literature sources.³⁴ More specifically, Seshadri and Sparks described their interactive materials databases for thermoelectrics and batteries, where users can select materials properties or other metadata for plotting predictive properties graphs.^{35,36} In this same paper,³⁴ the authors noted that the most successful present day databases adopted a standardized format for reporting properties or information. For example, crystallographic databases have adopted a CIF or crystallographic information file (CIF) format for all crystallographic structures.³⁷

Another example of a digital infrastructure framework that utilizes a standardized data format, is the Citrine Informatics platform detailed by O'Mara et al.³⁸ The Citration platform, which uses a physical information file (PIF) schema, is a hierarchical system that utilizes multiple relational databases and automated data extraction methods to create a platform that is both human operated and machine readable and searchable. In addition to considering data upload, storage, and retrieval issues, O'Mara and colleagues pointed out that having a hosted online infrastructure, which the user is not required to maintain and manage, encourages users to frequently and routinely access and upload to the system.³⁸

Ward et al. highlighted some of the benefits of storing materials science and engineering data, including the concept of data reuse, and their paper focused mainly on elements needed for good data stewardship in materials science and engineering.³⁹ They suggested that much of the data generated from experiments may be stored but is not accessible, yet this is often overlooked. A number of anticipated benefits of a MDI are described and include re-testing of hypotheses, verification of data results and analysis, and easy comparison with previous studies. Additional characteristics of a successful data archive include an open access policy and prevailing citations, and using discretion

when determining what data supporting a publication should be archived. Ward et al. further developed their data stewardship concept by outlining the needs for a well-developed citation or attribution system; implementation of data standards to enable discoverability, reuse, and exchange; and a potential embargo period for authors who deal with intellectual property restrictions. They also elaborated on a pathway forward for developing materials science and engineering repositories.³⁹

Although it could be said that all science and engineering disciplines face data infrastructure challenges such as the ones Dima et al., Seshadri and Sparks, and Ward et al. have pointed out, there are some unique issues to be considered that are particularly specific to materials science and engineering data. These were outlined by Kalidindi and De Graef in their 2015 paper.⁴⁰ The authors suggested that the majority of materials science and engineering data fall into three categories: synthesis and/or processing routes, hierarchical internal structure of materials and material products, and properties and performance characteristics. Of these categories, much of the data that falls in hierarchical internal structure of a materials is in the form of images (of the microstructure, for instance) generated by characterization tools, which are often very challenging to convert into digitally searchable data for storage in a digital infrastructure. This is an example of a unique data challenge in the field of materials science, because images within this discipline can be captured with a diverse set of tools (i.e., 2D surface scans or 3D tomography) and over a wide range of length scales. Depending upon which surface or internal feature is being considered, and the tool utilized, different information will be revealed within the same sample and different data will be captured.

As Kalidindi and De Graef noted, for a materials data specific infrastructure, challenges such as these must be addressed in order to facilitate mining of large data sets for extraction of useful information. In another paper, Kalidindi suggested that data be centered around process-structure-property linkages (PSP).⁴¹ Data can give way to information via trends in PSP linkages, which is then transformed into ‘knowledge’ (a more comprehensive understanding of linkages), and then finally ‘wisdom’, where PSP linkages can facilitate design and optimization of materials.⁴¹ The Kalidindi and De Graef paper also underlined the importance of standardization and automation of workflows, and cross-disciplinary collaboration, in order to build the most efficient and useful infrastructures. In agreement with elements of this paper is Agrawal and Choudhary’s perspective article, where they indicated that understanding PSP linkages and how they impact performance of the material is a key to building a materials database.⁴² The Agrawal and Choudhary paper further emphasized that international collaboration and standardized workflows are critical to materials informatics.

Two key components to a successful materials data infrastructure that are referenced repeatedly throughout the literature include: open or standardized data formats and the ability for the data to be citable. A specific example of a digital infrastructure that enables citable DOIs within its platform and uses standard formats is the European Commission Joint Research Center materials database (MatDB), which is described by Austin.⁴³ The data citation support within the MatDB is automated and has internal checks in place to ensure that the data set being published passes specific quality control checks. In this way, the MatDB is capable of preliminary verification of data prior to allowing a DOI to be issued for it. Austin describes additional systems with which MatDB is integrated, including the Gen IV Materials Handbook, which can ensure data is formatted according to technical specification.⁴³

In parallel with the aforementioned developments, experts have also convened through a range of activities and events resulting in multiple outputs that have driven the field forward. They are described in the next section.

Key Workshops and Events Addressing Materials Data Infrastructure

Appendix B provides a table with summaries of 17 recent workshops, events, and reports related to a materials data infrastructure.

In one of the earliest reports that discussed digital data and its importance to the materials science community, the 2008 National Research Council (NRC) report on Integrated Computational Materials Engineering (ICME),⁷ some key challenges were identified including data standards, intellectual property (IP) restrictions, and lack of multi-disciplinary collaboration.

The National Institute of Standards and Technology (NIST) held a workshop in 2012 that focused on data standards and the challenges associated with domain-specific data, and included 134 participating professionals representing government, academia, and industry.⁴⁴ Key themes identified specifically related to data infrastructure needs, and these included open-platform frameworks that ease the development and operation of simulation codes, as well as software that is modular, user-friendly, and applicable to broad user communities.⁴⁴ Data challenges found at the different length scales were also discussed and classified as low, medium, or high priority in terms of when they need to be solved.⁴⁴

A big data workshop convened in 2014 by the Defense Materials Manufacturing and Infrastructure (DMMI) Committee and published by the NRC emphasized the unique challenges associated with storing terabytes' worth of data.⁴⁵ Forty-five workshop participants representing academic and government laboratories, as well as industry, discussed six major themes: (1) data availability, (2) "big data" vs data, (3) quality and veracity of data and models, (4) data and metadata ontology and formats, (5) metadata and model availability, and (6) culture.

The 2015 "Building the Materials Data Infrastructure" workshop held by ASM International's Computational Materials Data (CMD) Network was aimed at summarizing and building upon prior workshops.⁴⁶ A major outcome was a prioritized, four-year timeline for future workshops that specifically addressed high-priority unmet materials community needs. Also that year, ASM International hosted the "In-Process Materials Data for Modeling" workshop, which focused on defining opportunities for companies and other organizations to collaborate in order to increase the affordability, accessibility, and availability of pedigreed data for modeling purposes, perhaps in the form of a "database collaborative."⁴⁷ Some specific materials dataset challenges included grain distribution and texture effects on mechanical properties and uncertainty in boundary conditions for modeling data. The specific data types mentioned in this report (mechanical, thermophysical, grain distribution, etc.)⁴⁷ resonate strongly with the types of data that respondents to the TMS-MRS Big Data Survey wished to see in a materials data infrastructure.⁵

In 2015, the Office of Science within DOE hosted a workshop to assess how data was acquired, analyzed, curated, stored and shared by large experimental facilities and observatories, including synchrotron and neutron light sources that produce large quantities of materials and chemical data. It was concluded that most of these facilities struggle to manage the exponentially rising rate of data they generate and that collaboration and sharing of data, tools, data analytics, and methodologies to manage large datasets is critical. However, insufficient infrastructure to facilitate such interactions is making data generated by these facilities increasingly difficult to use effectively.⁴⁸

An NSF-funded collaborative project initiated in 2015 and titled Rise of Data in Materials Research⁴⁹ focused on obtaining community input regarding the creation and use of materials cyberinfrastructures. Along with collaborations from a number of co-leaders of this project, there was an extensive two-day workshop and a subsequent town hall meeting, with recommended outputs published online.⁴⁹

The Center for Hierarchical Materials Design (CHiMaD) at Northwestern University hosted three workshops in 2016 focused on the MDI.⁵⁰ These workshops were organized along three major themes: database and discovery, building an interoperable materials data infrastructure, and materials data and analytics.

Data Infrastructure in Other Science and Engineering Communities

To provide some final background, before moving on to challenges and recommendations, examples of data infrastructures from other communities are provided in the sidebar on page 17. It should be recognized that these examples are for fairly uniform types of data. Alternatively, the MDI envisioned and discussed throughout this report entails a broader array of data types and an interactive ecosystem of multiple repositories, tools, and people.

Infrastructure Examples from Other Science and Engineering Communities

BaBar

An early example of a working digital infrastructure that dealt with vast quantities of data and the problems associated with storing, sharing, and mining them is the infrastructure built for BaBar. This collaboration effort was initiated at the SLAC National Accelerator Laboratory between hundreds of particle physicists located in several different countries (SLAC originally was an acronym for the Stanford Linear Accelerator Center).⁵¹ BaBar is a high-energy physics experiment that generates over a petabyte of data and requires an extensive data processing schema in order to function appropriately. The system relies on hardware located in several countries on an open-source framework.⁵² According to the authors, it is essential to maintain a simple overall design for data structure and workflow for an infrastructure that houses such a vast amount of data to be useful.

LSST

In the field of astrophysics, a data infrastructure project named the Large Synoptic Survey Telescope (LSST) database has been undertaken to store the hundreds of terabytes of data that telescopes receive.⁵³ The actual telescope is under construction and is anticipated to be able to map the entire sky each night for 10 years, identifying billions of stars, galaxies, and universe events. To accommodate the vast quantities of data that will be generated, a massively parallel processing system and database framework is being used for LSST. Open source codes are being implemented to encourage collaboration and enable a flexible system upon which multiple users can build.⁵⁴



Challenges

Building upon the various materials-data-related efforts that have occurred in recent years, the study team was charged with identifying what it viewed as the most notable challenges that hinder or prevent progress in developing a robust MDI. These challenges were identified not only to help guide the community in building a MDI, but to provide a foundation from which the team subsequently developed the detailed recommendations and tactics in the next section.

Challenges in developing the materials data infrastructure can be broadly considered to span technical, cultural, and policy issues. Each sub-area of materials also carries its own requirements and language, which can present hurdles to linking and unifying a MDI. To gauge the relative difficulty of each challenge identified, the team organized these challenges on axes of potential impact (abscissa) and probability for success (ordinate). The resulting graph, shown in Figure 2, should only be used as an approximate guide to the relative risks and impacts associated with overcoming various challenges in building the materials data infrastructure. Timeframes for addressing these challenges can vary widely. Although this variable was not directly addressed for each challenge, timeframes were considered in many of the tactical recommendations in the next section. Numbers corresponding to each of the 36 challenges identified are simply used for reference, and not to indicate in any way some agreed-upon priority by the study team.

In general, the four quadrants can be described as follows:

- **QUADRANT I** (*Higher Potential Impact, Higher Probability for Success*): these “no-brainers” are expected to be particularly appealing to most readers of this report since they have the highest likelihood for strongly impacting the materials data infrastructure

- **QUADRANT II** (*Lower Potential Impact, Higher Probability for Success*): this “low-hanging fruit” may not have as high an impact as quadrant I, but the likelihood of success of overcoming these barriers is expected to make them a relatively high priority in developing the materials data infrastructure
- **QUADRANT III** (*Lower Potential Impact, Lower Probability for Success*): these “tough sells” are expected to be some of the lowest priority challenges, but would nonetheless have a worthwhile impact on the MDI if successfully addressed
- **QUADRANT IV** (*Higher Potential Impact, Lower Probability for Success*): these “heavy hitters” have a lower likelihood of success, but they can have a notable impact on the MDI if challenges can be overcome.

In addition to arranging challenges on the axes and within the quadrants shown in the following graph, the study team also identified where each challenge might fall along the spectrum of materials data storage vs. sharing. This is indicated for each entry in Table I by the position of an “X” along a horizontal line in which storage is represented by the left side of the line and sharing by the right side.

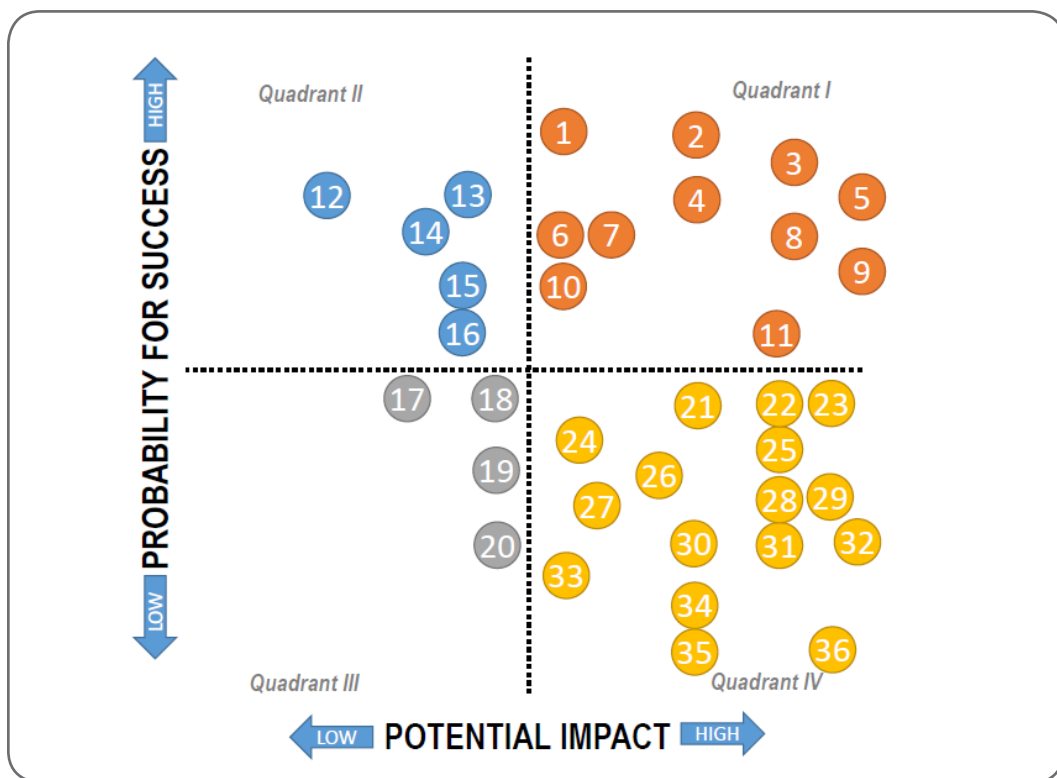


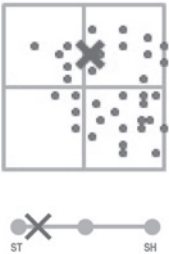



Figure 2. A plot of the relative probability for success vs. the potential impact of overcoming the 36 challenges identified by the study team.

