Federal Register Notice: 89 FR 51554, Federal Register :: Networking and Information Technology Research and Development Request for Information on Digital Twins Research and Development, June 18, 2024.

Request for Information on the National Digital Twins R&D Strategic Plan

David Elbert

Response to Networking and Information Technology Research and Development **RFI on Digital Twins Research and Development,** posted in Federal Register vol. 89, no 118, June 18, 2024, for NITRD, NCO, and the National Science Foundation (NSF).

July 28, 2024

Person Filing: David Elbert
Faculty Research Scientist
Hopkins Extreme Materials Institute (HEMI)
[redacted]

Chief Data Officer NSF PARADIM Materials Innovation Platform
PI Data at the Speed of Extreme Materials Discovery (DSEMD) ARL HTMDEC Award
Co-I and Data Lead NASA STRI IMQCAM Digital Twin Project
PI NSF VariMat Data CI Pilot for Materials Data
Executive Council Materials Research Data Alliance (MaRDA
PI Lead NSF MaRCN FAIROS Research Coordination Network

Response Text Organized by RFI Sections

Overview: Digital twins offer a variety of high value returns but, to date, have been most effective in industrial settings where process optimization requires interaction between tools and supply flows that are well defined. More complex systems stretching from fundamental research to deployment of findings remain largely out-of-reach of current twin concepts and technologies. This difference is due to details outlined in the references included in the RFI, but primarily centers on gaps in accurate physics-based models, data-driven models, appropriate datasets, and lack of accessible infrastructure to produce accurate results or meet the high-rate needs of these systems. Recent advances in data infrastructure and AI/ML availability, however, make investment in digital twins of complex system timely as well as critical given investment by other nations.

In planning investments in digital twins, it is worth noting that the highest value digital twins will be in spaces undergoing rapid evolution in component modeling and data. This means that a central tenet of digital twin development must be a forward looking, composable approach that can absorb new ideas and tools that we cannot even anticipate. The rapid turnover in AI epitomized by LLM applications in just the last two years makes it clear that a

digital twin plan must be ready to embrace transformative tools as they arise. This also argues to prioritize creation of conceptual frameworks that can adopt developments from outside the digital twin work itself. Indeed, it shoud be expected that most fundamental advances in physical asset development and characterization; and most advances in data and computing systems will come from researchers working in non-aligned areas without regard for digital twins. This reality means that a successful digital twins program needs composable pluggability and will fail if locked into early technical choices or intransigent investigators. In this sense, digital twins should be conceptualized as "living systems" and may always be in a state of evolution. An important implication of this realization is that a central gap in creation of digital twins reflecting the science enterprise is the current dependence on stateless precepts (e.g. RESTful or GraphQL-based web services) and lack of recognition that a digital twin is most functional when conceived as stateful. The most powerful strategies to build a digital twin ecosystem and tool set will require significant advances in how the scientific community understands their own work. Complex systems, including research and development, are neither linear not stateless. Digital twins and eventual autonomous digital twins will require transformative change in perspective by many investigators creating critical pieces of the ecosystem. To compete globally and provide leadership across a digital twin ecosystem will require leadership-level investment in digital twin research and development that embraces and confronts the challenges of stateful architectures.

The following information on specific gaps and challenges is provided in the topical outline suggested in the RFI.

Artificial Intelligence (AI): AI has quickly emerged as a powerful tool for understanding complex systems and providing opportunity for transformative change in the practice of science and the control of autonomous systems. AI will clearly be a critical technology across the hierarchy of digital twins from virtual twins through truly autonomous twins. While many gaps exist, the central challenge for AI applications in digital twins is the rapid pace with which AI concepts and tools are evolving. Planning the digital twin ecosystem must focus less on specific types of AI tools and most critically on nimble integration of the breadth of AI tools and the inevitable emergence of new, transformative tools. To do this will require careful attention to AI/ML advancement that integrates with domain problems and that creates composable tools with clear protocols for advancement and deployment. Interoperability of AI/ML tools will require high-level incentives and sustained community collaboration to avoid creation of new barriers as we move to overcome those inherent in the current ecosystem.

