DISCLAIMER

Data Cards for Standardized Metadata Across DOE-Aligned Data Initiatives: Toward Transparent, Interoperable, and Governed Dataset Documentation

Document Version: 1.0
Date: July 2026

Prepared by:

Diana McSpadden, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA

ORCID: https://orcid.org/0000-0002-8520-1631 | ROR: https://ror.org/02vwzrd76

Olga Kuchar, Oak Ridge National Laboratory, Oak Ridge, TN, USA

ORCID: https://orcid.org/0009-0000-2475-4710 | ROR: https://ror.org/01qz5mb56

Charlotte Neher, Sandia National Laboratories, Albuquerque, NM, USA

ORCID: https://orcid.org/0009-0009-2169-0032 | ROR: https://ror.org/01apwpt12

Victor Walker, Idaho National Laboratory, Idaho Falls, ID, USA

ORCID: https://orcid.org/0000-0001-5172-4410 | ROR: https://ror.org/00ty2a548

Jean Luca Bez, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ORCID: https://orcid.org/0000-0002-3915-1135 | ROR: https://ror.org/02jbv0t02

Laura Biven, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA

ORCID: https://orcid.org/0000-0002-5755-8449 | ROR: https://ror.org/02vwzrd76

Disclaimer

This is a versioned document intended to support the development of a shared Data Card, including both its structured metadata and narrative sections, across multiple DOE-aligned initiatives. Subsequent versions will be published as updates to the Data Card are released. It does not supersede applicable laws, regulations, DOE directives, laboratory policies, contract clauses, records requirements, classification guidance, or review-and-release processes. Nothing in this report should be interpreted as creating new legal authority, replacing site-specific governance determinations, or authorizing release, dissemination, or use of information outside applicable governing processes.

This project was funded by the United States Department of Energy. Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Review and Approval Status

Participating stakeholders have approved this report through technical and programmatic review.

Contents

Abstract

As data-intensive research, advanced computing, and artificial intelligence become increasingly central to scientific and operational workflows, the need for consistent, transparent, and machine-actionable documentation has grown correspondingly. Multiple DOE-aligned communities—including Office of Science, Genesis Mission, American Science Cloud (AmSC), National Nuclear Security Administration (NNSA) stewardship and governance, and related cross-laboratory collaborations—have independently developed metadata practices to support discovery, access, reuse, repository deposit, and compliance. However, fragmentation across these efforts creates barriers to interoperability, shared governance, and efficient reuse. This report presents a technical framework for a shared data card intended to standardize common metadata for datasets across DOE-aligned data initiatives while remaining extensible to discipline- and repository-specific contexts. The report synthesizes lessons from dataset documentation frameworks such as data statements, Datasheets for Datasets, and transparency work in shared artificial intelligence (AI) systems, while also grounding the proposed approach in public-sector data standards and policy guidance, including the ISO/IEC 11179 family, DCAT 3.0, DCAT-US, Office of Science data management and sharing guidance, and implementation-informed metadata standards from Project Alexandria. We describe the purpose, structure, and intended use of the shared data card; discuss how it was developed through a cross-program merging effort; and identify design principles that balance human usability, machine validation and use, metadata interoperability, governance metadata, and repository integration. We argue that a common data card should be understood not as a single-program artifact, but as a cross-program metadata bridge, capable of supporting discovery, accessibility, interoperability, reuse, governed use, and AI-related workflows. The rich and unambiguous information in the data card serves as context for a data object and is a key enabler of a long-term vision to make data findable, accessible, interoperable, reusable (FAIR), and actionable for humans and machines. The article concludes by outlining implementation considerations, expected supporting resources, and future directions for the continued refinement of shared dataset documentation practices.