Table I. List of challenges identified by the study team and ordered by the quadrant in which they are located in Figure 2. The numbers in the left column correspond to the challenge numbers in Figure 2.

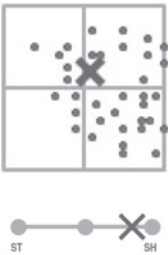


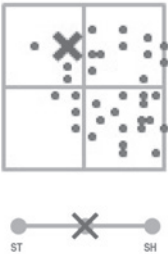
<p>All challenges were labeled in two ways: on axes that indicate probability of success and potential impact of overcoming the challenge (left) and on a line indicating the spectrum of materials data storage vs. sharing (right). In each case, an “X” specifies the relative placement of the challenge on these indicators.</p>		
QUADRANT I - (Higher Potential Impact, Higher Probability for Success)		
1	<p>No unified MSE community approach to its diverse challenges with materials data</p> <p>The materials science and engineering (MSE) community has many diverse types of data including those that describe length scale, characterization technique, materials class, and application, among others. Consequently, there is no single option to meet the broad needs of the community, and examples from other “big data” communities like astronomy do not translate uniformly. Thus, the implementation of open or common data standards, formats, and best practices for storing and sharing data remain a notable challenge.</p>	
2	<p>Mismatch between consumers and generators of specific materials data</p> <p>Given the thematic diversity and broad geographic distribution of materials research, it is difficult to identify who would benefit from access to a particular dataset. In this manner, the business case for MDI is not well defined and would benefit from articulation of costs and benefits associated with sharing different types of data.</p>	

3	<p>Lack of e-collaboration platforms</p> <p>The MSE community generates massive amounts of data and it is impossible for any single research group or organization to assemble all this information into one useful platform. Moreover, the expertise to curate and comb through a large dataset for useful information is commonly dispersed geographically, organizationally, and across disciplines. One way to address this issue is through the use of e-collaboration platforms and tools, which would significantly lower the barriers to storing and sharing data and facilitate networking and collaboration in the MSE community and across other disciplines. However, there is currently a lack of such platforms at both the organizational level (laboratory scale) as well as community level (internet scale).</p>	
4	<p>Lack of a qualified knowledge base on data management and analytics in the MSE community</p> <p>Materials researchers who also have expertise in software development, data management, and information science are needed. The MSE workforce is generally not well trained in statistics, data analytics, and database management and these are not typically emphasized in MSE curricula. Materials researchers are generally unaware of how to obtain targeted software development expertise to increase proficiency in building and using data management tools and machine learning models or algorithms that would further the knowledge base in this regard.</p>	
5	<p>Inadequate awareness of options and best practices for data storage</p> <p>Materials researchers are largely unaware of available repositories and best practices for properly storing data. Some institutions use laboratory computers, portable drives or other unreliable long-term storage solutions. Better education about suitable repositories and associated best practices are needed to ensure long-term availability and preservation of data, particularly for enabling and advancing future research and development opportunities.</p>	

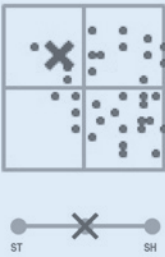
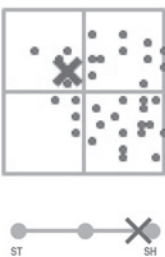
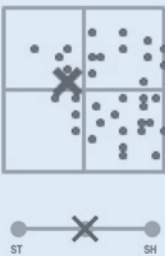
See page 21 for an explanation of the far right column of Table I.

6	<p>Limited focus on sustainable, long-term data storage strategies and support</p> <p>The materials data science and engineering community and funding agencies do not have clear strategies for realistically enabling sustained, long-term data storage. Although activities focused on long-term storage represent crucial investments for enabling a robust MDI, funding agencies instead commonly view them as outside the scope of specific research activities. Additionally, post-project commitments to keep and preserve data are rarely funded. Given the piecemeal nature of R&D funding, materials scientists and engineers are also less likely to value long-term data beyond the scope of a funded activity. Materials scientists and engineers, and funding agencies, should emphasize long-term data storage as a central objective in funded R&D activities.</p>	
7	<p>Insufficient participation of the computer science community in the MDI</p> <p>The MSE community is poorly integrated with the computer science community. This presents a real challenge to both communities as oftentimes what is needed by MSE does not advance the state-of-the-art in computer science. However, interfacing with experts in the computer science community could significantly accelerate the development of a MDI by providing critical recommendations and guidance that would help improve the state of data storage and sharing.</p>	
8	<p>Many required elements and solution pathways for the MDI are not defined in enough detail</p> <p>Though many groups have discussed challenges to creating the MDI, key elements such as providers, solutions, and required specialties are insufficiently defined with enough detail to enable implementation solutions to be developed. This may include specific tasks and pieces that are critical but are not well defined or universally agreed upon within the MDI community.</p>	
9	<p>Insufficient mechanisms for crediting data contributors</p> <p>Many data management policies are not developed with the objective of crediting scientists and engineers who share their datasets. In the absence of proper attribution, data contributors are less likely to share or preserve their datasets, particularly given the resources and investment required to generate data. Proper citation of data is essential for incentivizing data sharing and can deliver other benefits including: transparency of the data generation process, reproducibility and validity of experimental results to grow the MSE knowledgebase, increased data discovery, and prevention of redundancies in data generation.</p>	

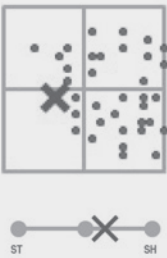
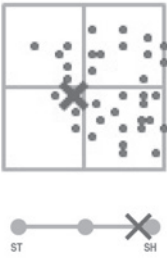
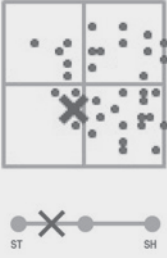
See page 21 for an explanation of the far right column of Table I.

10	<p>A pathway for industrial participation from large scale manufacturing in the MDI is unclear</p> <p>Data and knowledge from large-scale manufacturing does not easily feed back into the materials data infrastructure. There is a lack of both repositories and incentives for manufacturers that invest heavily in the underlying research and technology that generate data to share. This creates an impediment to sharing with the broader community even though it could greatly enhance the ability for materials innovations, discovery, and development.</p>	
11	<p>Scarcity of data sharing incentives</p> <p>Incentives to share data are scarce and often poorly communicated. Materials data producers can be hesitant to share data due to reasons such as fear of users mishandling data, a desire to own and control data, insufficient recognition of authorship, general risk aversion, and lack of financial compensation for the time and effort invested in data generation. Clearly articulated incentives for the MSE community can help bridge the disconnect between the perceived downsides and the potential benefits of sharing data within a MDI.</p>	
QUADRANT II (Lower Potential Impact, Higher Probability for Success)		
12	<p>User interfaces for uploading and downloading data can be challenging to design</p> <p>A user-friendly interface is important to facilitate data storage and sharing. However, most materials researchers are not well equipped to design the needed user interfaces.</p>	
13	<p>Success stories and proofs of concept are needed to demonstrate the value of data-driven materials science and engineering</p> <p>At present, with minimal elements of the infrastructure in place, it remains difficult to demonstrate the true value and capacity for innovation through a robust materials data infrastructure. Nonetheless, early proofs of concept will help not only advance our technical understanding of materials data issues, but can also be used as leverage to grow community support. In a similar vein, examples of failures should also be shared within the community to help highlight the gaps and deficiencies in the current materials data infrastructure.</p>	

See page 21 for an explanation of the far right column of Table I.

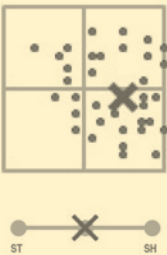
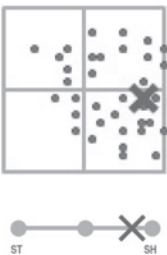
14	<p>Need for federated approaches to data sharing and storage</p> <p>The complexity and diversity of materials data and data sources drives the need for a federated data architecture. Unlike some of the health sciences that are funded almost exclusively by the National Institutes of Health (NIH), multiple agencies fund materials science research and development efforts. The lack of a sole funding entity thus poses an additional challenge to the federated approach to data sharing, which would benefit from a large national investment.</p>	
15	<p>Resources among data infrastructure providers are poorly integrated</p> <p>Due in part to the large diversity and heterogeneity of materials data, data infrastructure providers are not well integrated. There are currently very limited agreed-upon tools and methodologies for assessing materials data, which inhibits sharing and use.</p>	
16	<p>Insufficient options for long-term storage of “intermediate” data</p> <p>The MSE community would benefit from data platforms to help researchers upload, store, and access “intermediate” data for later use. For example, a thesis document commonly contains a set of text, tables, and figures but it is only a surface capture of the output. There may be a rich set of underlying, “intermediate” data that were not suitable for publication in the thesis. These data are often stored locally with minimal descriptions, making them difficult to be reused by other researchers, especially in the long term. Similarly, some datasets may not get published in a journal because they lack the scientific rigor for consideration. Currently, there are few viable options for storing and accessing these types of data.</p>	

See page 21 for an explanation of the far right column of Table I.

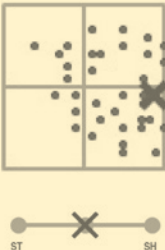
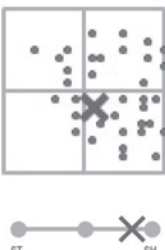
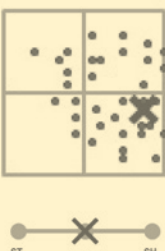
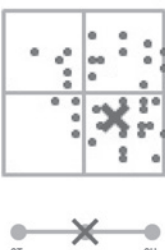
QUADRANT III (Lower Potential Impact, Lower Probability for Success)		
17	<p>Lack of robust APIs of connected systems and instrumentation</p> <p>The community would benefit from allowing application program interfaces (APIs) to access databases in an automated manner rather than via web interfaces. At present, users are mostly confined to the time-consuming task of pulling data from various resources and manually updating, particularly if any of the underlying databases are changed. Thus, APIs can play an extremely useful role in helping connect databases and instrumentation as well as reducing the time investment required by the data user or provider. This challenge is related to challenge #28.</p>	
18	<p>Complexity and disparate nature of materials data</p> <p>The complexity of materials data creates challenges for sharing among scientists and engineers as they can be represented in myriad ways across a range of time and length scales, spatial dimensions, compositions, instruments, models, end-use applications, and processing approaches, to name just a few. Furthermore, there is an ever-increasing amount of data captured by relatively newer techniques such as atom probe tomography, but they remain largely inaccessible since most of the underlying data is not typically published along with the final results; nor are they available to others outside of the research group or institution that performed the work. Across the MSE community, there are many examples where materials data are highly diverse and disparate. Facilitating access and pooling these disconnected data sources could help highlight gaps and eliminate redundancies in research and development efforts.</p>	
19	<p>Inadequate understanding of cost associated with materials data in the short- and long-term</p> <p>Generating or producing data can be costly, which directly impacts the time, financial support, and other resources that might be available to facilitate data storage and sharing. Furthermore, the growth in the amount of data along with development and maintenance of the chosen repository for their storage are indefinite. This leads to many difficulties in ensuring continued availability and usefulness amid issues such as personnel turnover and equipment or tool upgrades. Thus, quantifying the cost to store and sustain materials data over their useful lifetimes is difficult.</p>	

See page 21 for an explanation of the far right column of Table I.

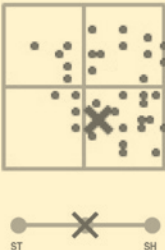
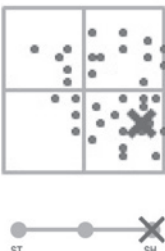

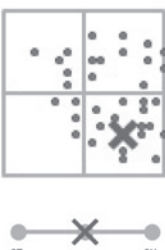
20	<p>Constraints of government technology export regulations</p> <p>To protect the security and economic interests of the United States, export control regulations limit the movement of technologies, information, and commodities to foreign entities. These regulations present unique challenges for all types of organizations operating in the United States including universities, industry, and federal laboratories. For example, cloud storage solutions—which permit the virtualization of resources and enable collaboration among data sharers—can also carry significant risk of violating export control regulations, and may also implicate the cloud storage provider facilitating the transmission of data. The penalties for non-compliance can invoke risk-averse behaviors among data contributors, thereby limiting the amount of data stored or shared. Ensuring compliance with federal law requires mechanisms that reduce these fears and promote a culture of sharing, such as common definitions or standardized interpretations of export control regulations.</p>	
----	--	--

QUADRANT IV (Higher Potential Impact, Lower Probability for Success)		
21	<p>Retraining the existing workforce</p> <p>In developing the MDI, the existing workforce will need to be educated and retrained on appropriate standards, data management practices, and useful tools as they develop. However, mechanisms for educating and retraining the workforce are currently limited, and there are minimal incentives to adequately address this issue.</p>	
22	<p>Limited data repository usage and availability of tools</p> <p>At present, there are a limited number of repositories, tools, and e-collaboration platforms that enable the storage, sharing, and reuse of data. Without a critical mass of others uploading data to existing repositories, there is little incentive to contribute one's own data or to develop tools that might be useful for a given repository.</p>	


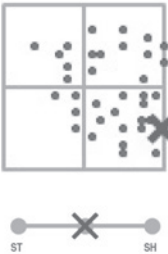

See page 21 for an explanation of the far right column of Table I.

