Building a CMMI Data Infrastructure

Workshop Organizers

Steven P. French
Dean and Professor of City and Regional Planning
Georgia Institute of Technology

Peter Voorhees
Professor of Material Science and Engineering
Northwestern University

sponsored by
Civil, Mechanical Manufacturing, Innovation Division
Engineering Directorate
National Science Foundation

February 6-7, 2017
Arlington, Virginia
# Table of Contents

Introduction ........................................................................................................................................... 3

Pre-workshop Survey ............................................................................................................................. 8

  Challenges to Establishing a Shared Data Infrastructure .................................................................. 8
  Data Available to the Repository from Participants .......................................................................... 10

Breakout Group Results ......................................................................................................................... 11

  Sustaining a Data Repository .............................................................................................................. 12
  Incentivizing Data Sharing .................................................................................................................. 13
  Innovative Data Creation and Fusion .................................................................................................. 16
  Metadata Schema and Vocabulary for Resource Discovery .............................................................. 17
  Using Data Management Plans and Existing Data Centers ............................................................... 18
  Alternative Repository Models ........................................................................................................... 19

Strategic Roadmap .................................................................................................................................. 20

References .................................................................................................................................................. 23

Appendix A. Workshop Agenda ............................................................................................................ 25

Appendix B. Workshop Participants ...................................................................................................... 28

Appendix C. Existing Data Repositories ............................................................................................... 31
Introduction

This report documents the results of the workshop on Building a CMMI Data Infrastructure held on February 6-7, 2017 in Arlington, VA. This workshop focused on how collecting and sharing data can improve the research capabilities of the disciplines represented within the Civil, Mechanical, Manufacturing, Innovation Division of the National Science Foundation. The workshop included engineers, material scientists, city planners, data scientists as well as publishers, librarians, and representatives of relevant professional organizations. A copy of the workshop agenda is included as Appendix A and a list of attendees is included as Appendix B.

The National Science Foundation (NSF) commissioned this workshop to develop a strategy for CMMI to realize the benefits by collecting, curating and sharing data to support its research community. A number of other disciplines, including astrophysics, earth sciences and health sciences, have realized significant benefits from sharing data within their the research communities. Sharing data within a research community has been found to lower research costs by reusing available data and to increase the rigor of scientific research through replication (Arzberger, et al., 2004). Sharing data can strengthen research collaborations, accelerate the pace of research and increase trust in the research enterprise. It can also create the extensive data resources needed to support machine learning and other analytic techniques that depend on large volumes of data. Establishing scientific data repositories to serve these communities in an era of rapidly expanding data represents a grand challenge to these research communities.

A primary conclusion of the workshop is that the CMMI should develop one or more data repositories to support its research communities. Developing one or more prototypes in the immediate future will provide the experience and expertise necessary to capitalize on the benefits of data sharing within these research communities.

Two substantive areas from the CMMI community, urban infrastructure systems and materials science, were the focus of the workshop. It should be noted that, while both of these domains
are a part of CMMI, they have some fundamental differences. Materials scientists primarily produce data on new materials and their properties through lab experiments and analytical modeling. As a result documentation of workflows and the details of laboratory procedures are especially important in this domain. While civil and mechanical engineers conduct some lab experiments on infrastructure systems (e.g. the Network for Earthquake Engineering Simulation), a vast amount of data that describes the characteristics and ongoing performance of urban infrastructure systems is collected and maintained by public agencies (e.g. state transportation agencies, local water and sewer authorities) or private companies (electric power utilities, telecom companies). Much of the research in this area requires getting access to previously collected data and addressing the privacy and security issues associated with using that data. Despite these basic differences, many of the issues identified were common between the two areas. In addition, while this workshop focused primarily on these two substantive areas, the strategies and approaches developed in this workshop should be useful to a wider set of disciplines.

As we enter the era of big data, there are unprecedented opportunities for deepening our understanding of new and previously unseen structures, relationships and behaviors across a wide variety of disciplines. Our world is awash in new data. However, the means and methods for acquiring and making this data available to support research are not well developed. This workshop highlighted the opportunity that CMMI has to be a leader in collecting, curating and sharing data to accelerate the pace of scientific discovery and knowledge creation. As we begin to think about a strategy to develop a data sharing approach, it is useful to think about data that is directly produced by NSF research projects as well that produced and captured by a variety of other public and private entities. Figure 1 shows data that are directly produced by NSF researchers and are, therefore, the easiest to assemble as compared to other kinds of desirable data that are produced and controlled by other entities.
In the first urban century, complex and, interdependent infrastructure systems have become critical to support human habitation. In the US and other advanced nations these systems are rapidly joining the Internet of Things (IoT) and evolving into cyber-physical systems (Gartner, 2013). As a result, more and more data are being produced on the performance of these systems under both normal conditions and under periods of severe stress caused by natural, technological and intentional hazard events. These systems interact dynamically with human activity through social, economic and political systems. Big data provides a unique opportunity to investigate the dynamic interactions between human and cyber-physical systems. To exploit these new opportunities, the NSF needs to develop a strategy to tap into this rich new source of data to support its research mission. Private corporations are launching similarly large-scale big data initiatives. For example, IBM is developing infrastructure for big data to be used in decision-making by many of their corporate clients (Perret, IBM, 2014).