Acronyms and Abbreviations

AI — Artificial Intelligence
AmSC — American Science Cloud
ASCR — Advanced Scientific Computing Research
CUI — Controlled Unclassified Information
DCAT-US — Data Catalog Vocabulary, United States Profile
DCAT — Data Catalog Vocabulary
DOE — Department of Energy
DOI — Digital Object Identifier
FAIR — findable, accessible, interoperable, reusable
IEC — International Electrotechnical Commission
ISO — International Organization for Standardization
ISOO — Information Security Oversight Office
IRB — Institutional Review Board
JSON — JavaScript Object Notation
LinkML — Linked Data Modeling Language
ML — Machine Learning
NNSA — National Nuclear Security Administration
NSTC — National Security Technology Center
ORCID — Open Researcher and Contributor ID
OSTI — Office of Scientific and Technical Information
PID — Persistent Identifier
PHI — Protected Health Information
PII — Personally Identifiable Information
ROR — Research Organization Registry
SHACL — Shapes Constraint Language
SQL — Structured Query Language
UCNI — Unclassified Controlled Nuclear Information
W3C — World Wide Web Consortium
YAML — YAML Ain’t Markup Language

1. Introduction

Scientific and technical communities increasingly rely on datasets not merely as supplementary materials but as core research assets, operational inputs, and AI-learning resources. As datasets propagate across repositories, discovery catalogs, collaborative platforms, computational environments, and publication workflows, their usability depends not only on technical access but also on the quality and consistency of the metadata that describe them. Data cards encode explicit context for a dataset by combining structured metadata with computable natural language. Structured components enable federated discovery and seamless catalog interoperability. Meanwhile, computable natural language maintains data knowledge that may have previously been implicitly understood, including details concerning generation, processing, and governance. Together, they provide the actionable context required for humans and artificial intelligence (AI) systems to find, access, interpret, cite, and reuse data, and, when appropriate, integrate it into AI and machine learning workflows.

Currently, existing documentation practices remain uneven. Different groups adopt different templates, define different minimum required fields, and emphasize different goals. Some metadata schemas are built primarily for repository deposit; others for AI transparency; others for public release, governance, or machine interoperability. This diversity reflects legitimate local needs, but it also creates friction when multiple groups attempt to share data, align catalogs, support federated discovery, or standardize stewardship practices.

The data card described in this report addresses the challenges above by serving as a common metadata artifact intended to support standardized, cross-program dataset documentation across Department of Energy (DOE)-aligned efforts, including but not limited to Genesis Mission-related implementations, the Office of Science, National Nuclear Security Administration (NNSA), and Advanced Scientific and Computing Research (ASCR) data governance and stewardship initiatives. The shared data card schema and template are not proposed as a replacement for repository-native metadata or for legal or governance review processes, nor do they replace domain-specific schemas. Instead, it is intended to provide a common, structured, and extensible layer of metadata and context to support interoperability across these communities.

This work is motivated by multiple intersecting developments, including the growing recognition that data documentation is a critical part of responsible and reproducible research. This recognition has produced several important prior frameworks, including Data Statements for natural language processing (NLP) [1], Datasheets for Datasets [2], and Data Cards [3], all of which emphasize that datasets require explicit documentation of motivations, provenance, intended use, and risks. Complementary work in model reporting, including Model Cards [4], has underscored the importance of pairing technical artifacts with contextual information about their appropriate use. While these frameworks each contribute essential elements, including machine actionability by Bender and Friedman [1] and transparency and human interpretability in Gebru et al. [2] and Data Cards by Pushkarna et al. [3], they do not fully capture the broader documentation objectives needed to support governed, reproducible, and machine-readable data practices within the DOE environment.

At the same time, the rise of machine learning and increasingly autonomous AI systems, including agentic workflows, has intensified interest in dataset suitability, representativeness, and governance, especially in high-impact and high-risk settings [5]. It has motivated richer contextual documentation and, at the same time, enabled computable natural language to generate, maintain, and interpret that documentation at scale. In parallel with the development of these capabilities, the DOE and the Office of Science have increasingly emphasized data management and sharing obligations tied to public access, reproducibility, repository selection, and long-term preservation as part of our national investments [6-9]. Standards efforts such as Data Catalog Vocabulary (DCAT) 3.0 [10], DCAT-US [11], and DataCite [12] now offer mature vocabulary and schema concepts for describing datasets, catalogs, distributions, access rights, provenance, and related metadata in interoperable ways. As natural language has become computable, conventional structured metadata can be enhanced with textual context to convey details that are better represented in narrative form. The result is an unambiguous, more semantically rich representation of the data, readable and interpretable by both agentic systems and humans.