23	<p>Lack of a clear, unified vision of how the MDI will benefit the community</p> <p>The MDI requires a clear vision to communicate how it will deliver benefits to the community through advancements in technology and ultimately aid critical areas such as health, national security, and infrastructure. In this regard, it is important to think about the vision holistically. That is, the MDI is an ecosystem and more than just an assembly of hardware and software.</p>	
24	<p>Insufficient consensus on uncertainty quantification (UQ) methods in the MSE community</p> <p>A key aspect to sharing useful data with the community is being able to accurately quantify its uncertainty. Yet, there is no well-defined set of uncertainty metrics to enable validation of experimental and computational data. This also hinders collaboration with non-materials scientists or engineers who are less familiar with the methods used to obtain relevant materials data.</p>	
25	<p>Lack of developed, agreed-upon ontologies for materials domains</p> <p>Other scientific disciplines that have found success in developing a data infrastructure have also developed ontologies as part of the process. (In computer science and information science, an ontology refers to a formal naming and definition of the types, properties, and interrelationships of the entities that exist for a particular domain.) Ontologies in the materials data domain would help the MSE community work toward standard representations of data and other elements within the MDI. There are currently no agreed-upon ontologies that can be applied even in sub-domains of the materials community. Moreover, there is not yet a clear pathway for how these would be developed.</p>	
26	<p>Underdeveloped data management approaches for MSE knowledge</p> <p>The MSE community outputs many types of data, but with insufficient structure for managing that data, it is difficult to identify important gaps, trends, or other statistically significant phenomena. Additionally, current data management approaches within the MSE community do not encourage machine-ready, machine-readable, or computable data, making it a challenge to employ that data when conducting computational or machine-driven experimentation. There is a need to establish methodologies that better structure data for storage and enable machine readability, which would help improve the usability and impact of the MDI.</p>	

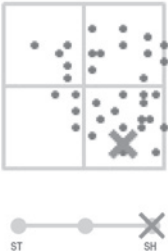

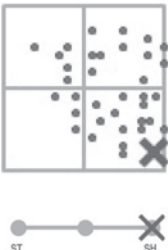
See page 21 for an explanation of the far right column of Table I.

27	<p>Need for standardized components and documented workflows to enable data extraction and reuse</p> <p>Workflows that generate and handle data define the necessary software or tools, inputs, and parameters needed to efficiently and effectively manage and reuse such data. Yet, best practices for documenting workflows are not integrated into conventional R&D approaches. By properly documenting workflows and integrating them with standardized MDI components, a user has a much better chance of locating data that can be used in their work.</p>	
28	<p>Poor interconnectivity of data platforms, which inhibits creation of materials data ecosystems</p> <p>To date, most data platforms in materials are not designed for interoperability. Much of this has to do with the variety of domains covered by materials science and engineering (e.g., timescale, length scale, composition, processing parameters). The lack of interoperability begins at the source of the data, with limited common or open standards for data output from scientific instruments, and extends outward. The community will continue to be challenged in creating the underlying ecosystem for the MDI without a set of standards or best practices that promote increased interconnectivity of materials data platforms. This challenge is related to challenge #17.</p>	
29	<p>Lack of mechanisms or use-metrics to indicate when old data is updated</p> <p>As materials-related projects progress, it is a given that some data will eventually be updated due to availability of new tools, approaches, or new knowledge. However, there are no commonly accepted best practices for sharing these updates with people who may have used this data to inform their own work. The MDI will benefit from a mechanism where users can be informed of updates, as appropriate.</p>	
30	<p>Lack of existing long-term sustainable business models for individual elements of the MDI</p> <p>Some long-term (i.e., >10 years) sustainable business models for individual elements of the MDI exist, e.g., for-profit modeling tools from Materials Design, ThermoCalc, the ICSD database, alloy phase diagram books, and others. However, for most data, such business models are either nonexistent or not well understood.</p>	

See page 21 for an explanation of the far right column of Table I.

31	<p>Inadequate IT security and outdated operating systems</p> <p>Despite their best efforts, many materials data producers often lack the resources to ensure up-to-date operating systems and adequate IT security to protect the data they generate. These vulnerabilities can impede progress in developing the MDI since they could lead to a security risk for the data producers and others. Best practices should be developed to help mitigate the risk associated with these potential issues.</p>	 <p>The scatter plot shows data points distributed across four quadrants, with a concentration in the top-right. An 'X' marks a point in the bottom-right quadrant. Below the plot, a horizontal timeline with dots at 'ST' and 'SH' has an 'X' between them.</p>
32	<p>Lack of funding and career opportunities for materials data management</p> <p>As the MDI develops, it will require funding and personnel dedicated to data management (as opposed to just materials R&D). At present, there are very limited financial resources devoted to storing and curating materials data at the individual or research group level. Some academic institutions are beginning to provide solutions through their libraries, but with limited utility for the diversity and heterogeneity of materials data. Ultimately, more resources are required to ensure that materials data producers are properly managing the data they generate and thus able to fully participate in the MDI.</p>	 <p>The scatter plot shows data points distributed across four quadrants, with a concentration in the top-right. An 'X' marks a point in the bottom-right quadrant. Below the plot, a horizontal timeline with dots at 'ST' and 'SH' has an 'X' between them.</p>
33	<p>Ambiguity of federal agency data policies</p> <p>Federal agency requirements for data storage and management are often unclear or too generalized for researchers to confidently follow. Data management policies may vary between funding agencies, and many do not clearly specify how compliance will be monitored, which can remove the motivation for managing data throughout its lifecycle. Some materials scientists and engineers lack proficiency in data, computer, or information science, further complicating their ability to comply and follow through with data management plans. Researchers may find it difficult to justify the time needed to manage data, especially in the absence of mandates or incentives.</p>	 <p>The scatter plot shows data points distributed across four quadrants, with a concentration in the top-right. An 'X' marks a point in the bottom-right quadrant. Below the plot, a horizontal timeline with dots at 'ST' and 'SH' has an 'X' between them.</p>

See page 21 for an explanation of the far right column of Table I.

34	<p>Lack of well-defined data sharing norms among publishers and funding agencies</p> <p>Data sharing policies vary widely across funding agencies, and there are no universal standards or norms utilized by agencies and publishers alike. For many professionals, sharing data between trusted colleagues or peers is a common practice as some consider data to have no value if it cannot be reproduced or validated by others. Yet, data sharing norms depend on the type of discipline and research organization. There is a need for funding agencies and publishers to develop policies that facilitate sharing of multiple data types and set clear expectations for data accessibility, particularly within the MSE community.</p>	
35	<p>Inadequate career incentives to share data</p> <p>Whether a person is working in academia, industry, or government, there is typically minimal personal incentive to store and organize materials data. With no value placed on data storage and sharing, it will be extremely difficult to build a robust MDI that includes the necessary participation from each of these sectors. This challenge is related to #11, but is more focused on individuals.</p>	
36	<p>A data sharing culture is hindered by issues such as intellectual property and privacy</p> <p>Concerns including intellectual property and privacy can hinder opportunities for data sharing in the MSE community. In some cases, it is such uncertainty about how such IP and privacy laws would be executed that makes materials scientists and engineers averse to sharing. More guidance is needed in how to navigate potential issues involving copyright law, contract law, and privacy as they pertain to a person's work. This would help mitigate the risks associated with storing, sharing, and reusing data, and empower people to actively participate in the MDI.</p>	

See page 21 for an explanation of the far right column of Table I.



Recommendations

Using the myriad challenges identified in the previous section as input, multiple recommendations were identified to help build a sustainable materials data infrastructure (MDI). Particular focus has been given to issues associated with materials data storage and sharing. Table II provides a summary of the eight overarching recommendations, and includes multiple corresponding tactics for addressing each of these recommendations. Following Table II is a discussion for each of the high-level recommendations and accompanying tactics or actions, with details including tactical strategies, timeframes, recommended implementers, estimated costs, and possible sources of financial support to accomplish these activities.

As mentioned in the Preface section, it is hoped that many readers of this report will identify recommendations, suggested activities, and tactical details for which they have interest and relevant expertise. The goal is to enable the community to implement these recommendations and make significant progress in the next three years toward building and participating in a robust MDI.

Additionally, once the reader has decided where to engage, they need not be limited to the information provided in the recommendations and tactics here, but can also do a deeper dive into additional related information in the other resources provided and referenced in this report. Along these lines, the recommendations and tactical details provided in this section should not be considered all-inclusive. Scientists, engineers, leaders, and policy makers who read this report should feel challenged to not only undertake some of the actions suggested here, but also consider developing other recommendations and actions toward building a robust materials data infrastructure that addresses the challenges presented in the previous section.

Table II: Priority Recommendations**Recommendation 1: Strengthen the MDI core in repository, registry, and tool development**

- Tactic #1: Develop and deploy robust repositories
- Tactic #2: Develop and deploy registries for MDI repositories and tools
- Tactic #3: Develop analytical and visualization tools that enhance the speed and capabilities of materials data use and analyses
- Tactic #4: Launch and sustain e-collaboration platforms
- Tactic #5: Develop automated data/metadata capture tools for scientific instruments

Recommendation 2: Sustain and grow MDI-dedicated funding programs

- Tactic #1: Support the MDI community by leveraging and coordinating current federal programs
- Tactic #2: Develop and launch new funding programs

Recommendation 3: Create, execute, and monitor incentive mechanisms

- Tactic #1: Establish incentive mechanisms for materials data sharing
- Tactic #2: Execute and monitor the incentive programs to energize and sustain MSE community involvement

Recommendation 4: Develop demonstration projects and cross-disciplinary community efforts that enhance and accelerate adoption of the MDI

- Tactic #1: Establish materials-data-driven design projects that enable property prediction, to enhance MDI adoption
- Tactic #2: Establish materials-data-driven projects for manufacturing, to enhance MDI adoption
- Tactic #3: Launch targeted community efforts to help achieve MDI critical mass in the MSE community
- Tactic #4: Fill gaps in the MDI materials data domain

Recommendation 5: Establish a MDI ecosystem and business cases

- Tactic #1: Develop a reference architecture concept for the MDI
- Tactic #2: Develop and demonstrate business cases for data storage and sharing
- Tactic #3: Explore concept of a materials data “app store”

Recommendation 6: Develop and invest in education and training programs for the MDI workforce (providers and users)

- Tactic #1: Integrate MDI into existing MSE curricula and build new cross-disciplinary curricula
- Tactic #2: Conduct outreach and training programs for professionals

Recommendation 7: Create MDI consortia and working groups

- Tactic #1: Create a long-term coordinating and advisory body comprising relevant MDI stakeholders
- Tactic #2: Create Community of Practice (CoP) groups around the MDI

Recommendation 8: Define and establish clear policies and guidelines associated with the MDI

- Tactic #1: Establish an interagency (federal) council to foster consistent data preservation and sharing policies
- Tactic #2: Increase emphasis on data management plans for funding support considerations

Recommendation 1: Strengthen the MDI core in repository, registry, and tool development

Platforms and tools for materials data management are essential for building a robust MDI that enables data-storage and data-sharing in support of materials-related research and development activities and innovations. Dedicated, sustained, and integrated efforts are needed to build and enhance such platforms and tools. The tactics below are focused on key areas where significant progress is needed and can be achieved within the next three years. The recommended timeframe for implementation would roughly be the same for all of the tactics described below; that is, financial support should be allocated and secured within the next year, and it may take up to three years for such efforts to come to fruition in terms of deployment and adoption within the community.

Tactic #1: Develop and deploy robust repositories

This tactic is intimately related to a number of the other recommendations and suggested actions provided below, and similarly, it helps to provide an initial foundation for various issues underlying the creation of a robust MDI. Although it will be up to the individual repository developers to determine the detailed hardware, software, frameworks, governance, and functionality of the individual repositories, as guided by their individual research and data needs, following are some overarching guidelines for such development and implementation. Additionally, some examples of current materials data repositories are also provided.

An effective data repository must be capable of performing tasks that integrate seamlessly together, including capture, storage, curation, retrieval, and sharing of the materials data. Additionally, the MDI will require a federated approach of many such individual repositories that are interoperable. Repositories have to account for data and metadata captured both from laboratory-notebook types of structured and unstructured small data sets, as well as massive materials data sets, using well-written scripts and interoperable interfaces. Uncertainty quantification (UQ) must be addressed and tracked during data capture and subsequent data handling processes, for both experimental and modeling data. Well-designed data schema should also allow for flexible repositories that can evolve to accommodate new types of information.

It is essential, especially for data ingestion, discovery, retrieval, and sharing, to develop user friendly repository interfaces that feature thoughtfully designed application program interfaces (APIs) and graphical user interfaces (GUIs). This includes development and leveraging of API protocols for data and metadata interoperability across both new and existing resources, to enable a federation of repositories and repository tools that increase usability, accessibility, and sharing. A measure of success will be the adoption by the community of multiple robust materials data repositories that are interlinked and user friendly.

Federal agencies such as the National Institute of Standards and Technology (NIST) should continue to play a leading role in deploying materials data repositories, in addition to universities and technology companies. Likewise, it is suggested that federal agencies including NIST, the National Science Foundation (NSF), the Department of Energy (DOE), and the Department of Defense (DoD) provide financial support for enhancement and development of such repositories. Implementers should include university researchers, experts at software companies, engineers at materials-data-generating corporations, researchers at federal laboratories, and other users of materials data.

Examples of some existing materials data repositories are provided in the sidebar below, and although such efforts represent a strong step forward in building a robust MDI, most of them are still in relatively early stages of development and deployment. Additionally, they represent only a fraction of the materials data produced by the community. These and other existing materials data repositories can serve as models to provide a valuable foundation and lessons learned for those who are developing repositories, and those developing the frameworks and tools to enable interoperability among repositories within the MDI.

Examples of Data Repositories for the Materials Community

NIST-CHiMaD Repository

The NIST-CHiMaD Proto Data Repository (<https://phasedata.nist.gov>) contains a variety of (structural) phase-based experimental and computational data to support the CALPHAD (CALculation of PHase Diagrams) assessment process. These phase-based data include, but are not limited to, thermodynamic quantities, phase equilibria data, and diffusion quantities. The data are generally only semi-structured and contain a variety of 1-, 2-, and 3-dimensional data. The repository consists of three parts. First, experimental and computational alloy phase-based data can be searched for and added to the repository using the Materials Data Curation System (MDCS).⁵⁵ By using community-developed data schemas, the MDCS platform allows users to enter data in multiple formats that can then be combined to search, reuse, and transform the data. The data can be entered either using a web interface dynamically generated based on the data schema selected or using a representational state transfer (REST)-API. Currently this part of the repository emphasizes experimental data associated with the development of a Cobalt-based CALPHAD thermodynamic database, which is being developed through CHiMaD (the Center for Hierarchical Materials Design), as well as a variety of diffusion data. This includes data associated with self-diffusion, impurity diffusion, and a variety of alloy diffusion couples. In conjunction with the diffusion data, a graphical interface is provided to compare experimental and computational self-diffusion coefficients.