From the semiconductors that comprise computer chips to the metals in jet turbine blades, advanced materials are central to nearly every technology in use today. It is now possible to design materials computationally using data derived from experiments and computer simulations. For example, new compounds can be discovered using large databases that are created through quantum mechanical calculations of the energy for formation of compounds.
Databases produced by rapid throughput experimental assessment of materials properties and compound formation are now commonplace. In addition, modern materials characterization tools, such as electron microscopes and X-ray beamlines, now routinely yield TB-sized datasets. These databases are then used to inform a computational design process that yields materials with desired properties in less time and cost than by the classical experiment-based Edisonian approach. The issues faced by the materials community parallel those of the infrastructure community: How can such large datasets be shared? How can they be explored? How can they be easily accessed?

In the last decade we have moved from an environment of data scarcity to one on data abundance, or perhaps even data overload. More and more data are being produced on the structure and function of urban areas and materials. This includes data that describe infrastructure networks and the interacting behavior of urban residents. The advent of high efficient density functional methods, rapid experimental synthesis, and state-of-the art instrumentation has led to a wealth of materials data that must be leveraged to design new materials that are desperately needed by industry. In both the infrastructure and materials areas, the data has high spatial resolution and is collected longitudinally over time, so it also has high temporal resolution. We have entered the era of "Big Urban and Materials data."

In the infrastructure area, over the past generation cities and counties have developed extensive base information in the form of traditional GIS and relational databases. These data sets include parcel-level land records, building inventories, infrastructure systems and street networks as well as detailed demographic information. In addition to this rich set of traditional data a wide variety of unstructured data is now being created from sensors, video cameras, drones and social media. Urban infrastructure systems and smart buildings are being monitored continuously by embedded monitoring instruments, mobile sensors and increasing by the cell phones of citizen sensors (Thakuriah, 2016). In addition, social media postings (Facebook, Twitter, Instagram, etc.), surveillance cameras, drones, cell phone location data, license plate readers, transit access cards and credit card transaction records provide a dynamic view of
human activity that can be connected with the performance of the city’s infrastructure systems and their performance (Hasan et al, 2013). While much of this new data is unstructured, most of it contains either a time stamp and/or geolocation data that allows it to be combined with more traditional structured data sets. Combining this rich and diverse amalgam of unstructured data with more traditional data can provide a more comprehensive, highly detailed, real time view of urban infrastructure systems and how they interact with the behavior of urban residents. This presents a unique opportunity to understand the dynamics of human-cyber-physical system interactions.

Similarly, over the past decade a number of materials databases have been established that stretch from those created by quantum mechanical calculations of materials properties, to compilations of data on material performance. Databases of images of material microstructures are readily available, as are data from experimental equipment, such as small angle scattering results from X-ray synchrotrons. Materials data is thus highly heterogeneous, and this heterogeneity poses a challenge to accessing and storing data as well as creating schema to guide potential users.

Unlike traditional urban data that was primarily created by state, federal and local government agencies, much of the new unstructured data is created and held by private entities. Materials data that are a result of government funded research must be made available. However, data created by companies are frequently not shared, because such data can often give the company a competitive advantage. Gaining access to this proprietary data is desirable, but may be difficult and expensive. Moreover, in the infrastructure area much of this data is so detailed that it creates significant privacy issues. How we address the ownership and privacy issues will be important if we hope to make infrastructure and materials big data an integral part of the research data landscape.

The astrophysics and structural biology communities now share data collected by instruments and individual investigators. It is routine for biologists to upload data on the structure of a
molecule when the paper describing the result is submitted for publication. Large astronomical databases, such as the Sloan Digital Sky Survey and the Hubble Space Telescope archive, show that data can be used by a large community of users who are not connected with the PI’s who acquired the original data, thus increasing the impact of these instruments. These communities have shown that widespread sharing of scientific data is possible and effective. Thus, it is clear that in order to realize the full potential impact of big data, it will be necessary to explore methods to incentivize the infrastructure and materials communities to share data more broadly.

Pre-workshop Survey
Prior to the workshop participants were asked to identify the major challenges to developing a data repository and were asked what kind of data they could provide to such a repository. Participant responses are summarized below.

Challenges to Establishing a Shared Data Infrastructure
The amount of data available on infrastructure systems, manufacturing, and materials science is rapidly growing, and the existence of a central data repository would greatly benefit the CMMI research community. However, there are many challenges to initiating, expanding, and maintaining such a data repository. Due to the breadth and depth of the data available for a repository, there are concerns regarding data discovery, accuracy and avoidance of duplication. Therefore, systematic maintenance and screening of data is necessary, as well as efforts to clean and validate the data, and maintain data currency in order to support data evolution on timescales much shorter than data retention periods within the repository.

Initiating a central data repository for the CMMI research community also necessitates taking advantage of existing, long-running repositories in an effort to reduce duplication of data that may already be available to the CMMI research community, and to improve interoperability between repositories. Prior to initiating a central data repository, existing data repositories will need to be examined in order to better understand their data standards and protocols for
archiving, linking, and generating metadata. It may also be necessary to look to existing data repositories for developers and ways to incentivize data suppliers and users.

A cultural change regarding data ownership and overcoming biases against sharing data in the CMMI research community will be crucial in attracting data suppliers and users of the repository. It is vital that the repository is able to gain recognition and citations in order to establish its credibility and draw more data suppliers and users to the repository. Preserving and ensuring confidentiality when necessary, for locational and identification purposes, will also be crucial when attracting data suppliers. Gaining access to proprietary sources of data, and overcoming data security limitations will also be necessary in order to provide users with unique data sources, which will be important in attracting and maintaining users.

