Scientific Data Management Beyond Traditional Computing Boundaries

Patrick Widener, *Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*

Laura Biven, *Jefferson Laboratory, Newport News, VA, 23606, USA*

Ian T. Foster, *Argonne National Lab, Lemont, IL, 60439, USA; University of Chicago, Chicago, IL, 60637, USA*

Beth Plale, *Indiana University, Bloomington, IN, 47405, USA*

Sarp Oral, *Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*

Rafael Ferreira da Silva, *Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*

Abstract—Scientific data management is undergoing a fundamental transformation driven by the convergence of AI/ML workflows, distributed computing and storage environments, and exponential data growth. This paper examines eight key developments reshaping research data management: fluid data movement, user-centric storage semantics, distributed storage solutions, integration into compute facility infrastructure, active preservation systems, enhanced data protection and control mechanisms, AI data readiness, and data and workflow provenance. We analyze how these developments address current limitations while enabling new capabilities for cross-facility collaboration and AI-driven research. Our analysis provides insights for research facilities and funding agencies working to modernize scientific data infrastructure while maintaining security, reproducibility, and accessibility.

The management of scientific data is undergoing a fundamental transformation, driven by exponential growth in data volumes and increasingly distributed research environments. As highlighted in recent federal frameworks such as the National Science and Technology Council's guidance on research infrastructure [1], [2], organizations face mounting challenges in coordinating data infrastructure across facilities while allowing seamless access, analysis, and preservation of research output. These challenges are compounded by the emergence of new data-intensive paradigms such as artificial intelligence (AI) / machine learning (ML) workflows, real-time analysis requirements, increased automation, and cross-facility collaborations that generate unprecedented volumes of experimental and observational data [3]. The convergence of edge computing, cloud platforms, and traditional research infrastructure adds additional complexity to an already challenging landscape, requiring new approaches to data management that can adapt to rapidly evolving scientific needs while maintaining security, reproducibility, and accessibility. This evolving landscape requires a critical examination of current data management approaches and their limitations.

Traditional data management paradigms, built on linear progression through creation, analysis, and archival stages, are becoming insufficient for modern scientific workflows [4]. This insufficiency is due to the emergence of multiscale workflows: from in situ HPC simulations and data analysis; local data acquisition and steering; and distributed workflows for time-critical analysis or distributed learning, for example. Other sources of complexity include the increase in automatization resulting in challenges for transparency and trustworthy results; and the need for robustness and sustainability across distributed infrastructures,

XXXX-XXX © 2025 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725; by Jefferson Science Associates, LLC under contract DE-AC05-06OR23177; and by UChicago Argonne, LLC, under contract DE-AC02-06CH11357—all with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

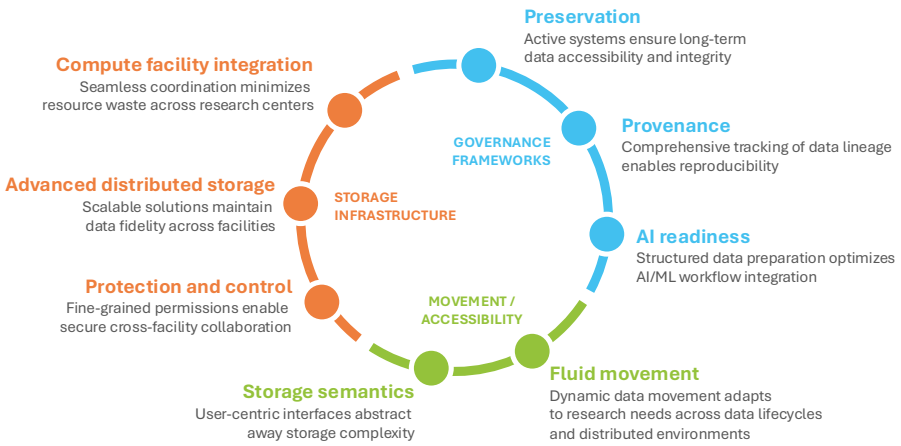

FIGURE 1. Key components transforming scientific data management, organized into three fundamental domains: Movement/Accessibility, Storage Infrastructure, and Governance Frameworks. Each component represents a critical area of development in modern research data systems.

multiple organizational domains, and a wide range of timescales [5], [6]. Traditional approaches that treat data lifecycle stages as distinct phases with clear boundaries are particularly challenged by the requirements of distributed computing environments, where data may need to be accessible simultaneously for multiple purposes across different locations and computing platforms. Furthermore, the growing importance of near real-time analysis, experimental steering, and collaborative research demands data management systems that can support more flexible and dynamic data flows while maintaining the strict requirements for data provenance, security, and long-term preservation that are essential for scientific research.