Secondly, after selecting the element of interest, the user can select the experimental and computational data of interest and the Arrhenius ($\log D$ vs $1/T$) plot is automatically generated. The third part of the repository includes published experimental unary, binary, and ternary thermochemical and thermophysical property data that are searchable through the NIST Thermodynamics Research Center.⁵⁶

Materials Commons

The Materials Commons⁵⁷ is an e-collaboration platform and information repository for use by the structural materials community, and has been developed by the Predictive Integrated Structural Materials Science (PRISMS) Center at the University of Michigan.⁵⁸ In collaborative projects on the Materials Commons, researchers upload the results of experiments and computations as they are performed, along with the provenance information that describes how the experimental and computational processes were performed and what data they produced. Using the Materials Commons enables researchers to safely store data, share it among collaborators, programmatically analyze the results, and then make the completed datasets public, with provenance so that others may understand, use, and cite the data. The Materials Commons follows the normal scientific workflow researchers use in developing their data to seamlessly capture the provenance of the information.

In a Materials Commons project, data is stored as original data files and also parsed to store structured data in a NoSQL database that can capture materials data and provenance in a way meant to accurately represent scientific knowledge, integrate experimental and computational data, and enable fitting constitutive and process models. Within the Materials Commons data model, processes link changes in sample attributes and describe processing-structure relationships, while the set of sample attributes associated with a particular experimental or computational sample describe structure-property relationships. These links can be made across Materials Commons projects to describe integrated workflows and hierarchical models. Future development is planned to enable “cloning” a published dataset so that a new project can be constructed to reuse data from an existing project while maintaining all provenance.

Citration

Citration⁵⁹ is a cloud platform, developed by Silicon Valley-based Citrine Informatics, which closely couples large-scale materials data and artificial intelligence (AI) to rapidly extract insight across previously isolated datasets. The platform exists in two versions: Open Citration, which is free for academic and government researchers and contains one of the world's largest collections of open materials data, and Enterprise Citration, which is a for-pay software product used by companies. Underpinning Citration is a unique business model—nonprofit researchers may use the advanced data infrastructure and AI tools available on the Open platform without paying for access or storage, and companies on Enterprise Citration are able to mine across both platforms while keeping their proprietary data completely private.

Citration enables materials scientists to gain leverage in their work by using AI to harvest insights from a large number of experiments and simulations across laboratories and institutions, representing potentially millions of human-hours of cumulative research effort. Citration benefits from network effects—the more people use the platform and store data there, the smarter and more effective it becomes for all users. Citrine Informatics believes that these leverage and network effects will drive use of the platform by providing incentives for data contribution that strengthen over time, and lead to a critical mass of users across the materials community.

Tactic #2: Develop and deploy registries for MDI repositories and tools

To enhance the broad use of a federation of materials data repositories and the sharing of materials data, it is important to develop and maintain registries of data repositories and tools to facilitate the discovery of data resources. Examples of registries include the NIST Materials Resource Registry (NMRR)⁶⁰ and the multidisciplinary Registry of Research Data Repositories.⁶¹ Some key metrics to gauge the utility and adoption of registries are the number of repositories and tools listed as well as the sustained availability of these tools and repositories. It is suggested that NIST continue to take the lead on this tactic, in coordination with other federal agencies, with support as needed from professional societies. Organizations such as NIST, NSF, DOE, and DoD could provide financial support to develop and sustain such materials data registries. In order to avoid duplication, it is expected that the number of such registries needed might be very small, and such efforts should be tightly coordinated with one another. In fact, NIST's vision is that separate, but federated, instances of the NMRR exist across the community. Total investment needed in this arena is estimated to be on the order of \$1M per year, to develop and sustain such registries over time.

Tactic #3: Develop analytical and visualization tools that enhance the speed and capabilities of materials data use and analyses

It is important to develop new tools, as well as to leverage existing ones, that encompass machine learning, data mining, and data visualization. These analytical tools should be communally available and capable of leveraging large datasets, as described in the NSF Task Force Report on Data and Visualization.⁶² Associated materials data analytics codes ideally should also be open access and built with open source software where feasible (e.g., Python's open machine learning suite: "scikit-learn").⁶³ Tools that automate the capture of metadata for material science datasets in machine-readable formats from computation are also essential. One indicator that will shed light on the effectiveness of such tools will be the amount of data reuse and citations that result from employing such tools; therefore, wherever possible these metrics should be tracked for usage of these types of tools.

The National Science Foundation, DOE, and DoD could provide fiscal support for such tool development and adoption, which may require an initial investment on the order of \$1M, and subsequent support to ensure sustainability. Developers of these tools include the same groups that will implement repository development: university researchers, experts at software companies, engineers at materials data generating corporations, researchers at federal laboratories, and users of materials data in other organizations. Development of these tools is encouraged to occur in conjunction with development of repositories.

Tactic #4: Launch and sustain e-collaboration platforms

Launching and sustaining e-collaboration platforms (also sometimes called e-science gateways) will help build a vibrant community of practitioners and users of a MDI. Robust e-collaboration platforms are essential to lowering the activation barriers to data capture, storage, and sharing. They will not only enhance data sharing, but will also enhance and accelerate the development of the required MDI repositories, registries, and tools, as well as their linkage within a federated system. Such e-collaboration platforms will especially support cross-disciplinary interaction. As envisioned here, one difference between simple data sharing and e-collaboration lies in the degrees of separation from the collaborator to the person collecting the data (the "owner"), as well as in the broader nature of the e-collaboration. For instance, e-collaboration platforms can provide researchers and product

developers the ability to work with others' datasets without necessarily directly collaborating with the data owner. Consequently, these novel platforms will dramatically increase the impact and utility of the aggregated datasets.

Similarly, although data sharing commonly has some component of collaboration, there are also ways to collaborate without sharing data. For example, researchers may collaborate on tool development for the MDI. E-collaboration platforms could also involve extracting information from workflows for use by the community. E-collaboration also can provide mechanisms for insights from the community on how best to share data in different cases, based on feedback of what platforms or tools have been of greatest use.

Although e-collaboration platforms could take any number of different forms with different sets of functionalities, some examples and further guidance are provided here. At present, many examples of e-collaboration platforms exist for specific, limited tasks. They include Google Docs and ShareLaTeX for writing, Mendeley and ResearchGate for research documents, and GitHub and Sourceforge for software development. In the materials community, example platforms that offer e-collaboration functionalities include the Integrated Collaborative Environment,⁶⁴ Materials Commons,⁵⁷ nanoHUB,⁶⁵ Materials Innovation Network (MATIN),⁶⁶ and Timely and Trusted Curation/Coordination (T2C2).⁶⁷ Some measure of success of such platforms would be not only the number of participants, but the number of interactions, as well as any tangible outputs or products that result from these e-collaborations. Financial support for such platforms and activities is needed from government organizations, which could include NSF, DOE, and DoD, and it is estimated that these may entail long term investments on the order of \$3–5M, particularly for continued support to ensure sustainability. Implementers of these platforms would include universities, federal laboratories, and industry.

Tactic #5: Develop automated data/metadata capture tools for scientific instruments

It is imperative to develop new data capture tools for scientific instruments used in the experimental characterization and testing of materials. However, oftentimes these instruments output data and metadata in closed or proprietary formats that hamper ingestion into a data repository, reuse, and subsequent sharing. Yet, these instruments develop huge amounts of critical experimental data in the materials community, and manual intervention in the capture and analysis of this data in coordination with the repositories and other tools in the MDI is woefully inefficient. Automated data and metadata workflow among these instruments, repositories, and tools is critical for fully utilizing the MDI. Just a few examples of the characterization and testing domains in which these scientific instruments are employed include electron microscopy (e.g., transmission electron microscopy, scanning electron microscopy, electron backscatter diffraction, etc.), x-ray diffraction, x-ray tomography, x-ray photoelectron spectroscopy (XPS), mechanical testing, 3D atom probe analyses, and serial sectioning, among others.

Key elements needed to support data capture tools will include scripts for data extraction and open data standards associated with the scientific instrumentation. Developers of these tools should include the scientific instrument manufacturer communities as well as university researchers, experts at software companies, and researchers at federal laboratories. NIST, professional societies, university researchers, researchers at federal laboratories, and instrument users within industry could

all play a key role in developing data standards associated with such scientific instruments. NIST has previously facilitated a similar activity for the manufacturing community by establishing the MTConnect Institute in collaboration with equipment manufacturers and their industrial customers. The MTConnect standard enables uniform data flows from disparate industrial equipment, allowing manufacturers to easily collect and analyze their shop floor data. Organizations including NSF, NIST, DOE, DoD, and the instrument manufacturers could provide fiscal support for such tool development and adoption, which may require an investment on the order of \$1M–\$2M. The incentive for the investment in time and money by the instrument manufacturers would be the greater usability and demand for their instruments, once users can more effectively integrate them within the materials data infrastructure.

Recommendation 2: Sustain and grow MDI-dedicated funding programs

Funding opportunities are essential to new, breakthrough technological advances. Federal agency support would help the MSE community build the MDI by developing and implementing enabling platforms, tools, and resources described elsewhere in this report. To stimulate the initiation of small projects and community-wide efforts, agencies could make significant contributions in a relatively short period of time (i.e., within 12–18 months). Although suggestions for sources and degrees of support for some specific activities are provided in the tactics in Recommendation #1, as well as a number of the other recommendations below, some suggested possible methodologies or pathways for organizing support are provided in the following two tactics.

Tactic #1: Support the MDI community by leveraging and coordinating current federal programs

Given the recent efforts of federal agencies to support open data infrastructure components for data-driven approaches, some MDI-supporting programs already exist. Yet they often remain scattered, with limited ability for individual efforts to integrate into a federated MDI that leverages the full sum of data, repositories, registries, tools, and capabilities. Although the Materials Genome Initiative (MGI) has provided great progress in this regard,^{1,2} the MGI efforts include a broader scope than a specific, dedicated focus on the MDI.

A primary thrust of this tactic is to increase and coordinate MDI development through existing federal programs, and to also underscore the importance of those that have a broader charter than just materials data, such as NSF's Data Infrastructure Building Blocks (DIBBs) or Software Infrastructure for Sustained Innovation (SI2) programs. Such programs lead to the development of important research-enabling tools that are necessary to build, support, and sustain the MDI. By funding activities to help contribute to a more robust MDI, agencies are helping facilitate the storage and sharing of data produced by principal investigators and thus prolonging the useful life of data beyond the timescale of a single research project.

Better coordination is also needed between federal agencies to coordinate existing materials-data-related programs. Implementers of this tactic would be program officers, division directors, and directorate leadership at various government agencies, who are encouraged to work together across disciplines, organizational structures, and agencies. This effort could be initiated by a small group

of champions from within these organizations working together to take the lead in forming an interagency funding coordination group for the MDI. This could be the same or possibly a different group than the broad coordinating and advisory body described in Recommendation #7, Tactic #1. But these two groups should work in coordination with one another, perhaps having some common members or liaisons between them.

Tactic #2: Develop and launch new funding programs

Federal agencies are encouraged to expand the role of MDI investments in their existing and new funding initiatives. Funding opportunity announcements (FOAs) could be issued with a special focus on the MDI. One example is FOAs for MDI seed collaborative demonstrations, which could include smaller funding opportunities that would help build the MDI. One suggestion would be a project or set of projects on automatically extracting materials information from journals and documents. More significant funding could be invested in development of specific repositories, registries, tools, and e-collaboration platforms for the MDI. Likewise, new funding could be provided for coordinating community-wide activities such as development of standard, sustainable access to materials science datasets. Additionally, new MDI programs that promote public-private partnerships could be funded.

Concurrent with Tactic #1, new multidisciplinary programs can be launched, perhaps through the integration of existing programs, to help promote cross-disciplinary partnerships. For example, federal agencies could launch new MDI programs across multiple divisions or directorates to encourage multidisciplinary work that is needed to build a highly effective MDI. Such large interdisciplinary programs could be undertaken by any of the federal agencies and might include: (i) building robust e-collaboration platforms through the support of multiple cross-disciplinary research groups, perhaps working together in a large team effort (see Recommendation #1, Tactic #4); (ii) supporting large multi-disciplinary teams to develop repository linkages focused on application program interfaces and graphical user interfaces for interoperability (see Recommendation #1, Tactic #1); (iii) establishing large, collaborative materials data-driven projects (involving academia and industry) for manufacturing innovations, with emphasis on accessing, creating, and employing materials processing and manufacturing datasets to demonstrate data-driven process modeling in the manufacturing enterprise (see Recommendation #1, Tactic #2).

The new programs suggested under this tactic, when taken together, would represent a large investment in the MDI, which could total on the order of \$50M or more, and would involve multiple government agencies including NSF, NIST, DOE, and DoD. To a lesser extent, private foundations might also be a source of at least some support.

Recommendation 3: Create, execute, and monitor incentive mechanisms

For a robust and impactful MDI to be created and broadly adopted, strong incentives are needed that appeal to the entire materials data community of users and providers. A major incentive that inherently exists for such a MDI is the promise for an accelerated pace of breakthroughs and materials-related innovations, technologies, and products. As an example, predictive computational

modeling and simulation is enabled by the sharing of data between experimentalists and modelers, which allows for rapid validation of computational models and results, and can lead to important innovations. However, such advantages need to be more widely understood and appreciated by the community.

Likewise, more concrete incentives that directly impact individual stakeholders within their organizations need to be created and promoted. Such incentive mechanisms can provide a major catalyst for attracting more skilled people into this endeavor and promote more vibrant MDI-related collaborations within the community.

Tactic #1: Establish incentive mechanisms for materials data sharing

To achieve the goal of this tactic, many stakeholders, professional societies, and publishers in particular, should help recognize groups or individuals who share materials data. (For more information on the role of publishers in the MDI, see the sidebar on Publishing Issues on page 44.) More specifically, one mechanism whereby this can be done is establishing common practices in citations to data or data-related articles (e.g., Thomson Reuters, Data Citation Index). The Force11 Group¹⁵ has developed a Joint Declaration of Data Citation Principles that establishes a foundation for such practices. Since citations are often a key metric for tenure consideration at universities, as well as for promotion within government laboratories and some companies, this could be a key incentive. Publishers, possibly working with professional societies, could implement such practices for materials science and engineering.

Another mechanism for incentivizing the community to share data is to solicit and increase the publication of MDI success stories in technical journals or professional society communications. These articles would include a description of how data sharing contributed to a specific technological breakthrough or innovation. Importantly, this may also help reduce the perceived risk of legal issues that commonly hinder materials scientists and engineers from sharing data (see sidebar on Legal Issues on page 45 for additional detail).