Search and fusion techniques, such as cross-indexing heterogeneous data sources will need to be developed to ensure that the repository is user-friendly and is organized efficiently to support data discovery. Balancing access to raw data sources with desired analytics and varying computational needs will further ensure user satisfaction and attract users to the repository.

Although there are numerous challenges to initiating, expanding, and maintaining a data repository, it is an undertaking that should prove beneficial to the CMMI research community. Protocols for maintaining data accuracy and credibility will need to be established, and existing, long-running data repositories will likely need to be examined to ensure adherence to standards. Further development of ontologies to better use heterogeneous data sources will be necessary to improve efficiency of the repository and to attract and maintain users. These efforts will likely aid in shifting the perception of data sharing, thus making the data repository a vital part of the CMMI research community.
Data Available to the Repository from Participants

Participants were also asked to list the kinds of data they have that could be made available to a CMMI data repository. Below is a list of the data that participants indicated they currently have available and would be willing to contribute to a repository.

Survey and Census Data

- Household surveys from various parts of the country dealing with mitigation actions, disaster impacts, dislocation/displacement, and recovery issues
- Jurisdiction (counties and cities) surveys examining mitigation planning and mitigation/adaptation actions related to development and land-use regulations
- Travel survey and human activity pattern data
- British Geological Survey data

Social Media Data

- Twitter data on human behavior during disasters
- Twitter data on human behavior during fires
- Facebook data on users’ responses to various architectural designs
- Filtered social media data, e.g., Twitter data, verified to be relevant to disaster events and critical infrastructure damage

Materials Science Data

- Finite element models of structures
- 3D microstructures
- Grain Boundary Properties
- X-ray tomography of phase transformations in materials
- Phase diagrams, diffusion coefficients, thermal conductivity, phase precipitation kinetics

Data on Buildings

- BIM models
- Building controls data
- Building energy use data
- Smart Cities data for selected metro areas
- Detailed building inventory data sets for selected metro areas
- Simulated building and bridge responses to stochastic excitations, e.g., earthquakes
**Infrastructure Network Data**

- Infrastructure inventory databases
- Infrastructure network data
- Data on the physics governing network flows by infrastructure
- Infrastructure network node and edge data (facility locations and characteristics)

**Disaster and Natural Hazards Data**

- Input and output for testbed simulations of disaster recovery
- Experiment data on how people participated in the public-private partnership for disaster management
- Post-hazard bridge damage-functionality and recovery

**Transportation Data**

- Regional Travel Survey Data
- State registration and inspection records: Measurements and results from vehicle safety tests
- Vehicle emissions test results

This list provides an initial idea of the type of data that might be included in a CMMI data repository.

**Breakout Group Results**

Workshop participants were divided into five breakout groups and asked to address key issues related to developing a shared data repository. The results of those discussions are summarized below. The five breakout groups were:

1. Sustaining a Data Repository
2. Incentivizing Data Sharing
3. Innovative Data Creation and Fusion Technologies
4. Metadata Scheme and Vocabulary for Resource Discovery
5. Using Data Management Plans and Existing Data Centers
Sustaining a Data Repository

Building a sustainable data-sharing platform requires long-term strategic planning. It will also require the development of a sustainable business model that will enable the long-term operation and maintenance of the repository. The breakout group discussed three distinct business models to support data repositories:

1. Supported by a central organization
2. Users pay to access the data, and
3. Users pay to store their data in the repository.

Which of these models will be most successful for CMMI will depend on the needs of the research community. The Data Research Alliance has done a study of the revenue streams of a number of data repositories. Their report is available at [https://www.rd-alliance.org/final-report-income-streams-data-repositories.html](https://www.rd-alliance.org/final-report-income-streams-data-repositories.html)

Additionally, it is important to understand the business requirements of stakeholders to know what would they want and need from a data repository. These characteristics should be built into the repository from the start. This will require surveying the community and end-users as a part of the design process. This will help determine when and why researchers will be willing to share data. It is also important to clearly understand the value added proposition for users, both those who contribute data and those who use it. This will be especially important if there are fees associated with contributing or accessing the data.

The data repository should have a clearly defined scope. This scope will identify what data will be included in the repository as well as what data will not be included. Sustainability and incentives are closely related. There must be clear incentives for data sharing (e.g., credit for data citation, improved research performance, access to new data, etc). Principal investigators within CMMI primarily focus on their specific research area, and, therefore, it may be necessary to go beyond this core group to develop a functional and sustainable data repository.
There is wide diversity in data produced and consumed by the CMMI research community. One data model may not fit all the users. It may be best to consider a federation of data repositories that meet the needs of different research communities. Such a federation of repositories will spread the cost of maintaining the repository across different user groups and organizations. In addition, a federated set of repositories will be more flexible and allow for the creation of new databases as new fields emerge.

Attendees indicated that a successful repository requires a governance structure that will ensure long-term viability. In addition, the repository design must consider system reliability and on-going support. The user community must trust that the repository will be updated and maintained for the long term. The required NSF data management plans (DMP) may be a useful method to incentivize NSF-funded projects to contribute data to a shared repository.

In a broader sense, given that similar data access discussions are occurring across multiple research areas, there is an opportunity for multiple NSF programs (and non-NSF programs) to combine resources to launch a combined effort to develop a data repository or set of repositories.