To address these challenges, this paper examines key developments transforming scientific data management across the computing continuum. These developments span three main areas: data movement and accessibility, storage infrastructure, and governance frameworks. Drawing from recent literature and expert insights, we identify eight transformative developments reshaping how scientific communities handle data, with a particular focus on the convergence of traditional approaches with emerging distributed computing paradigms (Figure 1):

- ***Data Fluidity Across Lifecycle Stages*:** Moving beyond the traditional three-stage lifecycle (creation, working, archival) to enable more dynamic and flexible data progression across distributed computing environments, especially for AI-driven laboratories and cross-facility collaboration.
- ***User-centric Data Storage Semantics*:** Emphasizing access to data through semantics that are aligned with user needs (e.g., data consistency, granularity, and access control) instead of storage system characteristics.
- ***Advanced Distributed Storage Solutions*:** Development of new storage paradigms to handle datasets too large for local storage while maintaining "keep-every-bit" requirements for ultra-high-fidelity analysis and reproducibility across distributed facilities.
- ***Integration of Data Management with Compute Facilities*:** Evolution of high-end computing facilities to support seamless data organization, management, and access across experimental, observational, and computing facilities, minimizing network latency and resource waste.
- ***Active Preservation Systems*:** Reimagining "dead" archival storage to address accounting and reproducibility challenges, suggesting a more dynamic approach to long-term data preservation and access.
- ***Data Protection and Control*:** Support collaborative data environments with rich access control semantics and shared governance models, without which important problem areas will lag behind and data sharing will be hindered.
- ***AI Data Readiness*:** Where once the problem was finding enough data to make AI-driven processing viable, the next challenge will be generalized approaches to making data ready to use in model training and other tasks.
- ***Data and Workflow Provenance*:** Enabling reproducible and trustworthy scientific results through comprehensive tracking of data origins, trans-

forms, and workflow execution details, while supporting performance optimization through standardized provenance capture and organization.

FLUID DATA MOVEMENT

Scientific data management has traditionally proceeded linearly through distinct lifecycle stages: data creation (from modeling, simulation codes, and scientific instruments), working data (optimized for tasks like AI training and visualization), and archival storage (in write-once repositories). However, the emergence of AI-driven laboratories, cross-facility collaborations, distributed computing environments, and multiscale workflows fundamentally challenges this linear model, as modern scientific workflows require data to be accessible simultaneously across multiple locations and computing platforms. The traditional segmented approach, while historically effective, is proving insufficient for modern scientific computing's need for dynamic and fluid data movement.

Several challenges are driving the need for more fluid data-movement approaches. First, as data set sizes grow, it's increasingly likely that different logically connected data items will exist in different parts of the data lifecycle; preserving logical connections between data items as they evolve will be difficult given the current state of the art. Second, the sheer volume of data being generated means that some datasets are too large (or cannot be reduced effectively) to copy to local storage, yet must still meet requirements for high-fidelity analysis and reproducibility. Third, new AI/ML workflow disrupts traditional assumptions about data locality and access patterns. Fourth, cross-facility collaborations require data to be accessible across institutional boundaries while maintaining security and performance. In addition, the convergence of edge computing, cloud platforms, and traditional research infrastructure adds complexity to data movement and staging decisions.

Looking ahead, several new approaches show promise for helping to address these challenges. Advanced data orchestration systems could automatically optimize data placement and movement based on workflow requirements and resource availability. New distributed caching architectures could help bridge performance gaps between facilities while minimizing unnecessary data transfers. Furthermore, semantic-aware data management systems could help abstract away the complexities of physical data location and movement while preserving essential properties such as reproducibility and provenance tracking. New AI-based approaches suggest the opportunity for semantic indexing of large quantities of diverse, distributed data, in ways that may greatly enhance scientists' ability to discover, integrate, and exploit previously unrelated data sources.