Awards and other forms of recognition can be a strong incentive for data sharing. To establish awards, prizes, and other recognitions, the University Materials Council (UMC)⁶⁸ could play a key role in developing possible performance metrics to measure success and provide guidance on how to use such metrics for awards and recognition. Some examples could include highlighting downloads on repositories and linking them to “badges” or “achievements” related to data sharing; such metrics could also be considered in tenure evaluations. Likewise, professional societies typically confer many different types of awards to their members and could establish some awards based on data sharing or impact. Conference organizers and journals could also establish similar awards. Awards of these types could include monetary awards on the order of \$1–\$10K, or be solely driven by community recognition.

Government agencies could also support awards. One recent example of this approach is the Materials Science and Engineering Data challenge, supported by the Air Force Research Laboratory (AFRL) in partnership with NIST and NSF. It encouraged data sharing and industry participation and winners were also recognized in a special symposium at the Materials Science and Technology Conference in 2016. Future challenges could involve a similar level of support of \$50K or less,

and could be coordinated by organizations including the federal agencies, professional societies, publishers, and private organizations.

The aforementioned incentives should be established in a short time frame of approximately 24 months or less.

Tactic #2: Execute and monitor the incentive programs to energize and sustain MSE community Involvement

After creation of incentive programs or mechanisms, execution is recommended as a twofold approach: short-term activities and long-term performance monitoring. While the former could result in immediate community excitement and activity, the latter, using some of the performance metrics developed from Tactic #1, will encourage the community members to sustain their engagement. Both execution of activities and monitoring of metrics related to the incentives described in Tactic #1 could be undertaken by university departments, the UMC, federal agencies, industry, and professional societies.

Recommendation 4: Develop demonstration projects and cross-disciplinary community efforts that enhance and accelerate adoption of the MDI

The phrase “demonstration projects” is used in the present context to indicate projects that demonstrate the value and utility of the MDI. Demonstration projects, especially when paired with cross-disciplinary community efforts, can help to accelerate the deployment and adoption of the MDI. These efforts will contribute to establishing the core technologies employed in the MDI and identify challenges that developers and practitioners must address.

Tactic #1: Establish materials-data-driven design projects that enable property prediction, to enhance MDI adoption

This tactic is anticipated to launch over a 12–18 month timeframe. It involves developing projects that would demonstrate materials property prediction via process-structure-property linkages and leverage materials data from across supply chains. These materials data-driven property predictions could be applied to improve existing materials, design new materials, repurpose existing materials for new applications, as well as develop materials processing methodologies. Some key steps would likely include identifying target product applications and aggregating data from relevant databases and data sources. This might include secure or anonymous aggregation of important industrial data.

Collaborations between scientists and engineers in government, industry, and academia are envisioned, and required support is estimated to be in the range of \$500K per project.

Tactic #2: Establish materials-data-driven projects for manufacturing, to enhance MDI adoption

Making materials-related datasets available to support manufacturing is an important step toward enabling new materials and manufacturing breakthroughs. This tactic focuses on accessing, creating, and employing materials processing and manufacturing datasets to demonstrate data-driven process

modeling in the manufacturing enterprise. An underlying goal of such demonstration projects is to identify areas where data could most easily be made non-proprietary. These might be relatively large collaborative projects, involving academia and industry, which have a duration of 24–36 months and cost in the neighborhood of \$5M per project. Even one or two projects of this nature could make a significant impact on enhancing the value, application, and adoption of the MDI. Similar efforts could also be undertaken, or leveraged within existing activities, such as Manufacturing USA and its manufacturing innovation institutes.

Publishing

In response to the science and engineering community trend toward open data, publishers have been working to make research data more accessible, discoverable, and reusable. Many publishers have updated their data policies and been working to accommodate the varied stakeholder interests including the needs of researchers, data repository producers or users, funders, and service providers. Consequently, many journals now range from encouraging data sharing to requiring data availability prior to acceptance.⁶⁹

Additionally, new journals have emerged that include opportunities for publishing descriptions of scientifically valuable data sets. There are generic examples, such as Scientific Data and Data in Brief, as well as domain-specific journals such as Integrating Materials and Manufacturing Innovation (IMMI).

Despite the growing list of modes and outlets for publishing data, hurdles remain that inhibit widespread adoption of research data sharing practices. An interview with Anita de Waard, Vice President of Research Data Collaborations at Elsevier, gave valuable insight into how one publisher is framing the issues. Dr. de Waard shared three major barriers to sharing science and engineering data:⁷⁰

- *Incentives to submit and publish data are unclear. Funding agencies and research institutions alike support the sharing of data; however, there are not many well-articulated incentives to publish detailed datasets from a data producer's perspective.*
- *Standardized methods of sharing are not available. Much work needs to be done to standardize the storage and sharing of data; there is currently too much freedom in approaches, which can inhibit the reuse of data.*
- *Very few examples currently exist to help illustrate the vision for published data. Success stories of great science being enabled through the storage and sharing of data are not prominent.*

To help overcome barriers to sharing data, publishers such as Elsevier have developed a variety of tools. For example, Elsevier encourages the use of Mendeley Data, an institutional repository to link data to papers with the assignment of a digital object identifier (DOI), as well as Hivebench, an electronic lab notebook to assist in data capture and reporting. Similarly, Wiley has partnered with Figshare, a London-based data repository organization, so that data can be easily uploaded and shared during submission of a manuscript.⁷⁰

Legal Issues

Legal issues are commonly cited as one of the barriers that hinder members of the materials community from openly sharing data.⁵ Michael Madison, a faculty member who teaches intellectual property and patent law courses within the School of Law at the University of Pittsburgh, and has published on the concept of the “knowledge commons,”⁷¹ shared some thoughts on what may benefit the development of a MDI:⁷²

- Technical issues should be prioritized over legal issues in the early stages of development of a data infrastructure. Legal issues are difficult to isolate and should not be viewed as a hindrance, but rather a way to help guide a community and ensure accountability where needed.
- Three areas of law are most likely to come up in the course of a developing data infrastructure:

Copyright. The relevant legal trigger for copyright in data is the concept of “originality.” Copyright exists if the person or people preparing the data do so in a way that is, in some non-trivial way, creative. The more factual, ordinary, or standardized the organization of the data, the less likely it is that copyright could be present. Data-based works, if they are “original,” earn at most a limited copyright, which means that verbatim copying of most or all of the dataset is prohibited but copying individual items or entries is not prohibited. Data creators and curators can claim that they own proprietary rights in data, based on copyright, and those claims often go unchallenged, for practical reasons. But as a matter of first principles, datasets often are subject to limited or zero copyright. Nonetheless, when researchers generate data and store it in a repository, it is possible that the publisher of that data may assert proprietary rights in the data, in addition to or on top of any proprietary rights claimed by the researchers, which adds to the complexity of what labels or licenses might be attached to data that one is interested in using. Individuals must therefore be alert in identifying and managing any restrictions on data they use.

Contract law. Separate from copyright, terms can be created that are more like binding contracts and incorporate terms of use, terms of access, and end user license agreements. These types of terms have a variety of implications when building a data archive. Terms of use are largely viewed as a binding mechanism even though sometimes researchers aren’t aware when they click through “agreement” clauses that they have agreed to them. If multiple people are using and contributing to a data commons, it is the responsibility of the manager to ensure people do not misuse the data once they have access. A data manager needs to ensure that researchers uploading data have not committed to binding contract law with other sets of data when trying to upload to another data repository.

Privacy. In the sponsored research world, personally identifiable data (PID) is a major liability; sharing data that is observational, experimental, or from a clinical trial can give rise to a host of problems. There is also the data lifecycle to consider regarding PID, as after initial consent to collect that data from the subject, it has to be stored and maintained in a way compliant with any PID restrictions. Further complicating matters, there is no set of clear rules available regarding who is liable if something goes wrong. Sponsors funding research generally have expectations about data sharing but you as the researcher have to ensure you are being compliant with any PID restrictions.

- Ultimately, it is not possible to fully eliminate legal risks entirely; however, with the proper guidelines in place, one can manage the risks and maintain an ethical data-sharing model.

***Tactic #3:** Launch targeted community efforts to help achieve MDI critical mass in the MSE community*

“Critical mass” in the present context refers to the number of implementers and users of the MDI needed to make it self-propagating, or organic, in its widespread adoption, growth, and sustainability. Although some specific types of community efforts are discussed in this tactic for reaching critical mass of users and implementers, many of the other recommendations and tactics in this report could also contribute to obtaining such critical mass, which will in turn accelerate the widespread adoption of the MDI.

Targeted efforts toward reaching a critical mass could encompass four different elements: (1) the service provider-level of MDI integration, (2) individual, user-level engagement of the MDI, (3) institutional-level MDI-focused working groups or committees, and (4) a broader community wide-level engagement (such as data-digitization projects).

In terms of service-provider-level engagement, researchers, small companies, industry stakeholders, and professional societies could pilot a program to build, curate, and monetize commercially valuable materials datasets and resources. This activity is also connected to Recommendation #5. Activities to stimulate integration among MDI service providers would also be useful and could include conducting MDI integration summits coordinated by service providers, federal agencies, or professional societies.

Some user-level activities could include hackathons, boot camps, and webinars, organized by federal agencies, professional societies, or industry. Coupled with Recommendation 2, agencies might also support a set of groundbreaking materials-research projects that are fundamentally enabled by large-scale data and a software infrastructure, and that merge information from traditionally separate data silos.

At the institutional level, MDI-focused working groups or committees could be launched by universities, federal laboratories, industrial enterprises, or professional societies. Examples of the types of activities that these groups might undertake include crowdsourced platforms for designing and developing components of the MDI, and specific tool recommendations of greatest local community need.

From a broader, community-wide level of engagement, a Wikipedia-like effort could be used to excite the community and help focus participation on the areas of greatest common interest. An online platform could be launched and seeded by a group of MDI champions across academia, government, and industry.

***Tactic #4:** Fill gaps in the MDI materials data domain*

This community-wide effort has two phases. The first short-term phase (6 months or less) is for preliminary investigation and gap analysis to identify and prioritize gaps in existing materials data resources. This could be accomplished by professional societies conducting and analyzing membership surveys, should cost on the order of \$10K–\$20K per project, and might involve two or three such projects. The surveys could be broken down by technical materials data areas (e.g., thermodynamics data, microstructure data, kinetics data, refractive index data, and mechanical property data).

The second phase (8–36 months) should focus on demonstrating the ability of stakeholders to fill critical materials data gaps identified in the first phase via simulations, experiments, existing legacy data, and data from the literature. One example could be employing machine learning data science to fill in some key gaps in materials data. Industry, academia, federal laboratories, database providers, and data scientists could all contribute to this activity. The cost is estimated to be \$1–\$2M distributed over a few projects and would presumably be covered by some combination of federal agencies (e.g., DOE, NSF, DoD). Private entities and consortia with a vested interest in filling specific materials data gaps might also provide support. This tactic will not only help provide a more robust MDI but also help demonstrate the value and utility of the MDI to the broader community.

Recommendation 5: Establish a MDI ecosystem and supporting business cases

More than just a federated system of technical platforms, a MDI requires establishment of an ecosystem, which includes a community of users and providers interacting within a digital infrastructure. The ecosystem is critical for sustainability and widespread adoption of the MDI, and must include viable MDI business cases for the members within such a community.

Tactic #1: Develop a reference architecture concept for the MDI

This overarching recommendation would involve creating a concept for a MDI reference architecture in order to provide a template, or model, for others to begin to follow, implement and refine. This will help facilitate a better understanding of the integrated technical, community, and business ecosystem challenges to be resolved. An example of a reference architecture has been provided by NIST for developing a cloud computing architecture.⁷³ In developing a reference architecture for a MDI for data storage and sharing, elements to consider include: the functional requirements, roles of major actors, and definition of functional interfaces (such as APIs) and their interactions. These elements would be represented schematically in such a reference architecture concept in order to provide high level guidance to the MDI community. Some of the challenges and tactical recommendations associated with these elements are touched on in other sections throughout this report. Developers of the reference architecture concept would necessarily be dedicated thought leaders in this area. It is envisioned that this could be a 12-month project, and that funding support for this effort might be provided by one of the federal agencies. The proposed reference architecture concept could be published in an appropriate outlet (e.g., in an agency report or peer-reviewed publication).

Tactic #2: Develop and demonstrate business cases for data storage and sharing

This tactic is focused on building increased confidence in the storage, use and open sharing of materials data by demonstrating strong MDI business cases. This could start with awareness through a campaign of advertisements (sponsored) or newsworthy highlights (free) in journals such as *JOM* or *IMMI*. Demonstrated MDI business case examples should include the following: 1) unique data storage and sharing strategies that were successfully employed to convert materials data to knowledge and resulted in materials innovation(s); 2) cost or efficiency improvements gained through MDI approaches, e.g., reducing internal hardware and software expenditures in

creation/acquisition of data because those data are already available for open use; and 3) various other incentives for sharing data that were realized in the specific business cases. As these materials-data-related business cases become more prevalent and better known within the community, it will help incentivize more users and developers to actively participate in the MDI ecosystem.

Tactic #3: Explore concept of a materials data “app store”

An online marketplace could be a potential business component of the MDI ecosystem providing long-term sustainability of technologies and data. Such an “app store” could promote the use and sharing of materials data via access to and community evaluation of various MDI-relevant repositories and tools. It might be an expansion of the registry concept described earlier, and could be arranged in an inclusive, user-friendly, and intuitive format. It could involve products for free or for purchase and would provide greater flexibility to access or purchase data or tools of interest. The look and structure of such an online store is only conceptualized here but it is envisioned to consist of the following three elements:

1. **Platform** – A platform makes it easy to contribute through its interface and facilitates the download and use of tools and data. An early demonstration of such a store could be realized through e-Collaboration platforms. An on-line store could also have some of the same functions of an e-Collaboration platform, but without the need for collaboration workspaces.
2. **Store** – A store enables the ability to conduct commerce, document metrics, perform analytics, provide rankings, and list reviews. It allows developers to package and make their resource(s) available and facilitates payment or some other type of reward.
3. **Content** – The content includes a well-defined set of technical tools and materials data useful to a particular sub-domain.

Implementers could include small businesses involved with materials data and related tools, as well as users and providers of materials data and tools from various sectors. This is seen more as a potential opportunity for entrepreneurs rather than a specific government investment.

Recommendation 6: Develop and invest in education and training programs for the MDI workforce (providers and users)

As considered in other workshops and reports,^{2,46,74} it is important to educate and train the present and future workforce in the value, use, and core technological underpinnings of materials data storage and sharing. Creating a strong workforce capable of building and exploiting the MDI is foundational for its ultimate success and sustainability. This entails engaging a combination of disciplines including MSE, information science, and computer science, since the MSE community currently lacks the trained workforce to fully develop the requisite MDI technologies. Education and training programs that will equip the current and future MDI workforce are therefore the subject of this recommendation.