The shared data card effort described in this report grew from the practical need to reconcile multiple overlapping metadata traditions and requirements and enable the agentic discovery, access, and use of data. The challenges, however, did not allow us to simply adopt one of the existing frameworks. Some communities prioritize discovery and access; others prioritize governance and controlled sharing; others prioritize machine validation or AI suitability; still others prioritize domain-specific extensibility or compatibility with publication and repository systems. The resulting work is therefore neither a purely theoretical standard nor a single-community template. It is a merged and evolving metadata artifact shaped by technical implementation constraints, stewardship feedback, governance review, and standards alignment.

To promote transparency and consistency, the data card template, LinkML schema [13], examples, and supporting documentation have been disseminated through the Genesis American Science Cloud (AmSC) GitLab repository [14] and began being publicly accessible as an archive of the v1.2 tagged release [15]. These resources support data card metadata validation, reuse, and ongoing refinement. The repository also serves as the collaborative workspace for ongoing development and refinement of the shared data card framework.

The article proceeds as follows. Section 2 reviews related work on dataset documentation, AI transparency, and open/public-sector metadata standards. Section 3 situates the data card as a dataset- or collection-level metadata artifact and clarifies its role relative to repositories, governance systems, and standards. Section 4 describes the development of the shared template and the key design choices made during the merging process. Section 5 outlines the proposed structure of the data card and the capabilities it is intended to support. Section 6 discusses the relationship between the data card and existing metadata frameworks, including DCAT, Alexandria, Office of Scientific and Technical Information (OSTI), and Office of Science guidance, as well as the Genesis Mission's sensitivity and governance metadata. Section 7 addresses implementation, validation, and submission considerations. Section 8 describes related efforts within the Genesis Mission and outlines a proposed phased approach to enable autonomous science. Section 9 concludes with implications.

2. Related Work

2.1 Dataset Documentation as an Object of Design

The need to document datasets as meaningful technical artifacts has become increasingly prominent across machine learning, AI, and data-intensive research. A major contribution in this area is Data Statements for Natural Language Processing by Bender and Friedman [1], who argue that dataset documentation should explicitly characterize language variety, speaker demographics, annotation context, and provenance in order to mitigate bias, improve scientific clarity, and enable more responsible deployment of NLP systems. Their work treats documentation not as optional metadata but as a structured statement about the conditions under which a dataset was created and under which it can be expected to generalize.

A parallel line of work includes developing README files as lightweight yet essential documentation artifacts that support data understanding, reuse, and long-term stewardship. Best-practice guidelines from institutions such as Cornell Data Services [16], Princeton Research Data Service [17], and Dryad [18] emphasize that README files should clearly articulate a dataset's purpose, structure, provenance, file organization, processing steps, and any known limitations. Unlike high-level metadata schemas, READMEs are intentionally human-readable and adaptable, enabling researchers to convey contextual knowledge that would otherwise remain implicit or be lost over time. This approach treats READMEs not merely as ancillary descriptions but as integral components of responsible data management, preserving critical methodological details, clarifying assumptions, and facilitating reproducibility by helping future users interpret the data accurately and safely. This emphasis is evident in the adoption of READMEs as a requirement for depositing data in the Oak Ridge Leadership Computing Facility's [19] research data public repository, Constellation [20].

A broader and highly influential framework is Datasheets for Datasets [2], which proposes that every dataset be accompanied by a structured document describing its motivation, composition, collection process, preprocessing, uses, distribution, and maintenance. Importantly, Gebru et al. [2] emphasize that dataset documentation should not be treated as a fully automatable afterthought. Rather, the process of completing the datasheet is itself valuable because it prompts reflection on design assumptions, trade-offs, risks, and downstream consequences.

Pushkarna et al.'s Data Cards [3] extend these traditions by framing dataset documentation as a user-centered transparency artifact. They argue that documentation should function as a boundary object across dataset producers and users with varying levels of expertise, and should support both overview-level understanding and more detailed, structured inquiry. Their work, focused on human consumers of the described Data Card, emphasizes modularity, extensibility, accessibility, and "purposeful and transparent" documentation, and introduces generative frameworks to identify and organize information relevant to dataset consumers.