Predictions:

- Data management systems will evolve to support continuous data progression across lifecycle stages, automatically optimizing placement and replication based on access patterns and computational needs.
- New distributed caching and staging mechanisms will emerge to efficiently handle data too large for local storage while maintaining reproducibility requirements.
- Cross-facility data orchestration systems will automate the movement and positioning of data to minimize latency and resource waste while preserving security and access controls.

RETHINKING STORAGE SEMANTICS

Semantics capture interaction and use of data storage systems and thus can play a crucial role in the future for how research data infrastructure systems serve their users' needs. Historically, data access patterns and capabilities have been limited by the underlying physical storage systems and their characteristics. However, modern scientific applications and workflows increasingly require more flexible and user-centric approaches that prioritize how researchers interact with and utilize the data, rather than being constrained by storage system limitations. Semantic systems also present opportunities for alignment with FAIR principles for enhanced reusability of data and results.

Current state-of-the-art approaches are evolving toward a more sophisticated model that separates storage implementation from data access patterns and semantic understanding. Modern data management systems are implementing semantic layers that enrich metadata, standardize data characterization, and enable AI-ready accessibility while abstracting away the underlying storage complexities. These systems emphasize pre-model explainability, ethical data handling, and computability, enabling researchers to focus on their analytical needs rather than storage mechanics. Advanced semantic modeling approaches are being developed that can handle heterogeneous data sources through standardized metadata schemas, provenance tracking, and machine-readable documentation. This evolution reflects a broader shift toward

semantic data management that prioritizes research workflow needs while maintaining scientific rigor and reproducibility.

Several key challenges remain in fully decoupling data semantics from storage modality. First, legacy scientific applications and workflows often have deep dependencies on traditional filesystem interfaces, making it difficult to transition to new data models. Second, achieving consistent performance and access patterns across heterogeneous storage systems while maintaining security and access controls is technically complex. Third, the diversity of research workflows and access requirements makes it challenging to design abstract interfaces that can efficiently support all use cases while hiding storage complexity.

Looking ahead, several approaches show promise for addressing these challenges. Developing standardized APIs and interface protocols that abstract storage details while preserving essential semantic capabilities would enable broader interoperability. Implementing intelligent data placement and caching strategies that automatically optimize for different access patterns could help bridge performance gaps. Furthermore, fostering collaboration between operators of research data infrastructure to establish common frameworks for semantic data access could accelerate the adoption of storage-independent approaches.

Predictions:

- Research infrastructure will widely adopt semantic layers that abstract storage implementation details while preserving domain-specific metadata and access patterns.
- Data management systems will implement AI-ready semantic frameworks that automatically enrich metadata, track provenance, and enable explainable analysis across heterogeneous storage systems.
- Storage systems will evolve to decouple user-facing data semantics from physical storage characteristics through standardized APIs and automated optimization of access patterns.

NEXT-GEN DISTRIBUTED STORAGE

The landscape of scientific data storage is undergoing a fundamental transformation, driven by the convergence of traditional HPC workloads with emerging AI/ML applications. While solid-state storage has largely replaced rotating disks in computational science data centers, modern workflows spanning from edge devices to leadership-class facilities present challenges beyond hardware transitions. Current storage systems primarily optimize for either traditional scientific workloads (with sustained write bandwidth and sequential access patterns) or newer AI/ML workloads (featuring small, random accesses with less data locality), leading to infrastructures that struggle to efficiently serve both simultaneously. Additionally, existing approaches to data movement and staging based on copying entire datasets to local storage are becoming impractical as dataset sizes grow exponentially.

Several critical challenges must be addressed in next-generation storage systems. First, application developers face opaque trade-offs between different storage hardware, protocols, and performance characteristics, leading to suboptimal resource utilization. Second, the need for distributed, shared storage facilities, such as for the DOE High Performance Data Facility (HPDF), introduces new complexities in data movement and access patterns through its hub-and-spoke model. Third, the traditional POSIX file system model, originally designed for locally attached storage, has become a limiting factor in distributed scientific environments. These challenges are further complicated by the need to maintain distinct handling of metadata and data, which have fundamentally different consistency requirements and update patterns.

Looking ahead, several approaches show promise for addressing these challenges. Advanced storage systems could leverage AI-driven profiling and inference to automatically optimize data placement and access patterns. New storage abstractions could better separate metadata operations from data operations, allowing each to be optimized independently. Additionally, the integration of previously “lookaside" services (such as databases and indexing systems) directly into the storage fabric could enable more efficient metadata management and query capabilities.