Incentivizing Data Sharing
Sharing data is common in many research communities, yet in the disciplines within the CMMI this is not the case. As noted by Piwowar, Day, and Frdsma in their 2007 study on data sharing in the cancer clinical trial field, “Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available” (Piwowar et. al, 2007). The objective of this group was to identify mechanisms that can be used to make data sharing the norm in the community and to identify the benefits of data sharing that can appeal to individual researchers. The most advantageous approach is to reward investigators for sharing data as opposed to penalizing those that do not.

There are a number of reasons why the research community should support data sharing. These benefits include stronger collaborations, increased research productivity, enhanced
transparency and trust in the research enterprise, and improved citation metrics from published works (Dawes, 1996 and Ball et. al, 2004). Perhaps the most compelling is that data sharing can improve individual and group productivity. In most research projects a large amount of time and energy is expended in data collection, leaving limited time to analyze all the data that has been collected. Even when the data is fully analyzed, new hypotheses or approaches may emerge at a later date that can suggest new ways to analyze the existing data. Also, the larger amounts of data available by pooling data across many projects can enable machine learning and other advanced analysis techniques that can enhance research productivity and lead to new discoveries. Ball, Gavin and Brazma acknowledge there are challenges of sharing biological data, but use this field to illustrate these benefits, claiming that sharing data can improve productivity by providing a much larger dataset than is typically available to only one lab and to allow new methods and questions to be answered using an accumulated set of data (Ball et. al, 2004). In addition, by sharing data, they claim researchers can avoid duplicating efforts and increase their productivity. Once this is fully appreciated by the community, this increased productivity will be a particularly strong incentive to encourage data sharing.

Among researchers in the health field, sharing data is significantly related to an increase in citations (Piwowar et. al, 2007). Various factors may help explain why this correlation exists, but one that might explain it is the increased perception of the trustworthiness of data that is publicly available. These benefits of data sharing can appeal both to the larger community (transparency and trust) as well as to the individual researcher or group of researchers (increased citations).

In addition, to further incentivize sharing, researchers should be given credit for the data resources they make available by allowing citation of datasets, and by recognition or through awards from professional or academic societies. Peer recognition is strong incentive for data sharing.
Rules can also create a framework that requires data sharing. The field of genomics, generally considered to be a pioneer in the development of data sharing infrastructure, has identified continued challenges in data sharing. While funding bodies require data sharing in genomics research, development of rules and policies for how to protect study participants, sensitive data, and individual contributions continue to be challenging (Kaye, et al, 2009). Funding agencies can mandate data sharing as a condition of funding. However, these mandates would have to be monitored to insure compliance. Similarly, publishers can mandate that authors make their data available as a condition of publication. Ultimately, the research community has to agree on acceptable norms that govern behavior within the community.

This group identified the following list of incentives.

- Competitions/Challenges with prizes (bragging rights)
- Demonstration problems for the use of data - gives more citations
- Publicize examples of data sharing and showcase examples that demonstrate publishing good data will enhance research productivity
- Publishing data should provide a competitive advantage in securing future funding.
- Create funding supplements to support data sharing
- Make an e-lab book format that is easy to use.
- Provide data resources that are easy to access.
- Citations of various sorts to datasets such as: top cited, top downloaded, and top dependency.
- Agreeing upon ontologies and schemas for fields.
- In some cases, data of failed experiments can be as important as publishable data for creating accurate artificial intelligence (AI) models of a system.
• Data can be used to increase research productivity by using AI to suggest where new data is needed and can point the way to new experimental designs.

• Experiments as a service to provide needed data. However, a challenge will be to fund this as needed.

• Providing licenses for sharing, and explaining both data copyright and responding to frequently asked Q&A

• Establishing a data commons (e.g. NIH), and thereby incentivize putting data in the database.

• Showing that sharing data can improve the chances of getting a job, promotion, etc.

• By publishing metadata and thus encouraging people to contact an investigator for future collaborations

Innovative Data Creation and Fusion
Large-scale data fusion is creating a new way of doing science. Analytics is now a way to create data rather than just analyze data after it has been collected. One of the key benefits of combining large amounts of data from multiple sources is the ability to see new patterns and relationships that may not be apparent within a single project or data set. This is one of the basic premises of machine learning and other approaches to deriving insights from data. This is one of the most promising aspects of moving to a shared data model.

Historically, we have been studying critical infrastructure with imperfect data. Urban infrastructure systems are rapidly joining the Internet of Things (IoT) as instrumentation is added to transportation, water and sewer systems and to electric power grids. This system data can be combined with human behavior data drawn from social media to more closely monitor the patterns of urban metabolism. However, this data is often incomplete or privately held.
The nature of analytics has changed and fusion techniques are evolving to support data creation or generation that creates new data by linking formerly unrelated data streams. In terms of urban infrastructure we need to develop data fusion methods that can be applied to large-scale sensor networks and the Internet of Things.

When dealing with high resolution remote sensing, social media or travel data there are significant privacy concerns. We must develop better algorithms to prevent re-identification from linked data, for example of Vehicle Identification Numbers (VINs) from Department of Motor Vehicle data sets. We need to have the ability to link data, yet preserve the privacy of individuals and their vehicles. There is a need to formally describe technical anonymization implications of privacy (e.g., you cannot reverse engineer identity). Taking advantage of these new analytics requires a combination of domain knowledge and computational disciplines. We need to create multi-disciplinary approaches to studying privacy, including data scientists, social scientists, geospatial specialists and legal scholars working together.