Tactic #1: Integrate MDI into existing MSE curricula and build new cross-disciplinary curricula

Developing and launching new multidisciplinary curricula from scratch is complex and could be

hampered by long delays in the collaboration and administrative approval process. Thus, a two-pronged approach is recommended in which MDI-related concepts are integrated into existing MSE curricula in the short run, while new cross-disciplinary curricula are developed in the longer term. To establish effective curricula that provide the right mix of skills, the MSE community needs to leverage the wide range of existing resources and capabilities within academia. This encompasses many actions, or sub tactics, as described below.

As a first step, specific needs, opportunities, and benefits within MDI curricula need to be articulated in detail. Although this could be done at the individual department level within universities, it might better be accomplished by a focused workshop(s) (see also other workshop suggestions⁴⁵). Such a workshop could be held in conjunction with ABET advisory groups and coordinated by organizations including the UMC, NSF, professional societies, and ABET coordinators. Resulting reports could then be disseminated to university departments, as well as the broader community, with assistance from the UMC and professional societies. This type of report should leverage past workshops and reports on broader MSE workforce development (see, for example, Pollock et al.⁷⁵), but it should focus exclusively on materials data issues and include educational needs related to community-level best practices for a MDI.

Examples of existing cross-disciplinary curricula include FLAMEL (From Learning, Analytics, and Materials to Entrepreneurship and Leadership Doctoral Traineeship)⁷⁶ at Georgia Tech and D³EM (Data-Enabled Discovery and Design of Energy Materials)⁷⁷ at Texas A&M University. These programs are for graduate students interested in utilizing the latest data-driven techniques toward application of the design and manufacture of new materials. We note that curricula can come in many forms and should be taken to refer to any easily accessible educational materials and resources for student training. These might be online lectures, Massive Online Open Courses (MOOCs)⁷⁸, training workshops, software tools, specific modules for classrooms, etc., all of which can support rapid and efficient adoption of MDI in educational settings.

In addition to curriculum development and changes, internship programs or fellowships should be initiated to further integrate MDI concepts into the workforce. These internships and fellowships would give beneficial scientific and technical experiences to students and faculty in industry, federal laboratories, or universities, while focusing on different components of the MDI.

Tactic #2: Conduct outreach and training programs for professionals

While development and execution of curricula within universities is critical for the future workforce, retraining current professionals is also important to foster the development and adoption of the MDI and its use to solve real-world problems. Today's engineers and scientists need to leverage materials data and use the best data-related tools to solve science and engineering challenges in an innovative and fast-paced environment. Developing and delivering effective professional development and training strategies are thus needed to help researchers and engineers build expertise in MDI approaches.

Online offerings:

Online professional development activities should include webinars and discussion forums, among other efforts. These activities could be spearheaded by professional societies

working with members of industry, academia, and government, perhaps in association with the committees and other groups mentioned in Tactic #3 of Recommendation #4. Some of these efforts could be in the form of smaller webinars, such as the TMS webinar series on materials data curation, or alternatively a larger model such as MOOCs.⁷⁸ A number of these online courses could be stood up within 12–24 months, and could be developed by professional societies and the relevant volunteer committees therein, as well as by universities or government organizations.

Another possible activity is to establish dedicated online resources for materials data management education, perhaps analogues to the ESIP commons, which is a knowledge repository created by members of the Earth Science Information Partners (ESIP) community, and contains training modules on data and other issues.⁷⁹

In-person activities:

In-person environments such as short courses, boot camps (more advanced/intensive), and other immersive learning or professional development training programs should also be developed. These could perhaps have similarities to Data Carpentry courses⁸⁰ but would be more specifically tailored to materials data infrastructure issues. These courses can be developed by some combination of personnel from academia, service providers, federal laboratories, and professional societies, again possibly leveraging the committees and other groups referenced in Recommendation #4, Tactic #3. Some examples of training topics could include: (1) how to integrate state-of-the-art equipment with “Lab to Cloud” software for storing and sharing datasets; (2) how to increase productivity through workflow capture and data management tools; (3) using modern informatics tools, including teaching MSE researchers how to employ such tools to help promote use of their products (e.g., data, codes, and knowledge) by other potential stakeholders; (4) training on how to use modern citation practices so data and tool providers can get credit for their data and tool usage; and (5) training users in the application of machine learning methodologies to materials data. A fiscal model for development and execution would have to be established for each of these courses and could be based on funding support from either companies, registration fees, federal funding agencies, or some combination thereof. These activities should be developed and executed over the next 1–3 years.

Recommendation 7: Create MDI consortia and working groups

Federal agencies and professional societies should convene integrated working groups comprising members across government, industry, and academia to identify best practices and recommend coordinated approaches for MDI protocols, standards, tools, ontologies, etc.

Tactic #1: Create a long-term coordinating and advisory body comprising relevant MDI stakeholders

Creation of a MDI council or public-private consortium for consistent MDI stewardship is recommended. After determining governance and a charter, the council or consortium could convene

relevant stakeholders (e.g., government personnel, academics, equipment manufacturers, software companies, industry, cybersecurity experts, and publishers) to discuss a plan for coordinating MDI activities within the next 24 months. Although they would collectively develop their charter for best addressing the MDI, some suggested items to be considered include: (1) determining the key current and upcoming efforts and implementation personnel involved with MDI-related activities; (2) offering guidance and suggestions on coordinating their activities to support a more unified MDI with interoperable components; (3) developing general approaches and policies in support of the MDI; (4) identifying unified ways for publishing data; (5) addressing IT access and firewall security concerns associated with data sharing, access, and storage; (6) addressing integration and modernization of commercial infrastructure; and (7) developing clear, unified marketing and education messages to the MSE community. The challenges and recommendations in this report, and some of the other resources referenced herein, provide good information sources to scope the highest-priority items for this coordination group to address. This group could initially be convened by a government agency or professional society but would then become a self-sustaining entity, possibly in a model like the University Materials Council (UMC). Existing industrial and other materials-related consortia should be encouraged to interact with this coordination and advisory body wherever such leveraging is deemed useful.

Tactic #2: Create Community of Practice (CoP) groups around the MDI

The role of community of practice groups consisting of existing MDI providers could be quite important. These groups would work in parallel with the more overarching, long-term coordinating and advisory body described in the previous tactic, and would have much more specific objectives. These materials data communities of practice (CoPs) could leverage the activities of other, broader groups like the Research Data Alliance (RDA)¹⁷ and the Materials Accelerator Network,⁸¹ but would be focused on specific issues pertaining to the MDI. Funding agencies could provide resources, and professional societies could help establish these groups. For instance, CoP groups could be formed in conjunction with the committees referenced in Recommendation #4, Tactic #3. These CoPs could be convened within a year and produce community outputs and guidance within two years. One specific example is to formalize a community of practice organization to create data/metadata standards and protocols for data lifecycle, best practices, and reproducibility. Outputs of CoP groups could include data schema, practices, and ontologies, as well as interoperable data formats. An example CoP model is ESIP's collaboration areas.⁷⁹

Recommendation 8: Define and establish clear policies and guidelines associated with the MDI

More consistent, carefully developed policies are needed for stewardship of a MDI. Such policies not only provide a key driving force for the MDI, but avoid unnecessary activities and provide the best solutions for the MDI technologies and ecosystem. The overall goal of such policies is to help guide and regulate standard data services and infrastructures.

Tactic #1: Establish an interagency (federal) council to foster consistent data preservation and sharing policies

In particular, this council would help to develop unified data management plan (DMP) requirements

which would be based on FAIR principles—making materials science data findable, accessible, interoperable, and reusable. As described in the Background section, the FAIR principles are guidelines to provide both computers and humans support for data sharing, storage, and automated cataloging.⁸ The guidance and policies developed by this council should have an overarching theme of supporting consistency within the MDI; it is important that all the federal agencies supporting materials science coordinate on this effort. This council could be formed within 6 months and subsequently within one year develop initial materials-related DMP policies and plans which are consistent with broader DMP policies within their agencies. This council could also solicit advice from and coordinate with the broader advisory body described in Recommendation #7, Tactic #1.

Tactic #2: Increase emphasis on Data Management Plans for funding support considerations

This tactic encourages federal agencies to consider elevating the weight of Data Management Plan (DMP) scores in proposal evaluations, as well as evaluating past DMP performance of proposers and awardees. Without agency follow through, adhering to DMP policies is often a low priority. This could be implemented by all of the agencies supporting materials science (e.g. DOE, NASA, NIST, NSF, DoD), with the lead being taken by individual program officers in research programs that have materials data components. This increased emphasis on DMPs directly dovetails with developing unified DMPs in tactic #1, and would also have to be consistent with any broader agency policies. Another mechanism for consideration is for the federal agencies to mandate publication of underlying data supporting a peer-reviewed article, where “publication” is defined more broadly in this context as making it publicly accessible somewhere within the MDI. Such a mandate would be consistent with the Office of Science and Technology Policy Memorandum “Increasing Access to the Results of Federally Funded Scientific Research” from February 2013.¹⁰

Conclusion

Digital data will play a vital role in contributing to scientific and engineering discoveries and innovations associated with materials science and engineering (MSE) in the coming years. Both experimental and computational data are critical to such breakthroughs, yet the value and impact of much of the materials data that is produced is currently far from being fully realized. This is in large part due to a lack of concerted, coordinated materials data storing and sharing efforts across the MSE and related communities. In other words, it is essential that the relevant stakeholders come together to develop a unified, collaborative approach to storing, sharing, and optimally using materials data. A primary goal of this study has thus been to provide knowledge and guidance to help in the development of a robust materials data infrastructure (MDI). *The Materials Data Infrastructure (MDI) consists of three core digital components—repositories, tools, and e-collaboration platforms—as well as the technology, policies, incentives, standards, people, and related activities necessary to plan, acquire, process, analyze, store, share, reuse, and dispose of materials data* (see also Figure 1 of the Introduction).

To help provide key knowledge and guidance towards further development of a robust MDI, the experts contributing to this study have concentrated their efforts in a few areas: (1) identifying major challenges to long-term storage and sharing of materials data, (2) developing recommendations for overcoming the challenges that are identified, and (3) providing tactical suggestions for specific actions towards development of a robust MDI along with details on timeframes, implementers, estimated costs, and possible sources of financial support.

A major aim of this study report is to *stimulate direct action by a wide variety of people who read this report, and who may benefit from or contribute to a robust materials data infrastructure.*

In today's highly digital and integrated world, a coordinated materials data infrastructure will be a key enabler for accelerating materials-related science and engineering breakthroughs in the 21st century and beyond. It is our desire that the readers of this report will use it to elucidate ways in which they and their colleagues can best contribute to and realize the great benefits of such a MDI, and will act promptly to contribute to its development.

References

1. White House Office of Science and Technology Policy. *Materials Genome Initiative for Global Competitiveness*. (2011); https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf.
2. *Materials Genome Initiative Strategic Plan*. (National Science and Technology Council, Committee on Technology Subcommittee on the Materials Genome Initiative, 2014); https://mgi.nist.gov/sites/default/files/factsheet/mgi_strategic_plan_-_dec_2014.pdf.
3. Alliance Permanent Access to the Records of Science in Europe Network. (2015). <http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/>.
4. Schmitz, G.J. and Prah, U. "ICMEg – The Integrated Computational Materials Engineering Expert Group – A New European Coordination Action," *Integrating Mater. Manuf. Innov.* 3, 2 (2014).
5. "Preliminary Results of "Big Data" Survey Provide Perspective on Open Research Topics," *JOM* 65, 1072–1073 (2013); <https://link.springer.com/content/pdf/10.1007%2Fs11837-013-0724-y.pdf>.
6. "OMB Circular A-16 and Supplemental Guidance." (Reston, VA: Federal Geographic Data Committee, 2002); <https://www.fgdc.gov/policyandplanning/a-16>.
7. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. (National Research Council, 2008); <https://www.nap.edu/catalog/12199/integrated-computational-materials-engineering-a-transformational-discipline-for-improved-competitiveness>.

8. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L.B. da S., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C.'t, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Sci. Data* 3, 160018 (2016).
9. Stebbins, M. "Expanding Public Access to the Results of Federally Funded Research," White House Blog, Feb 2013, <https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>.
10. "Increasing Access to the Results of Federally Funded Science." (2013); https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
11. *NSF Public Access Plan: Today's Data, Tomorrow's Discoveries*. (National Science Foundation, 2015); <https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>.
12. *Plan for Providing Public Access to the Results of Federally Funded Research*. (National Institute of Standards and Technology, 2015); <https://www.nist.gov/sites/default/files/documents/data/NIST-Plan-for-Public-Access.pdf>.
13. CENDI. "Public Access Plans of U.S. Federal Agencies." (Accessed 03/23/2017); https://cendi.gov/projects/Public_Access_Plans_US_Fed_Agencies.html#PubAccPlans.
14. *NASA Plan for Increasing Access to the Results of Scientific Research*. (National Aeronautics and Space Administration, 2014); https://www.nasa.gov/sites/default/files/atoms/files/206985_2015_nasa_plan-for-web.pdf.
15. "FORCE11 Manifesto." (2011); <https://www.force11.org/about/manifesto>.
16. "Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0." *FORCE11*. (2014); <https://www.force11.org/fairprinciples>.
17. About RDA. (Research Data Alliance (RDA) (2016)); <https://www.rd-alliance.org/about-rda>.
18. All Recommendations & Outputs. (Research Data Alliance (RDA) (2016)); <https://www.rd-alliance.org/recommendations-and-outputs/all-recommendations-and-outputs>.
19. RDA/CODATA Materials Data, Infrastructure & Interoperability IG. (Research Data Alliance (RDA) (2013)); <https://www.rd-alliance.org/groups/rdacodata-materials-data-infrastructure-interoperability-ig.html>.
20. *Common Framework for Earth-Observation Data*. (Committee on Environment, Natural Resources, and Sustainability of the National Science and Technology Council, 2016); https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/common_framework_for_earth_observation_data.pdf.
21. Bristol, S. (U.S. Geological Survey), in discussion with J.A. Scott and L.T. Beringer (The Minerals, Metals & Materials Society), January 2017.
22. *National Plan for Civil Earth Observations*. (Office of Science and Technology Policy, Executive Office of the President, 2014); https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/NSTC/2014_national_plan_for_civil_earth_observations.pdf.
23. CyVerse. (Accessed 03/23/2017); <http://www.cyverse.org/>.

24. Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., and Antin, P. "The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences," *PLOS Biol.* 14, e1002342 (2016).
25. Antin, P. and Merchant, M. (University of Arizona), in discussion with J.A. Scott and L.T. Beringer (The Minerals, Metals & Materials Society), January 2017.
26. "Data Lifecycle Models and Concepts v11." (USGS Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services (WGISS), 2012); <https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2>.
27. Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martín, E., Montgomery, E.T., Ladino, C., Tessler, S., and Zolty, L.S. *The United States Geological Survey Science Data Lifecycle Model*. 12 (Reston, VA: U.S. Geological Survey, 2014); <http://pubs.er.usgs.gov/publication/ofr20131265>.
28. Tschoop, M.A., Coleman, S.P., and McDowell, D.L. "Symmetric and Asymmetric Tilt Grain Boundary Structure and Energy in Cu and Al (and Transferability to Other FCC Metals)," *Integrating Mater. Manuf. Innov.* 4, 11 (2015).
29. Olmsted, D.L., Foiles, S.M., and Holm, E.A. "Survey of Computed Grain Boundary Properties in Face-Centered Cubic Metals: I. Grain Boundary Energy," *Acta Mater.* 57, 3694–3703 (2009).
30. Olmsted, D.L., Holm, E.A., and Foiles, S.M. "Survey of Computed Grain Boundary Properties in Face-Centered Cubic Metals—II: Grain Boundary Mobility," *Acta Mater.* 57, 3704–3713 (2009).
31. Cahn, R.W. *The Coming of Materials Science*. (Amsterdam: Pergamon, 2003).
32. Boyce, D.E., Dawson, P.R., and Miller, M.P. "The Design of a Software Environment for Organizing, Sharing, and Archiving Materials Data," *Metall. Mater. Trans. A* 40, 2301–2318 (2009).
33. Dima, A., Bhaskarla, S., Becker, C., Brady, M., Campbell, C., Dessauw, P., Hanisch, R., Kattner, U., Kroenlein, K., Newrock, M., Peskin, A., Plante, R., Li, S.-Y., Rigodiat, P.-F., Amaral, G.S., Trautt, Z., Schmitt, X., Warren, J., and Youssef, S. "Informatics Infrastructure for the Materials Genome Initiative," *JOM* 68, 2053–2064 (2016).
34. Seshadri, R. and Sparks, T.D. "Perspective: Interactive Material Property Databases Through Aggregation of Literature Data," *APL Mater.* 4, 053206 (2016).
35. Energy Materials Datamining; <http://tomcat.eng.utah.edu/sparks/battery.jsp>.
36. Gaultois, M.W., Sparks, T.D., Borg, C.K.H., Seshadri, R., Bonificio, W.D., and Clarke, D.R. "Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations," *Chem. Mater.* 25, 2911–2920 (2013).
37. Hall, S.R., Allen, F.H., and Brown, I.D. "The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography," *Acta Crystallogr. A* 47, 655–685 (1991).
38. O'Mara, J., Meredig, B., and Michel, K. "Materials Data Infrastructure: A Case Study of the Citrination Platform to Examine Data Import, Storage, and Access," *JOM* 68, 2031–2034 (2016).
39. Ward, C.H., Warren, J.A., and Hanisch, R. "Making Materials Science and Engineering Data More Valuable Research Products," *Integrating Mater. Manuf. Innov.* 3, 22 (2014).
40. Kalidindi, S.R. and Graef, M.D. "Materials Data Science: Current Status and Future Outlook," *Annu. Rev. Mater. Res.* 45, 171–193 (2015).

41. Kalidindi, S.R. “Data Science and Cyberinfrastructure: Critical Enablers for Accelerated Development of Hierarchical Materials,” *Int. Mater. Rev.* 60, 150–168 (2015).
42. Agrawal, A. and Choudhary, A. “Perspective: Materials Informatics and Big Data: Realization of the ‘Fourth Paradigm’ of Science in Materials Science,” *APL Mater.* 4, 053208 (2016).
43. Austin, T. “Towards a Digital Infrastructure for Engineering Materials Data,” *Mater. Discov.* 3, 1–12 (2016).
44. Warren, J.A. and Boisvert, R.F. *Building the Materials Innovation Infrastructure: Data and Standards*; <http://dx.doi.org/10.6028/NIST.IR.7898>.
45. *Big Data in Materials Research and Development: Summary of a Workshop*. (National Research Council, 2014); <https://www.nap.edu/catalog/18760/big-data-in-materials-research-and-development-summary-of-a>.
46. *Building the Materials Data Infrastructure: A Materials Community Planning Workshop*. (Materials Park, OH: ASM International, 2015); http://www.asminternational.org/documents/10192/19715738/Materials+Community+Planning+Workshop+Report_030515/3960c291-d2d3-446c-95d7-89740a57a233.
47. *In-Process Materials Data for Modeling Workshop Report*. (Materials Park, OH: ASM International, 2015); http://www.asminternational.org/documents/10192/25764557/ASM+In-Process+Materials+Data+Workshop+Summary_2015-08.pdf/c5a6684b-91c3-4893-ae6d-474622c98853.
48. Bethel, W., Greenwald, M., van Dam, K., Parashar, M., Wild, S., and Wiley, H. “Management, Analysis, and Visualization of Experimental and Observational Data – The Convergence of Data and Computing.” (2016); https://science.energy.gov/~media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf.
49. Rise of Data in Materials Research. (Accessed 03/23/2017); <http://riseofdata.org/>.
50. Center for Hierarchical Materials Design (CHiMaD). (Accessed 03/23/2017); http://chimad.northwestern.edu/news-events/Event_Archives.html.
51. BaBar. (Accessed 03/23/2017); <http://www.slac.stanford.edu/BFROOT/>.
52. Becla, J. and Wang, D.L. “Lessons Learned from Managing a Petabyte,” *Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR 2005)* 70–83 (2005); <http://cidrdb.org/cidr2005/index.html>.
53. Designing for Peta-Scale in the LSST Database. (Accessed 03/23/2017); <http://aspbooks.org/custom/publications/paper/376-0003.html>.
54. Jurić, M., Kantor, J., Lim, K.-T., Lupton, R.H., Dubois-Felsmann, G., Jenness, T., Axelrod, T.S., Aleksić, J., Allsman, R.A., AlSayyad, Y., Alt, J., Armstrong, R., Basney, J., Becker, A.C., Becla, J., Bickerton, S.J., Biswas, R., Bosch, J., Boutigny, D., Carrasco Kind, M., Ciardi, D.R., Connolly, A.J., Daniel, S.F., Daues, G.E., Economou, F., Chiang, H.-F., Fausti, A., Fisher-Levine, M., Freemon, D.M., Gee, P., Gris, P., Hernandez, F., Hoblitt, J., Ivezić, Ž., Jammes, F., Jevremović, D., Jones, R.L., Bryce Kalmbach, J., Kasliwal, V.P., Krughoff, K. S., Lang, D., Lurie, J., Lust, N.B., Mullally, F., MacArthur, L.A., Melchior, P., Moeyens, J., Nidever, D.L., Owen, R., Parejko, J.K., Peterson, J.M., Petravick, D., Pietrowicz, S.R., Price, P.A., Reiss, D.J., Shaw, R.A., Sick, J., Slater, C.T., Strauss, M.A., Sullivan, I.S., Swinbank, J.D., Van Dyk, S., Vujčić, V., Withers, A., and Yoachim, P. “The LSST Data Management System,” *ArXiv E-Prints* (2015); <https://ui.adsabs.harvard.edu/#abs/2015arXiv151207914J/abstract>.
55. Materials Data Curation System. (Accessed 03/23/2017); <https://materials.registry.nist.gov/>.
56. NIST Alloy Data. (Accessed 03/23/2017); http://trc.nist.gov/metals_data/.

-
57. Materials Commons. (Accessed 03/23/2017); <https://materialscommons.org/mcapp/#/login>.
 58. Puchala, B., Tarcea, G., Marquis, E.A., Hedstrom, M., Jagadish, H.V., and Allison, J.E. "The Materials Commons: A Collaboration Platform and Information Repository for the Global Materials Community," *JOM* 68, 2035–2044 (2016).
 59. Citrination. (Accessed 03/23/2017); <https://www.citrination.com/search/simple>.
 60. Materials Resource Registry. (Accessed 03/23/2017); <https://mgi.nist.gov/materials-resource-registry>.
 61. Home Page of re3data.org. (Accessed 03/23/2017); <http://www.re3data.org/>.
 62. *National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization*. (National Science Foundation, 2011); https://www.nsf.gov/cise/aci/taskforces/TaskForceReport_Data.pdf.
 63. User guide: contents — scikit-learn 0.18.1 documentation. (Accessed 03/23/2017); http://scikit-learn.org/stable/user_guide.html.
 64. Jacobsen, M. D., Fourman, J. R., Porter, K. M., Wirrig, E. A., Benedict, M. D., Foster, B. J., and Ward, C. H. "Creating an Integrated Collaborative Environment for Materials Research," *Integrating Mater. Manuf. Innov.* 5, 12 (2016).
 65. nanoHUB. (Accessed 03/23/2017); <https://nanohub.org/>.
 66. MATIN. (Accessed 03/23/2017); <https://matin.gatech.edu>.
 67. T2C2: Timely and Trusted Curation and Coordination. (Accessed 03/23/2017); <http://t2c2.cs.illinois.edu/>.
 68. University Materials Council. (Accessed 03/23/2017); <http://umatcon.org/>.
 69. Data Sharing. (Accessed 03/23/2017); <https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-and-open-access/open-access/data-sharing.html>.
 70. de Waard, A. (Elsevier), in discussion with J.A. Scott and L.T. Beringer (The Minerals, Metals & Materials Society), January 2017.
 71. "The Knowledge Commons Research Framework. *Workshop Gov. Knowl. Commons* (2014). <http://knowledge-commons.net/publications/gkc/research-framework/>.
 72. Madison, M. (University of Pittsburgh), in discussion with J.A. Scott and L.T. Beringer (The Minerals, Metals & Materials Society), January 2017.
 73. Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., and Leaf, D. *NIST Cloud Computing Reference Architecture*. (National Institute of Standards and Technology, 2011); http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=909505.
 74. Pablo, J.J. de, Jones, B., Kovacs, C.L., Ozolins, V., and Ramirez, A.P. "The Materials Genome Initiative, the Interplay of Experiment, Theory and Computation," *Curr. Opin. Solid State Mater. Sci.* 2, 99–117 (2014).
 75. Pollock, T., Seshadri, R., Bahr, D., Cahill, D., Hemker, K., and Lesar, R. *Workshop on the Future of Graduate Education in Materials Science*. (University of California, Santa Barbara: 2015); http://www.umatcon.org/downloads/MSE_Grad_Education_Pollock.pdf.
 76. "From Learning, Analytics, and Materials to Entrepreneurship and Leadership Doctoral Traineeship Program (FLAMEL)." (Accessed 03/23/2017); <http://flamel.gatech.edu/>.
 77. "Data-Enabled Discovery and Design of Energy Materials (D3EM)." (Accessed 03/23/2017); <http://d3em.tamu.edu/about-d3em-scholarship/>.
 78. Rodriguez, C.O. "MOOCs and the AI-Stanford Like Courses: Two Successful and Distinct Course Formats for Massive Open Online Courses," *Eur. J. Open Distance E-Learn.* (2012). <https://eric.ed.gov/?id=EJ982976>.
-

79. ESIP Collaboration Areas. (Accessed 03/23/2017); <http://www.esipfed.org/collaboration-areas>.
80. Data Carpentry. (Accessed 03/23/2017); <http://www.datacarpentry.org/>.
81. Materials Accelerator Network. (Accessed 03/23/2017); <http://acceleratornetwork.org/>.

Additional Reading

- Borgman, C. L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. (2015).
- Briney, K. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*. (2015).
- Hey, A.J.G. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. (2009).
- Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., and Janée, G. "Dataone: Data Observation Network for Earth-Preserving Data and Enabling Innovation in the Biological and Environmental Sciences," *D-Lib Magazine* 17, 3 (2011).
- Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D.P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E.L., Simonsohn, U., Soderberg, C., Spellman, B.A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.J., Wilson, R., and Yarkoni, T. "Promoting an Open Research Culture," *Science* 348, 1422 (2015).
- Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., and Taufer, M. "Enhancing Reproducibility for Computational Methods," *Science* 354, 1240 (2016).
- Stodden, V. *Enabling Reproducible Research: Open Licensing for Scientific Innovation*. (Rochester, NY: Social Science Research Network, 2009); <https://papers.ssrn.com/abstract=1362040>.
- Ward, C.H. *Implications of Integrated Computational Materials Engineering with Respect to Export Control*. (2013); <http://www.dtic.mil/get-tr-doc/pdf?AD=ADA590928>.

Wong, T., Venkatesh, V., and Turner, T.J. “Data Infrastructure Developed for PW-8: Nickel Base Superalloy Residual Stress Foundational Engineering Problem,” *Proceedings of the 3rd World Congress on Integrated Computational Materials Engineering (ICME 2015)*, ed. W. Poole et al. (New York: Springer, 2015), pp. 247–259; doi:10.1007/978-3-319-48170-8_30.

For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. (Washington, DC: National Research Council, 2012); <https://www.nap.edu/catalog/13564/for-attribution-developing-data-attribution-and-citation-practices-and-standards>.

The First Five Years of the Materials Genome Initiative: Accomplishments and Technical Highlights (OSTP). (2016). <https://mgi.nist.gov/sites/default/files/uploads/mgi-accomplishments-at-5-years-august-2016.pdf>.

RDA & CODATA Legal Interoperability of Research Data: Principles and Implementation Guidelines. *RDA* (2016). <https://www.rd-alliance.org/rda-codata-legal-interoperability-research-data-principles-and-implementation-guidelines-now>.