Predictions:

- Storage systems will incorporate AI-driven modeling to dynamically optimize data placement across heterogeneous resources based on application patterns and needs.
- New storage abstractions will separate metadata and data paths, enabling independent optimization while maintaining consistency.
- Previously separate services (databases, column stores, graph stores) will integrate into primary storage systems, enabling unified data management.

COMPUTE FACILITY INTEGRATION

The integration of computing facilities has evolved from a focus primarily on computational performance to addressing challenges of data management across distributed scientific infrastructures. Modern research increasingly requires coordinated data flows among experimental facilities, observational platforms, analysis systems, and archive storage. Major research facilities have moved beyond treating data storage as simply an auxiliary service, developing comprehensive data ecosystems spanning the entire data lifecycle. The Department of Energy's Integrated Research Infrastructure (IRI) program exemplifies this evolution by seeking to create seamless integration across experimental, observational, and computational resources, which requires coordinated data management across distributed facilities while maintaining local autonomy.

However, significant challenges remain to achieve truly integrated data management across facilities. These include difficulties in maintaining consistent data representations and metadata standards across different domains, complexities in tracking data provenance through distributed multi-scale workflows, and challenges in implementing unified data governance policies across institutional boundaries. The dynamic nature of modern scientific workflows also creates tension between the need for flexible data access and the requirements for secure and efficient resource utilization. Furthermore, the exponential growth in data volumes continues to stress existing infrastructure and management approaches.

Several promising directions could advance the integration of data management across computing facilities in the near term. At a technical level, developing standardized approaches to data organization and cataloging would improve interoperability while reducing operational complexity. The adoption of container technologies and standardized workflow descriptions would improve portability, while automated data lifecycle management capabilities could optimize resource utilization and ensure appropriate long-term preservation. Common reference architectures and integration patterns, developed through coordinated multi-agency efforts, could significantly reduce implementation complexity. On the governance side, establishing shared frameworks for data access control and management policies would enable more effective cross-facility collaboration while maintaining security requirements. Additionally, implementing automated service discovery and orchestration capabilities would allow for more dynamic resource utilization, while coordinated approaches to data placement and movement across facilities could substantially improve efficiency.

Predictions:

- Automated orchestration systems that optimize data placement and movement while maintaining provenance across distributed resources will be implemented at multiple scales, including at Leadership Computing Facilities.
- Facilities will adopt standardized interfaces and protocols for seamless data flows, while preserving local operational autonomy and reducing infrastructure redundancy.
- AI-driven management systems will automate resource discovery and workflow optimization across institutional boundaries while enforcing security policies and governance requirements.

DATA PRESERVATION SYSTEMS

Traditional approaches to data preservation have treated archived data as dormant and infrequently accessed, a perspective reinforced by cloud storage tiers that prioritize cost savings over accessibility. Current systems primarily organize data around coarse-grain storage and retrieval of large-data objects, with limited support for rich metadata or sophisticated query capabilities. However, the emergence of AI/ML workflows, reproducibility requirements, and long-term preservation needs fundamentally challenge this paradigm, as modern scientific computing demands preservation systems that can support dynamic access patterns while maintaining data integrity and provenance over extended timeframes. These traditional approaches are increasingly misaligned with modern scientific workflows, particularly for AI model training and computational reproducibility.

Modern data preservation systems face several critical challenges. First, the reproducibility requirements of scientific workflows can require the indefinite preservation of not just results, but also input data and computational environments. Second, the frequent coupling of funding for data preservation to time-limited project grants creates sustainability challenges when project funding ends. Third, the rise of cross-facility collaborations and "born shared" data environments requires either new approaches to cost recovery and resource allocation for organic shared archival storage and sustainability, or protocols through which these environments can interoperate with a "bring-your-own" data preservation approach. Additionally, AI/ML work-

flows often need to access historical datasets in ways that do not align with traditional archival storage optimization strategies.

Looking ahead, several promising approaches could help address these challenges. Advanced preservation systems could incorporate richer metadata and indexing capabilities to support more sophisticated query and retrieval patterns. New funding and operational models could better align long-term data preservation, curation, and sustainability practices with institutional missions and resources. Furthermore, the development of more flexible cost recovery mechanisms could enable sustainable sharing of archived data across institutional boundaries.