**Metadata Schema and Vocabulary for Resource Discovery**

This group consisted mainly of materials scientists. Material scientists generate primary data through experiments or computational analysis. In this area it is important to integrate metadata into workflow. Therefore, a materials science repository must recognize that metadata is not static, but changes over the course of experiments and thus material scientists need tools that support continuous capture and editing of metadata. They need to alter the description of a dataset as the data evolves and integrate data curation into the lab workflow.

To capture various aspects of the workflow, material scientists need digital lab notebooks. They also need data dictionaries for the most commonly used terms and relationships. This requires clearly defined protocols and best practices for lab work. CMMI could collaborate with the Cyberinfrastructure Division to design and prototype the future digital of laboratory environment that would support automated data and metadata capture for material science and engineering.
Using Data Management Plans and Existing Data Centers

Data Management Plans can be a significant tool in attaining the goal of sharable and discoverable data for all CMMI programs. Data Management Plans (DMP) can be evaluated as part of the review of any new proposal. This would require specific criteria for evaluation versus simply a compliance approach. To be effective these criteria would need to be similar to those used for Intellectual Merit and Broader Impacts.

Considering data sharing as an evaluation criterion would require a significant culture change among reviewers and program managers. Past performance in data sharing from earlier projects could be evaluated as the publication of results is currently. Previous work should include a discussion of the dissemination of research data. Investigators with a strong record of accomplishment in broad data dissemination would be given credit as is currently the case for publications.

Alternatively, data sharing could be considered as a part of Broader Impacts rather than as a separate criterion. Currently, domain experts with little expertise in data sharing are writing and evaluating DMPs. More specific criteria are needed to evaluate DMPs and reviewers and investigators would need training to implement this culture change. To be effective NSF would need to develop a way to monitor and enforce DMPs.

To encourage data dissemination NSF should require proposals to specify funding to make the data produced by the project public. Investigators would need to be provided guidance on the costs associated with data sharing. NSF can also facilitate data sharing if it develops a data repository (or repositories). This would also lower the cost and increase the effectiveness of data sharing. NSF may also need to provide supplemental funding to make data public. Post grant awards like REU funding ($5-10K) or supplemental funds as part of the grant request from a pool within CMMI or the Engineering Directorate could be used to fund data sharing efforts.
Alternative Repository Models
The workshop presented several alternative models of existing data repositories. These included Earthcube. This repository supported by the NSF Geosciences Division contains a wide variety of geosciences data that is shared across the research community. The National Institute of Standards and Technology (NIST) has developed and supports an extensive materials data repository. (See https://materialsdata.nist.gov/dspace/xmlui/). The Urban Big Data Centre, funded by the U.K. Research Councils Economic and Social Research Council, is a consortium of 7 universities led by University of Glasgow, which operates a national data service by gathering, linking and disseminating a wide range of complex, heterogeneous and in some cases real-time datasets in transportation, housing, migration, education, the environment and other areas affecting urban futures (http://www.ubdc.ac.uk). Citrine is an innovative start-up company that collects materials data and applies machine learning to address materials questions for industrial customers. Participating researchers can access the data without cost, while revenues from industrial users support the repository.

Workshop participants were asked to consider four alternative repository models:

1. CMMI specific repository
2. Materials specific repository
3. Urban infrastructure repository, and
4. Federated repository that links a number of existing repositories.

While a CMMI data repository that includes all the disciplines within CMMI is theoretically feasible, the heterogeneity of the research community and variety of data types would make it difficult to build a coherent and easily searchable repository. Two separate repositories that focus on materials and urban infrastructure appear to be a better short-term objective. A larger, multi-agency federated repository is a more ambitious target, but would benefit greatly from the experience gained through building smaller, more focused repositories within CMMI.
Strategic Roadmap

The results of this workshop suggest that there are definite benefits to creating a data sharing repository to support the CMMI research community. The primary benefit of creating a repository will be increasing the efficiency of research by bringing data together in one place and providing discovery mechanisms to help researchers to find relevant data. The visibility of the shared resource will attract new data owners and stimulate innovations in science by linking and analyzing data that was not previously possible. By aggregating larger volumes of data in a repository, researchers will also be able to use innovative analysis techniques, such as machine learning, that are not possible with more limited amounts of data.