Appendix A: Acronyms & Abbreviations

AAAS	American Association for the Advancement of Science
AFRL	Air Force Research Laboratory
AI	artificial intelligence
ANL	Argonne National Laboratory
APIs	application program interfaces
CALPHAD	CALculation of PHase Diagrams
CHiMaD	Center for Hierarchical Materials Design
CoP	Community of Practice
D ³ EM	Data-Enabled Discovery and Design of Energy Materials
DMP	Data Management Plans
DoD	Department of Defense
DOE	Department of Energy
DOI	digital object identifier
ESIP	Earth Science Information Partners
FAIR	Findable, Accessible, Interoperable, Reusable
FLAMEL	From Learning, Analytics, and Materials to Entrepreneurship and Leadership
FOAs	funding opportunity announcements
GEO	(U.S.) Group on Earth Observations

GUIs	graphical user interfaces
ICME	Integrated Computational Materials Engineering
ICMEg	Integrated Computational Materials Engineering Group
LSST	Large Synoptic Survey Telescope
MDI	Materials Data Infrastructure
MDCS	Materials Data Curation System
MGI	Materials Genome Initiative
MOOCs	massive online open courses
MRS	Materials Research Society
MSE	Materials Science and Engineering
NASA	National Aeronautics and Space Administration
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NMRR	NIST Materials Resource Registry
NOAA	National Oceanic and Atmospheric Administration
NRC	National Research Council
NSF	National Science Foundation
NSTC	National Science and Technology Council
OSTP	White House Office of Science and Technology Policy
PID	personally identifiable data
PRISMS	PRedictive Integrated Structural Materials Science
RDA	Research Data Alliance
SEM	scanning electron microscopy
SI2	Software Infrastructure for Sustained Innovation
SLAC	Stanford Linear Accelerator Center
TEM	transmission electron microscopy
TMS	The Minerals, Metals & Materials Society
UMC	University Materials Council
UQ	uncertainty quantification
USGS	United States Geological Survey
XPS	x-ray photoelectron spectroscopy

Appendix B: Summary of Prior Workshops & Event Outputs

High-level summaries of some of the key example challenges and outputs discussed at recent workshops and events related to Materials Data Infrastructure.

ICME: A Transformational Discipline for Improved Competitiveness and National Security (2008)

Source: National Materials Advisory Board, Division on Engineering and Physical Sciences, National Research Council

URL: <https://www.nap.edu/catalog/12199/integrated-computational-materials-engineering-a-transformational-discipline-for-improved-competitiveness>

Synopsis and Highlights: This report was one of the first to discuss the value of digital materials data and to emphasize its importance to the materials scientific and engineering community. Several challenges and opportunities for building a digital database are mentioned. Some key challenges include lack of standards, IP restrictions on access to and sharing of data, and lack of multidisciplinary collaborations, whereas opportunities include creation of an open-access platform that integrates computational and experimental data.

Materials Genome Initiative for Global Competitiveness White Paper (2011)

Source: National Science and Technology Council, Office of Science and Technology Policy

URL: https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf

Synopsis and Highlights: This white paper introduced the Materials Genome Initiative (MGI), which aims to greatly reduce the time and cost of developing new, advanced materials within the United States by leveraging existing experiments and computational tools for an integrated approach. A major goal within the MGI is to develop a materials innovation infrastructure, composed of three critical pieces: computational tools, experimental tools, and digital data.

TMS/ASM Intersociety Scoping Session on Materials Data Management (2012)

Source: The Minerals, Metals, & Materials Society (TMS) and ASM International in conjunction with the 2012 Materials Science & Technology Conference

URL: <http://materialsinnovation.tms.org/2012IntersocietyScopingSession/IntersocietyDataSessionMST12ReportFinal.pdf>

Synopsis and Highlights: This workshop had two major goals: (1) to identify the primary categories of data associated with materials science and engineering with respect to structural materials and (2) to assess the current states of primary data sharing and tools. Several types of data were identified (raw, metadata, derived data, method, model verification and validation, design values, statistics) along with states of data sharing (familiarity, accessibility, community uptake and maturity of data and tools). One major crosscutting issue identified was the ability to share complex data in an appropriate format among the community.

Building the Materials Innovation Infrastructure: Data and Standards (2012)

Source: National Institute of Standards and Technology

URL: <http://nvlpubs.nist.gov/nistpubs/ir/2012/NIST.IR.7898.pdf>

Synopsis and Highlights: This workshop identified several key themes that related to data infrastructure needs and included the following:

- Accurate models of materials performance which are validated using experimental data
- Open-platform frameworks to ease the development/operation of simulation codes
- Software that is modular and user-friendly with applicability to broad user communities
- Data repositories built on community standards and outfitted with modern search, retrieval, and analysis tools

Length scale challenges were a major outcome of this workshop, and encompass macro, micro, nano/molecular, and atomic length scales.

Integrated Computation Materials Engineering (ICME): Implementing ICME in the Aerospace, Automotive, and Maritime Industries (2012)

Source: The Minerals, Metals & Materials Society

URL: <http://www.tms.org/icmestudy/>

Synopsis and Highlights: This report identified, prioritized, and provided detailed frameworks for implementing Integrated Computational Materials Engineering (ICME) in the automotive, aerospace, and maritime industries. A subset of recommendations within this report involve data infrastructure and standards and include: establish adequate standards, data, and integration (as it related to manufacturing supply chains). To do this research groups must set data standards and classifications, develop data/workflow strategies, and increase communication efforts between ICME stakeholders. Implementation strategies for data standardization were identified and some examples include:

- Catalog information on available standards, collect case studies
- Define taxonomies (related groups of data, materials properties, etc.)
- Ensure IP is protected

Workshop on the Materials Genome Initiative - The Interplay of Experiment, Theory and Computation (2012)

Source: Juan J. de Pablo, Barbara Jones, Cora Lind-Kovacs, Vidvuds Ozolins, Arthur Ramirez

URL: <http://dx.doi.org/10.1016/j.cossms.2014.02.003>

Synopsis and Highlights: This report summarizes key outputs resulting from a NSF workshop relating to the MGI held in December 2012. Recommendations were broken down into logistical and technical considerations that include, for example, promoting an all-encompassing MGI mode of research and creating networks of experts from different disciplines and industries. Key areas of opportunity and associated grand challenges by materials class were identified and additional recommendations concerning workforce development were discussed.

Big Data in Materials Research and Development: Summary of a Workshop (2014)

Source: Defense Materials Manufacturing and Infrastructure Standing Committee, Division on Engineering and Physical Sciences, National Research Council

URL: <https://www.nap.edu/catalog/18760/>

Synopsis and Highlights: This Defense Materials Manufacturing and Infrastructure (DMMI) workshop was held to discuss concepts of big data storage and the unique challenges associated with storing terabytes of data in various materials communities. The six major themes that emerged from this workshop were: (1) data availability, (2) “big data” vs. data, (3) quality and veracity of data and models, (4) data and metadata ontology and formats (5), metadata and model availability, and (6) culture.

Materials Genome Initiative - Strategic Plan (2014)

Source: National Science and Technology Council, Committee on Technology, Subcommittee on the MGI Initiative

URL: https://www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf

Synopsis and Highlights: The strategic plan response to the 2011 MGI announcement identifies four goals: (1) enable a paradigm shift in culture, (2) integrate experiments, computation, and theory, (3) facilitate access to materials data, and (4) equip the next generation materials workforce. Of these four goals, two (goals 2 and 3) are especially related to establishing a materials data infrastructure (MDI). More specifically, goal (3) details objectives and milestones centered on identifying best practices for implementation of a materials data infrastructure, and supporting creation of accessible materials data repositories.

Materials Genome Initiative: Materials Data (2014)

Source: Air Force Research Laboratory and National Institute of Standards and Technology

URL: <http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8038.pdf>

Synopsis and Highlights: This workshop convened multiple federal agency stakeholders from ARL, DOE-EERE, NIST, ONR, NASA, AFRL and members of the community to promote communication between groups supported under the MGI that had a focus on managing materials data. Participants presented their group's contributions towards building a materials infrastructure and in managing data.

Building an Integrated MGI Accelerator Network (2014)

Source: Materials Accelerator Network

URL: <http://acceleratornetwork.org/events/past-events/building-an-integrated-mgi-accelerator-network/>

Synopsis and Highlights: This workshop brought together 150 thought leaders and stakeholders from across the nation in areas of academia, industry, and government. It was co-organized by Georgia Institute of Technology, University of Wisconsin-Madison, and the University of Michigan. A national dialogue was initiated regarding the Materials Accelerator Network concept, as highlighted in the OSTP press release on the second anniversary of MGI, June 24, 2013. High priority recommendations included (1) Focusing on education and training of the future MGI workforce, (2) Compiling a knowledge base of existing federally funded MGI-related efforts, (3) Linking physical- and cyber-infrastructure that cuts across materials classes and application domains, (4) Establishing working groups and networks in and across these domains, (5) Defining effective foundational engineering problems for each application domain to rally MGI stakeholder collaboration and networking, and (6) Establishing a distributed materials information infrastructure.

Building the Materials Data Infrastructure: A Materials Community Planning Workshop Report (2015)

Source: ASM International

URL: http://www.asminternational.org/documents/10192/19715738/Materials+Community+Planning+Workshop+Report_030515/3960c291-d2d3-446c-95d7-89740a57a233

Synopsis and Highlights: Utilizing input from workshops prior to this event, an analysis was conducted that addressed urgent needs and opportunities in the materials data initiative. These include methods for making data from published articles available, creation of a data registry, business models for sharing data, data quality standards and protocols, creation of a trained workforce, and a map of data relationships and dependencies. Workshop participants created a prioritized four-year timeline for future workshops that specifically addresses the abovementioned unmet materials community needs.

- *Data management:* establish a materials data quality roadmap and materials community data registry
- *Data sharing:* develop business models to encourage participation in infrastructure, identify connections between publishing articles and data
- *Education, Training, and Outreach:* develop data management workforce training

Additionally, recommendations for how to initiate a professional society-led materials data community were made. These included best practice sharing for making articles, reports, and data available, coordinating access to existing databases, developing standards and protocols for materials data, and supporting integrated workshops on MDI areas.

In-Process Materials Data for Modeling Workshop Report (2015)

Source: ASM International

URL: <http://www.asminternational.org/documents/10192/25764557/ASM+In-Process+Materials+Data+Workshop+Summary+2015-08.pdf/c5a6684b-91c3-4893-aefd-474622c98853>

Synopsis and Highlights: This workshop focused on defining opportunities for companies and organizations to collaborate in order to increase the affordability, accessibility, and availability of pedigreed data for modeling purposes. High-priority activities were defined as:

- *Business activities-* collaborative database must communicate the value of in-process materials data from an engineering perspective to create cost-competitive subscription models and convince industry, government, and additional organizations to participate
- *Information technology activities-* develop formats and schema that represent process pedigree and microstructural detail for each type of data, develop a technical steering group with computer science expertise to ensure data security and control of database maintenance
- *Cultural and regulatory activities-* define peer review process to assess quality of data and promote validated datasets with a sharing mechanism using membership or licensing agreements and also form a legal team to understand Export Administration Regulations and International Traffic in Arms Regulations as it relates to the collaborative

The ultimate purpose of this collaborative project is to generate and manage competitive, pedigreed, in-process manufacturing data to be used in modeling and simulation of materials.

Report of the DOE Workshop on Management, Analysis, and Visualization of Experimental and Observational Data: The Convergence of Data and Computing (2015)

Source: Department of Energy – Office of Science

URL: https://science.energy.gov/-/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf

Synopsis and Highlights: The primary focus of this workshop was an assessment of how data was acquired, analyzed, curated, stored, and shared by large experimental facilities and observatories funded by the DOE. Workshop participants remarked that the majority of facilities struggle with an exponential increase in data generation and that effective communication and collaboration is key, but there is insufficient infrastructure available.

Materials Data Analytics: A Path-Finding Workshop (2015)

Source: ASM International, The Ohio State University, National Institute of Standards and Technology

URL: <http://www.asminternational.org/documents/10192/25925847/ASM+MDA+Workshop+Report+Final.pdf/0e29644e-a439-4928-a07a-8718817a46e4>

Synopsis and Highlights: This workshop was closely aligned with the MGI in its aim to identify challenges, applications, and opportunities for the advancement of Materials Data Analytics (MDA). It convened approximately 30 experts from across academia, government, and industry to articulate the current state of the art and ultimately identify critical pathways and actions to facilitate MDA approaches. The top five challenges to advancing MDA were identified as follows:

- Understanding uncertainty in data and models
- A lack of data/knowledge sharing
- Complexity of multi-scale optimization
- Limited decision-support resources
- Extracting knowledge from literature based resources.

Modeling Across Scales: A Roadmapping Study for Connecting Materials Models and Simulations Across Length and Time Scales (2015)

Source: The Minerals, Metals & Materials Society

URL: <http://www.tms.org/multiscalestudy/>

Synopsis and Highlights: This study focused on bridging materials models and simulations across length and time scales, and addressing the corresponding community's challenges. This included consideration of fundamental linkage models, quantitative computational codes, and model verification and validation, which all involve the generation and handling of materials data. Several programmatic and technical recommendations were identified from the roadmapping study revolved around the collection, storage, and sharing of data. Selected examples include:

- Establishment of a data infrastructure for multiscale materials data
- Support of open data mandates for authors to upload data into repositories as part of a journal submission requirement
- Create data analysis tools for high-throughput methods

Rise of Data Workshop (2015)

Source: University of Minnesota, University of Maryland

URL: <http://riseofdata.org/>

Synopsis and Highlights: Rise of Data in Materials Research is an NSF-funded project led by Professor Ellad Tadmor and Professor Ryan Elliott at the University of Minnesota and Prof. Ichiro Takeuchi at the University of Maryland. An online interactive website is available for the materials community to be more fully engaged and includes a repository of presentations given at the June 2015 workshop. Six major themes discussed in the Rise of Data project include: Materials Cyberinfrastructure, Enabling Infrastructure Creation, Data Management and Handling, Knowledge from Data, Education of Materials Researchers, and Grassroots Standards and Government Support. Workshop participants created a list of recommendations for each theme described, with the idea that recommendations will evolve over time from community input.

CHiMaD Workshops (2016)

Source: Center for Hierarchical Materials Design (CHiMaD) at Northwestern University

URL: <http://chimad.northwestern.edu/news-events/ChiMaD%20Data%20and%20Database%20Efforts1.html>

Synopsis and Highlights: The Center for Hierarchical Materials Design (CHiMaD) at Northwestern University, supported by NIST, has hosted three recent workshops on various aspects of materials data infrastructure. These three major themes included: (1) database and discovery, (2) building an interoperable materials data infrastructure, and (3) materials data and analytics for materials research.



TMMIS

The Minerals, Metals & Materials Society

*Promoting the global science and engineering professions
concerned with minerals, metals and materials*