Predictions:

- Preservation systems will evolve from passive storage to active digital preservation environments, incorporating automated verification, format migration, and accessibility features to ensure long-term data usability and scientific reproducibility.
- New organizational and funding models will emerge to manage “perpetual storage" requirements, including standardized approaches to transition preservation responsibilities as projects end and automated cost recovery mechanisms for shared preservation resources.

DATA PROTECTION AND CONTROL

Much research into collaborative computational science has assumed a relatively benign data protection environment. That is, data sharing has been presumed to be non-adversarial, with good intentions and good-faith interactions the default. Placing concerns of data protection and access control outside the scope of research into scientific data management has permitted significant advances in the state of the art. However, these concerns can not be set aside indefinitely.

The rapid development of data-driven techniques such as AI/ML inference has generated significant interest in applying state-of-the-art solutions to problem areas where data protection and access control cannot simply be ignored. Use cases spanning institutions such as those envisioned by the National Science Data Fabric (NSDF) [7] or the DOE High Performance Data Facility (HPDF) [8] will need to address the challenges of sharing across different administrative boundaries and user identity providers. Additionally, significant data handling occurs in environments where access constraints driven by regulation (e.g. for human subjects research), law (e.g HIPAA), or national security concerns are paramount.

A variety of data access controls will be needed to realize the full potential of wide-area infrastructure such as NSDF, HPDF, and various Integrated Research Infrastructure-style workflows. Variations include granularity of access; role-based and permissions-based access; and functional access such as with data visitation or the Five Safes framework [9]. Absent these services, the current state of data protection tooling provided by commonly available filesystems and data storage infrastructure will be insufficient. Also, scientific data management concerns also arise in high-consequence computational environments such as NNSA stockpile stewardship computing; it is unlikely that computing in these environments will benefit from the rapidly advancing state of the art in low-sensitivity scientific data management without direct research into verifiable data security mechanisms.

Looking ahead, several promising approaches could help address these challenges. Harmonizing or brokering authentication services through trusted networks is a first step toward interoperability in controlled access environments [10]. Frameworks will need to be flexible to facilitate a variety of access paradigms from role-based, permissions-based, functional access, or query-based access for a variety of data subsets.

Predictions:

- Given that data handling in both loosely and tightly controlled data environments requires similar scientific data management, innovation in data protection approaches will make it practical to apply similar techniques.
- Difficulties in implementing data management across administrative boundaries will result in convergence on single user identity providers or identity interoperability approaches.
- Data access controls will support a variety of access models for coarse and fine-grained objects, role-based and permissions-based access, and function-based access.

AI DATA READINESS

The appetite for data displayed by current AI model training approaches seems practically limitless. To take advantage of the power of AI-driven computational science both sustainably and scalably, the enormous amounts of scientific data currently generated,

collected and stored must be made *AI-ready* [11]. Leadership-scale AI readiness implies data preparation, training protocol design, and conversion to storage formats appropriate for use in high-performance parallel I/O. Data preparation may involve cleaning (for example, handling missing variables), automatic or manual labeling, and some degree of feature engineering to select and possibly alter a particular set of features. Training, testing, and validation splits in a data set are then determined according to the training protocol and exported in formats appropriate to the platform on which the training is to be performed.

Several challenges must be addressed in order to fully leverage scientific data in AI processes. Much data is generated by simulations or collectors without regard to the quality needs of model training. Insufficient quality data can lead to overfitting and other issues which can result in models poorly suited to their planned tasks. Data sizes from multiple research domains are growing, in some cases limiting the platform environments in which training can be performed. Different domains have adopted different approaches to model training workflows, complicating cross-cutting paths to AI data readiness.

Predictions:

- Automated data quality assessment systems will emerge to evaluate and score datasets for AI readiness across multiple dimensions including completeness, consistency, and feature richness.
- New workflow frameworks will standardize the preparation of scientific data for AI, including automated cleaning, labeling, and format conversion optimized for high-performance training.
- Cross-domain data preparation patterns will be established to enable transfer learning and model reuse across scientific disciplines while maintaining domain-specific context.