These strands of work share a core assumption: datasets are not self-explanatory. The scientific process, including how data is generated, processed, and intended for use, is best understood by examining *why* a human created the data and *how* they transformed it to serve their purposes. Documenting these motivations not only makes critical data context explicit but also reveals potential biases and risks associated with reuse. The provenance, structure, intended use, limitations, and governance conditions for data must be clearly communicated if others (humans and agents) are to use them responsibly. The shared data card described in this technical report is aligned with that assumption but focuses more explicitly on the problem of cross-program standardization across heterogeneous DOE-aligned initiatives.

2.2 AI Transparency and the Role of Artifact Context

The rise of AI has underscored the importance of dataset documentation as part of larger transparency and accountability regimes. The authors of "Model Cards for Model Reporting" [4] argue that trained machine-learning models should be accompanied by concise but structured reporting on their intended use, performance across relevant groups and conditions, and known limitations. Model Cards complement dataset-oriented documentation by focusing on the model rather than the dataset, but they reinforce the idea that technical artifacts should not circulate without accompanying contextual information.

More generally, work on transparency in enterprise AI systems has noted that explainability alone is insufficient; users often require access to structured contextual information, iterative interaction, and visibility into both system assumptions and data foundations [21,22,23]. Similar concerns arise in work on shared AI models, where transparency depends not only on model architecture and evaluation, but also on the visibility of contributions, training data, and development processes.

These literatures support one of the key premises of the shared data card effort: data assets should carry sufficient contextual information, both structured and unstructured, to support not only repository deposit and research reuse but also downstream AI and machine learning (ML) applications, governance checks, and machine-actionable workflows.

2.3 Data Governance, Metadata, and Accountability

Recent work on accountability for machine-learning datasets has argued that dataset development should be treated as a disciplined engineering process with distinct stages, documentation practices, and review responsibilities. Hutchinson et al. [5] argue that datasets function as a form of infrastructure, and that robust documentation is necessary to make visible the often-overlooked decisions, assumptions, and labor that shape them. Their argument supports treating dataset metadata not merely as descriptive labels but as part of an accountability framework.

This perspective aligns with governance-focused concerns in many public-sector and cross-laboratory environments, where datasets can be subject to review, compliance, licensing, rights management, records obligations, and controlled access. In these settings, metadata often need to support both operational reuse and governance workflows.

2.4 Public-Sector Open Data and Metadata Standards

Efforts to standardize metadata in the public sector have long been shaped by open data and data catalog initiatives. DCAT 3.0 [10] provides a World Wide Web Consortium (W3C) vocabulary for describing catalogs, datasets, dataset series, distributions, data services, and related metadata. Its purpose is to enable interoperability among data catalogs, support decentralized publication, and facilitate federated discovery. DCAT distinguishes between abstract datasets and their concrete distributions, and supports properties such as access rights, language, publisher, identifier, temporal and spatial coverage, versioning, provenance, quality information, and dataset series relationships.

DCAT-US [11] extends and profiles these concepts for U.S. federal data contexts, providing practical schema guidance for catalog and dataset representation. Together, DCAT 3.0 and DCAT-US offer an important interoperability foundation for any shared metadata effort intended to span catalogs, repositories, and public-sector data systems.

Researchers and DOE-funded entities engage with DataCite's framework via the platforms managed by the OSTI [12]. DataCite is a global nonprofit organization that provides Digital Object Identifiers (DOIs) for research data and other scholarly outputs. DataCite metadata schemas are compliant and encouraged through OSTI's DOE service.

In parallel, implementation-specific standards, such as those developed for Project Alexandria [23], provide examples of how common metadata structures can be adapted to specific governance and repository ecosystems. Alexandria provides a platform to store, catalog, and organize NNSA's non-proliferation research data to improve access and discovery, while ensuring proper access controls. Alexandria's handling of top-level information sensitivity, Controlled Unclassified Information (CUI) category structure, Basic vs. Specified distinctions, and limited dissemination controls offers a particularly valuable implementation reference for shared metadata efforts that must support both open and controlled datasets.

2.5 DOE and Office of Science Data Management Guidance

The DOE Office of Science has also articulated strong expectations around data management and sharing. The Guidance for Reviewers of Data Management and Sharing Plans [25] emphasizes that scientific data generated under funded research should be publicly shared and preserved in ways that support validation and replication of results, subject to limitations required by confidentiality, privacy, rights, and security. The guidance further stresses repository selection, machine readability, metadata quality, and use of persistent identifiers such as the Digital Object Identifier (DOI), Research Organization Registry (ROR), and Open Researcher and Contributor ID (ORCID).