A central challenge for AI components of digital twins is realistic understanding and mitigation of risks related to uncertainty in decisions and their implementations. This challenge is well known in autonomous systems where misinterpretation of new data or blind spots in models can cause deadly mistakes in systems such as autonomous vehicles and laboratories. In the same way, digital twins provide systematic decisions, deployment and guided production of materials and parts that may be used in mission-critical

applications. The mismatch in risk profile between traditional manufacturing, where a person routinely takes for granted the competence of a part in a care or plane, and the risk profile of a part validated by a digital twin that may be vulnerable to model or system deficiencies that have never been experienced before, makes focus on safety a priority that must temper the speed at which advances can be made using AI.

In a rapidly evolving, high-value space such as AI, there are also strong incentives to compete against other concept and tool developers. Digital twins leveraging AI will need to develop protocols to use AI as a service or share AI technologies in ways unseen in other research. Ultimately, top-down support for grassroots community efforts will be necessary to encourage community participation and to sustain efforts. The NSF has built a foundation for such effort with the inaugural cohort of 10 FAIR and Open Science Research Coordination Networks (https://new.nsf.gov/funding/opportunities/findable-accessible-interoperable-reusable-open/nsf22-553/solicitation). These FAIROS-RCNs have established a strong community base around data-centric work including many AI/ML and model-centric topics driven by shared data resources. These efforts should be sustained and coordinated resources provided to facilitate coordination that embraces AI development and tools across communities.

Business Case Analysis: Digital Twins that span the data lifecycle and technological readiness levels will facilitate high value return on data produced from basic research. The Materials Science domain is inherently interdisciplinary with ultimate drivers focused on delivery of products that need new materials capabilities. The Materials Genome Initiative (MGI) provides a natural partner for digital twin efforts focused on business cases with its focus on leveraging data-centric acceleration of new material discovery, design, *and deployment* critical to every societal Grand Challenge. MGI Challenges center on providing materials that play a central role in areas as broad as:

- Protecting and Improving Human Health
- Delivering Sustainable and Resilient Energy
- Thriving in Extreme Environments
- Enhancing Structural Performance
- Protecting the Environment
- Propelling the Information and Communications Technologies Revolution
- Advancing Critical and Emerging Technologies

Recently, MGI efforts have expanded in the realm of building success at higher technical readiness levels (TRL) meaning using data-driven techniques near the testing and deployment end of the materials development continuum. Research partners within the defense sphere (AFRL, ONR, and ARL) have inherent need to provide research outcomes that deliver outcomes that can be deployed in the field on compressed time scales. The new NASA Space Technology Research Institute for Model-based Qualification & Certification of Additive Manufacturing (IMQCAM; https://techport.nasa.gov/view/156318) specifically embraces a digital twin approach to product delivery. IMQCAM spans

university, research-institute, and industry partners to develop a digital twin provided certified and validated aerospace parts created by metal additive manufacturing. This effort combines multiple important digital twin themes with direct economic value aimed to leverage data and models to allow manufacturing of flyable parts without need for expensive physical testing and verification on Earth. The IMQCAM digital twin conception is an important touchstone, and may be unique in the world, for creating a truly deployable, autonomous twin spanning from fundamental research to manufacturing methods. IMQCAM leverages decades of NASA experience with twins and would be an ideal partner to lead broader efforts to ideate business case analyses.