We suggest a phased approach to developing a robust data sharing strategy for the CMMI community. In the short term (next 12 months), we suggest CMMI fund a research project to develop a prototype data repository for one focused area. Such a project should include active researchers from the domains represented within the CMMI research community and data scientists who have expertise in developing repositories in other domains. The project would build on this workshop to explore more deeply the kinds of questions that the research community needs to answer. Based on that analysis the project team would develop a schema for the repository and a robust set of query and data fusion tools to assist users in finding and creating useful data to support their research questions. This initial attempt would be built upon data created by the CMMI research community. The DMPs for new projects should require that data produced by the project be deposited in the repository along with the standardized metadata and embedded in appropriate resource discovery mechanisms. This activity would be considered an integral part of each research project and be specifically budgeted. Current projects and those completed within the last three years would be able to apply for supplemental funding to prepare their data for inclusion in the repository. This prototype repository project would include an assessment phase that would look at the usage patterns of the repository and determine its strengths and weaknesses and develop a set of best practices to guide future repository efforts.
Based on the lessons learned from this initial prototyping exercise, the medium term (next 2-3 years) would develop two extensive repositories designed to serve the materials science community and the infrastructure and natural hazards communities. These repositories should be designed as federations that draw on existing repositories run by other agencies. In the materials science community this repository would include the NIST materials repository, the Air Force repository as well as existing repositories at UC Berkeley and Lawrence Berkeley National Lab, Duke, Northwestern and Harvard. In this type of federated repository the emphasis would be on developing ontologies that bridge the semantic differences in the schemas of the separate repositories. As discussed above, in this materials domain it is important to capture workflow information as part of the metadata. This would be a particular focus of this particular repository. In this same timeframe CMMI would develop a repository that specifically addresses the needs of the infrastructure and natural hazards communities. Again, a federated approach should be taken. This repository would combine the infrastructure data included in existing repositories, such as the Natural Hazards Engineering Research Infrastructure, the NIST Center for Risk-Based Community Resilience Planning and the DHS Homeland Infrastructure Foundation Level Data (HIFLD). Like any federated approaches the focus would be on ontologies to combine these independently developed repositories and develop a software platform to support data query, fusion and analysis in a way that is scalable over time and can appropriately handle data volume and heterogeneity, as well as the computational and data analytics needs of a variety of users. If possible, a team that includes domain experts as well as data scientists should develop both of these repositories. Joint funding with Computer and Information Science and Engineering Directorate (CISE) should be explored and possibilities of linkages or partnerships with overseas data services, such as the Urban Big data Centre in the UK or the Australian Urban Research Infrastructure Network, should be investigated.

In the long term, CMMI should look toward developing a large federated data repository with other NSF divisions, other federal agencies, state and local government agencies and private utilizes and companies. Such a massive, searchable database would open new avenues of
inquiry to investigators across a number of disciplines and significantly increase the speed and scope of scientific discovery.
References


Research Data Alliance. Open up! On the scientific and public benefits of data sharing. 2015.


Day 1

8:00 Registration

8:30 NSF Introduction and Welcome (Barry Johnson, Acting Assistant Director, ENG).

8:45 Workshop overview and charge (French, Voorhees)

9:15 Keynote: Mark Parsons

10:00 BREAK

10:00 Data Commons Demo and Explorations (20 min. talk)
- Materials Genome (Data-driven materials design, Jim Warren, NIST)
- Urban Informatics (Piyushimita Thakuriah, University of Glasgow)

11:00 Group Discussion

12:00 Working Lunch (Intro slide from participants/follow the mic)

1:30 Concurrent Breakout Sessions on Key Themes (Break into two sessions of one hour in length?)

- **Sustainability (Megan Clifford, lead):** A major impediment to achieving the goal of sharable and discoverable data is the long-term financial viability of the databases in which the data is stored. All too often a database exists for the time span of a grant, only to be disappear once the grant ends. The focus of this group is to explore and suggest strategies by which open databases can become financially sustainable. A fundamental question is whether such systems should be as open access or gateway systems. Also, of interest are models in other fields that can be used in the CMMI community.

- **Incentivizing data sharing (Surya Kalidindi, lead):** In many communities it is common to share data, yet in the disciplines within the CMMI this is not the case. Are there lessons from these other disciplines that can be used in the CMMI? The objective of this group is to identify mechanisms that can be used to make data sharing the norm in the community. Are there approaches that reward PI’s for sharing data (as opposed to penalizing those that don’t)?

- **Innovative Data Creation and Data Fusion Approaches (Piyushimita Thakuriah, lead):** Many data sets are not readily available due to proprietary, privacy or security reasons. This group will look at innovative ways to produce and collect data such as instrumentation, crowd sourcing, and synthetic data creation. How can the field make progress when they have no access to the data? Is it possible to create synthetic data sets of high fidelity? Easy data fusion or linkage from different sources is also an issue.
• **Metadata-Schema, Vocabulary Workflow Tool Development for Resource Discovery (Laura Bartolo, lead):** The perception amongst many investigators is that simply posting data to a server constitutes sharing. The reality is that if data is not adequately described and, equally as importantly, done so in a manner that is remotely searchable, posted data will go unused. The focus of this group is to suggest approaches to describing data in ways that are easily searchable. Methods to easily download data from databases or workflow tools will be identified.

• **Using Data Management Plans and existing NSF data centers (Walt Peacock, lead):** Can or should existing NSF data centers, or other data repositories, be used in this regard? Are such data centers a necessary condition for formulating a reasonable data management plan? How can data management plans be used to attain the goals of sharable and discoverable data for all CMMI programs?

3:30 BREAK while BG leads collect their thoughts and prepare reports

4:00 Summary reports from Breakout Groups

6:30 Steering Committee Working Dinner

**Day 2**

8:30: Charge for Day 2 (French, Voorhees)

8:45: Steering Committee synthesis from breakout groups: summarize results and solicit comments

9:15 Keynote: NSF’s Role in Public Access, Data Management and Data Infrastructure (Patricia Knezek, NSF)

9:45 BREAK

10:15 How to make the Repositories Sustainable?

Citrine - Bryce Meredig
Elsevier - Anita de Waard
AFRL - Chuck Ward

11:15 Implementation Breakout Groups

Breakout groups develop NSF-style abstracts strategies for data collection, curation and sharing.