DATA & WORKFLOW PROVENANCE

The provenance of a data object is its overall history, including both its origins and transformations applied to it over time. Similarly, workflow provenance is the record of the design, execution, and outputs of a workflow. Together, data and workflow provenance can provide sufficient context and transparency for reproducible, trustworthy, and reusable scientific results. They can also address performance questions such as sources of latency, energy consumption, and self-consistency. Provenance at a mid-point in a workflow can help inform down stream data management and the advances outlined above, including abstracting data management to apply to semantic controls, automating workflow decisions, and managing access. Leveraging provenance information is currently challenged by a lack of organization and standardization. Provenance is currently captured, if at all, in a variety of platforms such as task logs, lab notebooks, data management plans, data sheets, and publications.

Looking ahead, we anticipate several promising approaches helping to capitalize on the opportunities of provenance. Community-based efforts in reproducibility and reusability can help to identify pieces of provenance information that are essential for the science goals. Standards for structuring provenance information as metadata will enable it to be used within workflows and for reuse. Instrumented workflows can autonomously capture performance-level information, while interactive tools can help to assimilate user information into the provenance chain [12]. Research methodologies need to evolve to capture provenance at every research stage, not just at data publication [13].

Predictions:

- Increasing provenance information will become standardized and organized as metadata, as needed by the user communities.
- Workflows will be instrumented to collect provenance information, and interactive tools will be developed to capture user input as provenance.

MATURITY AND EXPECTATIONS

We have identified several key areas where scientific data management must evolve to address the changing requirements of computational science. These areas vary in maturity and face different challenges for implementation and widespread adoption (Figure 2):

› **Established.** Distributed storage systems represent a highly developed area, backed by decades of academic and industry research that has addressed the technical challenges of wide-area data distribution and caching. Access semantics similarly demonstrates significant maturity, with object storage and key-value interfaces now widely implemented across the industry (frequently offered by vendors as complementary options alongside traditional POSIX-compliant parallel filesystem interfaces).

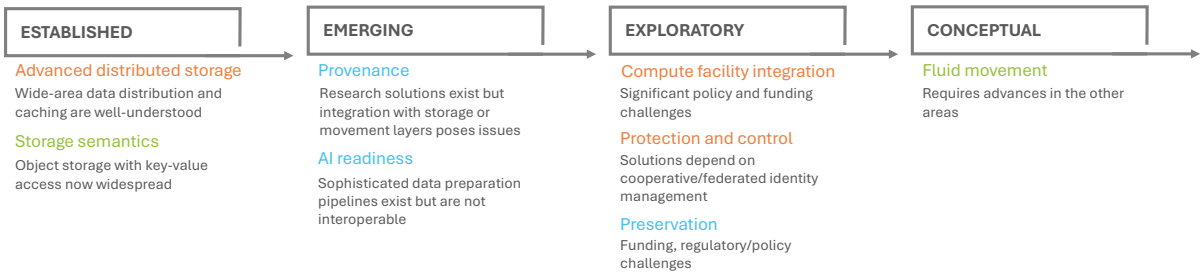

FIGURE 2. Our previously identified key components transforming scientific data management, with indications of the maturity of research efforts in each area with respect to enabling the next generation of scientific data management.

- **Emerging.** Substantial research exists on preserving provenance information in data-intensive computing environments [14], [15], but the main challenge lies in sustainably integrating this research with existing data storage and movement layers. These established systems often carry significant technical debt, creating implementation and interoperability issues. Regarding AI readiness, while tools for data preparation are abundant and generally well-integrated with common AI programming frameworks (primarily Python-based), the landscape remains fragmented. Almost every researcher has developed unique solutions to prepare data for AI processing. The key maturity challenge is not in figuring out how to make data AI ready, but rather extracting from the many existing ad hoc processing pipelines the common patterns and techniques for a number of different domain science areas.
- **Exploratory.** The remaining prediction areas we have identified are at a distinctly lower maturity level. These areas face significant "external" challenges including policy restrictions, regulatory requirements, and funding limitations. Facility integration efforts struggle particularly with the funding justification for infrastructure work aimed at deep cross-facility integration. This integration requires navigating the diverse administrative and operational structures that vary between facilities in both subtle and profound ways. Current efforts such as [16], [17] have proceeded incrementally by defining APIs available outside the facility network. Similarly, fine-grained access control faces substantial impediments; local identity management systems are typically deeply embedded, and comprehensive solutions likely depend on the development of more sophisticated federated identity approaches. Data preservation initiatives face dual challenges [18]: funding uncertainties (specifically, who bears the cost of data storage after projects are completed) and complex regulatory requirements (such as the 2022 Office of Science and Technology Policy guidance memo [19], which mandates specific retention and access protocols for data generated using US government funding).
- **Conceptual.** Data fluidity represents the integration of currently siloed data access techniques across the entire data lifecycle; it remains primarily in the research domain, distinguishing it from the more operationally-focused areas discussed elsewhere in this paper.