Data: A central challenge for digital twins remains machine operable discovery, sharing, and interoperability of data. FAIR data principles are gaining traction in many fields but continue to need high-level motivation, advanced efforts, and sustained support. In the Materials Domain, the Materials Research Data Alliance (MaRDA) is a US National Materials Data Network as envisioned in Goal 1, Objective 2 of the MGI Strategic Plan (https://www.mgi.gov/sites/default/files/documents/MGI-2021-Strategic-Plan.pdf). MaRDA drives US and international efforts through a community-led network focused on connecting and integrating materials research data and data infrastructure to realize the promise of open, accessible, and interoperable materials data. Each of these elements are aligned with the goals of the Materials Genome Initiative (MGI) but a broadly congruent with concepts of FAIR data across disciplines and addresses a central gap to realization of broad application of digital twins. Without community data and without broad adherence to data sharing in interoperable, reusable ways, digital twins cannot be implemented for the types of big, complex problems that makes them valuable. Recent reports make it clear that a computational model is not a digital twin and that twins require actionable data across the domain of fundamental parameters encompassed by the twin. As a network that promotes the convergence of ideas, people, data, and tools to accelerate discovery, MaRDA provides an exemplar for other domains as well as a contact point to link efforts that are interdisciplinary. Such alliances are required to enable new insights into materials mechanisms and lay the foundation for both human-centered and artificial intelligence-assisted approaches to materials and product design. The structure of MaRDA includes governance by an elected council that facilitates a broad base of stakeholder interaction. This structure emphasizes that community work must be done at the community level, with mission-driven priorities, yet supported with top-down efforts to link the community with policymakers and high-level priorities. In many ways, the complexity of data infrastructure and community interaction mirrors the complexity of a functional digital twin; success of twins will require interoperability of the human infrastructure needed to develop the hardware, software, and protocols that will ultimately comprise digital twin environments. MaRDA and international partners are central to creation of interoperable data formats and semantic context required for functional reuse. A successful campaign to develop digital twins must include expanded work on data, metadata, and focused creation of foundational, sustainable data resources that serve digital twin development and not rely on current focus on absorbing data created for unaligned research directions.

Ecosystem: A persistent gap in development and deployment of digital twins centers on cost and lack of access to high-performance, managed data center resources. While compute resources are addressed in digital twin reports, data infrastructure is rarely recognized as a separate and critical realm. Few researchers understand that secure data infrastructure performant at the level of petabytes (and beyond) is well beyond the realm of university IT staffs or even traditional HPC centers. Digital twins will mirror the distributed nature of research itself and will need high performance network and storage backbones with secure, yet seamless access from researchers. Current high-performance data infrastructure is concentrated in commercial cloud options; some national laboratories (the planned DOE High Performance Data Hub is particularly notable); and rare university facilities such as the Bloomberg Data Center in the Institute for Data Intensive Engineering and Science at Johns Hopkins which already hosts multiple community datasets and will expand capacity to nearly 100 PB by the end of 2024. Such facilities will need to be sustained and should serve as guides to substantial growth in data infrastructure resources that are accessible to research budgets and require minimal advancement in expertise within the research teams themselves. A summit to convene expertise in data IT and infrastructure including these leading lights is recommended to develop realistic mission options and a roadmap for delivering consistent with the imminent needs of digital twin development.

International: All projects I participate in endorse international collaboration. This is important not only to accelerate progress, but to create compatibility between the entrepreneurial approach of US research with the centralized, hierarchical approach of many international projects. Data and models are without geographic boundaries and digital twin projects will benefit by sharing resources and concepts with international allies.

Long Term: The scale and ambitiousness of digital twins provide opportunity for the long-term, sustained investment needed to address three central challenges and related gaps.

First is the need to build sustained community efforts in data interoperability and reusability that leverage FAIR data principles to create a ecosystem of high-value, foundational data with a specific focus on experimental data and linked data semantics. These efforts can be interwoven by creating community inspired and defined datasets that are then instantiated using the advanced research infrastructure available at National Laboratories, the NSF MIPs, centralized user facilities at federally funded laboratories (such as the Cornell High Energy Synchrotron Source, CHESS, and the National High Field Strength Laboratory, MagLab), the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS), and US investments in high-speed networking through Internet2. With digital twin leadership at the table, planning and realizing these community dataset can be coordinated and accelerated.

Second is the need to create the high-speed, high-availability data centers and versatile data-stack tools required to provide the data resources required for instantiation of digital twins. Centralizing these resources will provide opportunities for economies of scale in the

data center while also avoiding unmeetable costs and incongruent approaches in the diverse efforts working across the digital twin ecosystem. Data resources are complex and should be constructed to avoid concept and technology lock-in. Twin advances should be able to find willing and resourced infrastructure partners that will accelerate their efforts rather than provide the all-too-common barriers found today. One suggestion would be to create an ecosystem of providers with most resources being high availability, high speed data stacks and clusters using modern object storage and economical storage such as that pioneered and deployed by the Open Storage Network (OSN) (https://www.openstoragenetwork.org/), while other resources provide nimble sandboxes to try new technologies as they emerge and meet the needs of research applications creating new challenges.