12:00 Working Lunch in Breakout Groups

1:00 Alternative Solutions for Implementation - Reports from breakout groups

1:45 The Role of Data Management Plans (DMPs) Attendees will generate a checklist that reviewers could use in evaluating the DMP.

2:45-3:00 Closing Session
Appendix B. Workshop Participants
<table>
<thead>
<tr>
<th>Name</th>
<th>Email Address</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Georgia Tech Expenses</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Campbell, Carrie</td>
<td><a href="mailto:carelyn.campbell@nist.gov">carelyn.campbell@nist.gov</a></td>
<td>NIST</td>
</tr>
<tr>
<td>de Waard, Anita</td>
<td><a href="mailto:a.dewaard@elsevier.com">a.dewaard@elsevier.com</a></td>
<td>Elsevier (Publishing)</td>
</tr>
<tr>
<td>Foster, Ian</td>
<td><a href="mailto:foster@mcs.anl.gov">foster@mcs.anl.gov</a></td>
<td>U Chicago-Argonne Nat Lab</td>
</tr>
<tr>
<td>Gardoni, Paolo</td>
<td><a href="mailto:gardoni@illinois.edu">gardoni@illinois.edu</a></td>
<td>UIUC</td>
</tr>
<tr>
<td>Matthews, Scott</td>
<td><a href="mailto:hsv@cmu.edu">hsv@cmu.edu</a></td>
<td>Carnegie Mellon</td>
</tr>
<tr>
<td>Meredig, Bryce</td>
<td><a href="mailto:bryce@citrine.io">bryce@citrine.io</a></td>
<td>Citrine Informatics</td>
</tr>
<tr>
<td>Myers, Jim</td>
<td><a href="mailto:myersjd@umich.edu">myersjd@umich.edu</a></td>
<td>U of Michigan</td>
</tr>
<tr>
<td>Morovati, Meredith</td>
<td><a href="mailto:director@datadryad.org">director@datadryad.org</a></td>
<td>Dryad</td>
</tr>
<tr>
<td>Padgett, Jamie</td>
<td><a href="mailto:jamie.padgett@rice.edu">jamie.padgett@rice.edu</a></td>
<td>Rice</td>
</tr>
<tr>
<td>Pitera, Jed</td>
<td><a href="mailto:pitera@us.ibm.com">pitera@us.ibm.com</a></td>
<td>IBM</td>
</tr>
<tr>
<td>Shelden, Dennis</td>
<td><a href="mailto:dennis.shelden@design.gatech.edu">dennis.shelden@design.gatech.edu</a></td>
<td>Georgia Tech</td>
</tr>
<tr>
<td>Tien, Iris</td>
<td><a href="mailto:tien@ce.gatech.edu">tien@ce.gatech.edu</a></td>
<td>Georgia Tech</td>
</tr>
<tr>
<td>Ward, Charles</td>
<td><a href="mailto:charles.ward.4@us.army.mil">charles.ward.4@us.army.mil</a></td>
<td>Air Force Research Laboratory</td>
</tr>
<tr>
<td>Spanos, George</td>
<td><a href="mailto:gspanos@tms.org">gspanos@tms.org</a></td>
<td>TMS</td>
</tr>
<tr>
<td>Zhuang, Jun</td>
<td><a href="mailto:jzhuang@buffalo.edu">jzhuang@buffalo.edu</a></td>
<td>Univ at buffalo</td>
</tr>
<tr>
<td><strong>Northwestern Expenses</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bocchini, Paolo</td>
<td><a href="mailto:paolo.bocchini@lehigh.edu">paolo.bocchini@lehigh.edu</a></td>
<td>Lehigh</td>
</tr>
<tr>
<td>Runesha, Birali</td>
<td><a href="mailto:runesha@uchicago.edu">runesha@uchicago.edu</a></td>
<td>Chicago</td>
</tr>
<tr>
<td>Rohrer, Gregory</td>
<td><a href="mailto:gr2o@andrew.cmu.edu">gr2o@andrew.cmu.edu</a></td>
<td>Carnegie Mellon</td>
</tr>
<tr>
<td>Talukder, Ashit</td>
<td><a href="mailto:atalukde@unc.edu">atalukde@unc.edu</a></td>
<td>UNCC</td>
</tr>
<tr>
<td>Zhao, Ji-Cheng</td>
<td><a href="mailto:jzhao.199@osu.edu">jzhao.199@osu.edu</a></td>
<td>Ohio State</td>
</tr>
<tr>
<td>Budavari, Tamas</td>
<td><a href="mailto:budavari@jhu.edu">budavari@jhu.edu</a></td>
<td>Johns Hopkins</td>
</tr>
<tr>
<td>Toher, Cormac</td>
<td><a href="mailto:ch111@duke.edu">ch111@duke.edu</a></td>
<td>Duke</td>
</tr>
<tr>
<td>Chourasia, Amit</td>
<td><a href="mailto:amit@sandiego.edu">amit@sandiego.edu</a></td>
<td>UC San Diego</td>
</tr>
<tr>
<td>Shi, Xuan</td>
<td><a href="mailto:xuanshi@uark.edu">xuanshi@uark.edu</a></td>
<td>Univ of Arkansas</td>
</tr>
<tr>
<td>Macal, Charles</td>
<td>Macal, Charles M., @anl.gov</td>
<td>ANL/Univ of Chicago</td>
</tr>
<tr>
<td>Brison, Cate</td>