CONCLUSION

The transformation of scientific data management is driven by unprecedented challenges in data volume and complexity, changes in science workloads, and requirements for highly distributed computing. Our analysis of key developments identifies critical paths forward for the scientific computing community. These developments point to the future of more dynamic, intelligent, and integrated data management systems that include support for AI-driven orchestration. Significant challenges remain in security, sustainability, interoperability, and basic research topics, and progress beyond the current state of the art in each of these areas will be needed. Success will require coordinated efforts across research facilities, funding agencies, and technology providers. These efforts must balance flexibility with security, accessibility with preservation, and innovation with reproducibility to enable the next generation of data-intensive research while maintaining scientific rigor.

Acknowledgments. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, supported by the Office of Science of the U.S. Department of Energy under

Contract No. DE-AC05-00OR22725.

REFERENCES

1. M. Cooke, B. Miller, A. Vega, R. Hanisch, K. Murphy, J. O'Neil, A. Mitchell, L. Kewley, E. Steponaitis, K. Haworth, J. Galache, M. Benoit, and S. Gregurick, "NSTC Framework For Considering Data Infrastructure and Interconnectivity In and Among Research and Development Infrastructure Projects," National Science and Technology Council, Tech. Rep., 2024.
2. R. Agrawal, L. Biven, I. Chandramouliswaran, W. Chang, A. Dennis, H. Finkel, D. Hess, M. Lentz, H. Masson-Forsythe, A. Mannes, W. Miller, M. Seale, S. Sellars, S. Spengler, L. Ulmer, and A. Walton, "Innovating the data ecosystem: An update of the federal big data research and development strategic plan," Networking and Information Technology Research and Development, Tech. Rep., 2024.
3. R. Chard, J. Pruyne, K. McKee, J. Bryan, B. Raumann, R. Ananthakrishnan, K. Chard, and I. T. Foster, "Globus automation services: Research process automation across the space–time continuum," *Future Generation Computer Systems*, vol. 142, pp. 393–409, 2023.
4. R. Ferreira da Silva, R. M. Badia, D. Bard, I. T. Foster, S. Jha, and F. Suter, "Frontiers in scientific workflows: Pervasive integration with HPC," *IEEE Computer*, vol. 57, no. 8, 2024.
5. National Academies of Sciences, Engineering, and Medicine, *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*. Washington, DC: The National Academies Press, 2022. [Online]. Available: https://nap.nationalacademies.org/catalog/26532/automated-research-workflows-for-accelerated-discovery-closing-the-knowledge-discovery
6. Big Data Interagency Working Group, "Innovating the Data Ecosystem: An Update of the Federal Big Data Research and Development Strategic Plan," Networking and Information Technology Research and Development Subcommitee of the National Science and Technology Council, Tech. Rep., 2024.
7. J. Luettgau, H. Martinez, P. Olaya, G. Scorzelli, G. Tarcea, J. Lofstead, C. Kirkpatrick, V. Pascucci, and M. Taufer, "Nsdf-services: Integrating networking, storage, and computing services into a testbed for democratization of data delivery," in *IEEE/ACM 16th Int'l Conf on Utility and Cloud Computing*, 2023, pp. 1–10.
8. US DOE Office of Science, "U.S. Department of Energy Selects the High Performance Data Facility Lead," October 2023. [Online]. Available: https://www.energy.gov/science/articles/us-department-energy-selects-high-performance-data-facility-lead
9. T. Desai, F. Ritchie, and R. Welpton, "The Five Safes: designing data access for research," Working papers in Economics no. 1601, University of the West of England, Bristol, Tech. Rep., 2016.
10. US NIH, "NIH Researcher Auth Service." [Online]. Available: https://datascience.nih.gov/researcher-auth-service-initiative
11. T. Clark, H. Caufield, J. A. Parker, S. Al Manir, E. Amorim, J. Eddy, N. Gim, B. Gow, W. Goar, M. Haendel *et al.*, "AI-readiness for biomedical data: Bridge2AI recommendations," *bioRxiv*, pp. 2024–10, 2024.
12. S. Withana and B. Plale, "Patra Model Cards: AI/ML accountability in the edge-cloud continuum," in *2024 IEEE 20th Int'l Conf on e-Science (e-Science)*, 2024, pp. 1–10.
13. W. Dempsey, I. Foster, S. Fraser, and C. Kesselman, "Sharing begins at home: How continuous and ubiquitous fairness can enhance research productivity and data reuse," *Harvard Data Science Review*, vol. 4, no. 3, 2022.
14. M. Herschel, R. Diestelkämper, and H. Ben Lahmar, "A survey on provenance: What for? what form? what from?" *The VLDB Journal*, vol. 26, pp. 881–906, 2017.
15. Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Record*, vol. 34, no. 3, pp. 31–36, 2005.
16. B. Enders, D. Bard, C. Snavely, L. Gerhardt, J. Lee, B. Totzke, K. Antypas, S. Byna, R. Cheema, S. Cholia, M. Day, A. Gaur, A. Greiner, T. Groves, M. Kiran, Q. Koziol, K. Rowland, C. Samuel, A. Selvarajan, A. Sim, D. Skinner, R. Thomas, and G. Torok, "Cross-facility science with the superfacility project at lbnl," in *2020 IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP)*, 2020, pp. 1–7.
17. F. A. Cruz, A. J. Dabin, J. P. Dorsch, E. Koutsaniti, N. F. Lezcano, M. Martinasso, and D. Petrusic, "Firecrest: a restful api to hpc systems," in *2020 IEEE/ACM Int'l Workshop on Interoperability of Supercomputing and Cloud Technologies (SuperCompCloud)*, 2020.
18. P. Widener, A. May, T. Singleton, and O. Kuchar, "Challenges for monitoring and data analytics in a leadership public data repository," in *Proc. 5th HPC Int' Workshop on Monitoring and Operational Data Analytics*, May 2024.
19. A. Nelson, "Ensuring Free, Immediate, and Equitable Access to Federally Funded Research," 2022. [Online]. Available:

https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf

Patrick Widener is a Senior Research Scientist at the Oak Ridge National Laboratory. His research interests include data and metadata management in computational science and workflows. He received his PhD in Computer Science from the Georgia Institute of Technology, and is a Senior Member of the ACM and IEEE. Contact him at widenerpm@ornl.gov.

Laura Biven is Chief Data Officer at Jefferson Laboratory. Her interests include the development of infrastructure and data management practices for FAIR data and computing ecosystems. She received her PhD in applied mathematics from the University of Warwick, UK. Contact her at biven@jlab.org.

Ian T. Foster is Director of the Data Science and Learning Division at Argonne National Laboratory, and Professor of Computer Science at the University of Chicago. His research interests include high-performance and distributed computing. He holds a PhD in Computer Science from Imperial College, UK. He is a Fellow of the AAAS, ACM, BCS, and IEEE. Contact him at foster@anl.gov.

Beth Plale is the Michael A and Laurie Burns McRobbie Bicentennial Professor of Computer Engineering at the Indiana University Bloomington. Her research interests include AI accountability, high performance and distributed computing, and data management. She received her PhD degree in Computer Science from State University of New York Binghamton. She is a Senior Member at IEEE and ACM. Contact her at plale@iu.edu.

Sarp Oral is a Distinguished Research Scientist and Section Head at the Oak Ridge National Laboratory. His research and development interests are parallel I/O and file system technologies, benchmarking, high-performance computing and networking, fault-tolerance, scientific end-to-end data lifecycle, workflows, and computing. Sarp holds a PhD in Computer Engineering from University of Florida. He is a Senior Member of IEEE. Contact him at oralhs@ornl.gov.

Rafael Ferreira da Silva is a Senior Research Scientist and Group Leader at the Oak Ridge National Laboratory. His research interests include parallel and distributed computing systems, with a primary focus on scientific workflows. He received his PhD degree in Computer Science from Institut National des Sciences Appliquées de Lyon. He is a Senior Member of the ACM and IEEE. Contact him at silvarf@ornl.gov.