The third critical long-term need is development of bidirectional data flows that mirror the conceptual framework of a research enterprise comprising physical assets, synthesis of materials, modeling of structure-property relationships, physics-based modeling, data-driven modeling, automated data analysis, automated data curation, and AI/ML decision makers. The vision and foundation for such a framework is well known in computer science and autonomous systems research, but rarely appreciated in physical, chemical, and biological sciences. Digital twins require flexible, fast interaction between changing systems. An appropriate framework should be stateful with decoupling of resources though use of asynchronous subscription modes. Current science automation focuses on use of stateless, RESTful systems that are easy to implement and rely on local, modest performance computing. Alternatives are already arising in the Materials Domain and include the creation of nascent autonomy in the Open Materials Semantic Infrastructure (OpenMSI) project that has created tools to provide routine access to the Apache Kafka streaming infrastructure as a backbone for event-driven, loosely coupled infrastructure. OpenMSIStream (https://doi.org/10.21105/joss.04896) provides a robust adaptation of a standard streaming ecosystem to address challenges in laboratories producing large, diverse data volumes. Such efforts should be bolstered by reducing the barrier to stateful approaches by providing sustainable, economical brokering for mediation of data transfer between physical and virtual parts of the digital twin. New tools will arise and be adopted only when barriers such as these and seamless use are solved.

Regulatory: A valuable by-product of providing data infrastructure and streaming data tools is the ability to apply data governance. This will include the ability to assess and control data quality but will also provide ways to automate regulatory compliance. There remain challenges to establish a global regulatory framework that respects intellectual property in data, but we should not wait for clarity on governance before adopting an infrastructure that allows measurement and deployment of that governance.

Responsible: Responsible and ethical use of the data resources and models of a digital twin will be enhanced by the transparency created through contextual metadata and data streaming.

Standards: Standards and protocols are clearly critical, but the field is moving too quickly to wait for years of community agreement and mediation. The focus should be on transparency and function so research may move quickly, and interoperability can be assured. It is clear that AI/ML methods will provide new ways to create ontologies far faster than humans have proven able. Visibility of data and methods across disciplines and endeavors will be the rate determining step in development of tools to create standards or obviate their need in specific cases.

Sustainability: The persistent challenge of all great data and infrastructure advances is sustainability. The strong potential for commercialization of digital twins suggest that there are opportunities for partnerships with industry partners, but resources that are of fundamental importance to the national interest should receive support to be sustained. The framework for sustainability of data resources is well established in the community authored TRUST principles for digital repositories. These are identified as Transparency, Responsibility, User focus, Sustainability, and Technology and were reported in an article in Nature in 2020 (https://doi.org/10.1038/s41597-020-0486-7). These have been instantiated in detail for repositories and US research should adopt them through the Core, Trust, Seal digital repository certification (https://www.coretrustseal.org/). The community cannot be expected to share data and models if the hosts are not trustworthy.

VVUQ: Digital twins provide a unique opportunity to leverage data across the research lifecycle and prevent the need to recreate data for validation of deployed assets. To do this, VVUQ methods should be expanded to encompass datasets and be deployable across the instantiated digital twin.

Workforce: Workforce development will need continual attention but there are two central challenges. First will be training advanced computing and data techniques across the curriculum in domain fields. Second will be meeting salary competitiveness to bring skilled computer and data engineers into the digital twin enterprise within domain science. It currently is virtually impossible to recruit a postdoc or transition a masters student to the PhD at university, academic salary rates. A good engineering Postdoctoral Fellow earns $65-80,000 per year. My recent MS graduates in Data Science and Engineering receive offers at $120-150,000 per year. It is reasonable to think that the intellectual draw of a good project might induce someone to defer some income, but it is unrealistic to think that strong candidates will accept significant differences in pay.

Summary: In summary, digital twins create significant opportunities but are restrained by significant gaps and challenges. Recent workshops provide an excellent advance in understanding these gaps. With appropriate resources that meet the directions I have tried to highlight here, the US can lead digital twin development and application through the next decade.