<td><a href="mailto:cbrinson@northwestern.edu">cbrinson@northwestern.edu</a></td>
<td>Northwestern</td>
</tr>
<tr>
<td><strong>Organizers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steve French</td>
<td><a href="mailto:steve.french@design.gatech.edu">steve.french@design.gatech.edu</a></td>
<td>Georgia Tech</td>
</tr>
<tr>
<td>Peter Voorhees</td>
<td><a href="mailto:p-voorhees@northwestern.edu">p-voorhees@northwestern.edu</a></td>
<td>Northwestern</td>
</tr>
<tr>
<td><strong>Steering Committee</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laura Bartolo</td>
<td><a href="mailto:laura.bartolo@northwestern.edu">laura.bartolo@northwestern.edu</a></td>
<td>Northwestern</td>
</tr>
<tr>
<td>Megan Clifford</td>
<td><a href="mailto:mclifford@anl.gov">mclifford@anl.gov</a></td>
<td>Argonne National Lab</td>
</tr>
<tr>
<td>Surya Kalidindi</td>
<td><a href="mailto:surya.kalidindi@anl.gov">surya.kalidindi@anl.gov</a></td>
<td>Argonne National Lab</td>
</tr>
<tr>
<td>Piyushmita Thakuriah</td>
<td><a href="mailto:Piyushmita.Thakuriah@glasgow.ac.uk">Piyushmita.Thakuriah@glasgow.ac.uk</a></td>
<td>University of Glasgow</td>
</tr>
<tr>
<td>Walt Peacock</td>
<td><a href="mailto:wapeacock@gmail.com">wapeacock@gmail.com</a></td>
<td>Texas A&amp;M</td>
</tr>
<tr>
<td>Name</td>
<td>Email</td>
<td>Agency</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Joanne Culberston</td>
<td><a href="mailto:jculbert@nsf.gov">jculbert@nsf.gov</a></td>
<td>NSF/CMMI</td>
</tr>
<tr>
<td>Bruce Howard</td>
<td><a href="mailto:howard@helix.nih.gov">howard@helix.nih.gov</a></td>
<td>NIH</td>
</tr>
<tr>
<td>Barry Johnson</td>
<td><a href="mailto:bwjohnso@nsf.gov">bwjohnso@nsf.gov</a></td>
<td>NSF/ENG</td>
</tr>
<tr>
<td>Patricia Knezek</td>
<td><a href="mailto:pknezek@nsf.gov">pknezek@nsf.gov</a></td>
<td>NSF/MPS/AST</td>
</tr>
<tr>
<td>Alexis Lewis</td>
<td><a href="mailto:alewis@nsf.gov">alewis@nsf.gov</a></td>
<td>NSF/CMMI</td>
</tr>
<tr>
<td>David Mendonca</td>
<td><a href="mailto:mendonca@nsf.gov">mendonca@nsf.gov</a></td>
<td>NSF/CMMI</td>
</tr>
<tr>
<td>Jack Meszaros</td>
<td><a href="mailto:jmeszar@nsf.gov">jmeszar@nsf.gov</a></td>
<td>NSF/OSTP</td>
</tr>
<tr>
<td>Siddiq Qidwai</td>
<td><a href="mailto:sqidwai@nsf.gov">sqidwai@nsf.gov</a></td>
<td>NSF/CMMI</td>
</tr>
<tr>
<td>John Schlüetter</td>
<td><a href="mailto:jschluet@nsf.gov">jschluet@nsf.gov</a></td>
<td>NSF/MPS/DMR</td>
</tr>
<tr>
<td>Steven Schmid</td>
<td><a href="mailto:sschmid@nsf.gov">sschmid@nsf.gov</a></td>
<td>NSF/CMMI</td>
</tr>
<tr>
<td>Greg Shannon</td>
<td><a href="mailto:gregory.shannon@hq.dhs.gov">gregory.shannon@hq.dhs.gov</a></td>
<td>DHS</td>
</tr>
<tr>
<td>Erin Walsh</td>
<td><a href="mailto:erin.walsh@hq.dhs.gov">erin.walsh@hq.dhs.gov</a></td>
<td>DHS</td>
</tr>
<tr>
<td>James Warren</td>
<td><a href="mailto:james.warren@nist.gov">james.warren@nist.gov</a></td>
<td>NIST</td>
</tr>
<tr>
<td>Eva Zanzerkia</td>
<td><a href="mailto:ezanzerk@nsf.gov">ezanzerk@nsf.gov</a></td>
<td>NSF/GEO</td>
</tr>
</tbody>
</table>
Appendix C. Existing Data Repositories
A good list of existing research data repositories can be found at

http://www.re3data.org/
**Large Project-based Repositories**

**UC Berkeley & LBNL**
- 58,123 compounds
- 41,952 band structures
- 1,243 elastic tensors
- Phase diagrams
- REST API

**Northwestern**
- 285,000+ compounds
- Crystal structure
- Phase diagrams
- python API

**Duke**
- 51,367 compounds
- Thermodynamic props
- Magnetic props
- Crystal structure
- Phase diagrams
- REST API

**Harvard**
- 2.3 million organic semiconductors
- Solar cell performance
- Principle energy levels

**OQMD:**
*An Open Quantum Materials Database*

**CEPDB – the Harvard Clean Energy Project Database**