This guidance is directly relevant to the shared data card because it frames metadata not just as descriptive catalog information, but as infrastructure necessary for public access, reproducibility, repository selection, and stewardship. It also aligns with the National Security Technology Center (NSTC) "Desirable Characteristics of Data Repositories," [26] which emphasizes metadata, curation, provenance, unique persistent identifiers, and broad but measured reuse.

2.6 Metadata Registry Standards and Structured Metadata Models

A related body of work relevant to the shared data card comes from metadata registry standards, particularly the ISO/IEC 11179 family [27,28,29,30,31,32,33,34,35,36,37,38], developed jointly by the International Organization of Standardization (ISO) and the International Electrotechnical Commission (IEC). Whereas dataset documentation frameworks such as datasheets and data cards emphasize transparency, context, and responsible use, ISO/IEC 11179 focuses on how metadata can be structured, identified, classified, governed, and reused consistently across systems and organizations.

The shared data card is not a full implementation of the ISO/IEC 11179 metadata registry. However, it adopts many of the same underlying ideas in a lightweight, portable form suitable for dataset-level documentation. These include the use of structured identifiers, names, definitions, administrative metadata, relationships among resources, classification schemes, provenance, stewardship, and lifecycle information.

Several parts of ISO/IEC 11179 are especially relevant. ISO/IEC 11179-1 [27] provides the general framework for understanding metadata as structured descriptions that support common understanding and reuse. ISO/IEC 11179-30 [33] identifies basic metadata attributes appropriate for contexts where a full registry is unnecessary, which closely matches the role of the shared data card. ISO/IEC 11179-33 [36] addresses dataset-level metadata, including identifiers, descriptions, access rights, provenance, distributions, and assessments. ISO/IEC 11179-34 [37] extends these ideas to computable data, including workflows, software environments, inputs, and outputs. Other parts of the family, including ISO/IEC 11179-3 [29], ISO/IEC 11179-6 [31], and ISO/IEC 11179-32 [35], provide useful concepts for registration, stewardship, lifecycle state, and controlled classification.

This standards family is relevant not because the shared data card seeks to replicate a registry, but because it provides a mature conceptual basis for structuring metadata in consistent and interoperable ways. In that sense, the shared data card can be understood as a standards-informed metadata profile: one that draws on metadata registry practice while remaining simpler, more portable, and easier to use across heterogeneous DOE-aligned environments.

3. The Shared Data Card as a Data-Level Metadata Artifact

The shared data card is best understood as an artifact that may describe a single data file, multiple data files, datasets, collections of datasets, or versions of the artifact. Authors of data cards must consider the appropriate resolution for their cards. If too granular, there may be an excessive number of data cards, increasing the burden of documentation, maintenance, and stewardship. Data cards created at too coarse a resolution may lose important details. The data card is not primarily a file-level manifest, nor is it a replacement for repository-native records or governance systems. Instead, it resides at the confluence of several domains: the domain of dataset development, the domain of repository submission, the domain of cataloging and discovery systems, the domain of governance and review processes, and the domain of autonomous data utilization.

3.1 Characteristics of the Shared Data Card

As a data-level artifact, the card is intended to provide a coherent description of an artifact as a whole: its identity, purpose, provenance, access conditions, governance context, and suitability for different forms of use. It should therefore support at least the broad functions outlined in Table 1 and shown in Figure 1.

Table 1. Shared Data Card Characteristics

Characteristic	Description
Primary object of description	Dataset or collection-level data asset
Representation format	YAML and Markdown, contained in a .md file

Characteristic	Description
Audience	Human and machine authors, stewards, repositories, discovery engines, validators
Creation	Intended to support human, machine, and human-in-the-loop creation and updates.
Main goals	Discovery, accessibility, interoperability, reuse, governance, and AI use/reuse support
Governance role	Metadata support for governance, not replacement of governance authority
Extensibility	Yes, including domain-specific metadata extensions
Validation support	Yes, via LinkML and related tooling

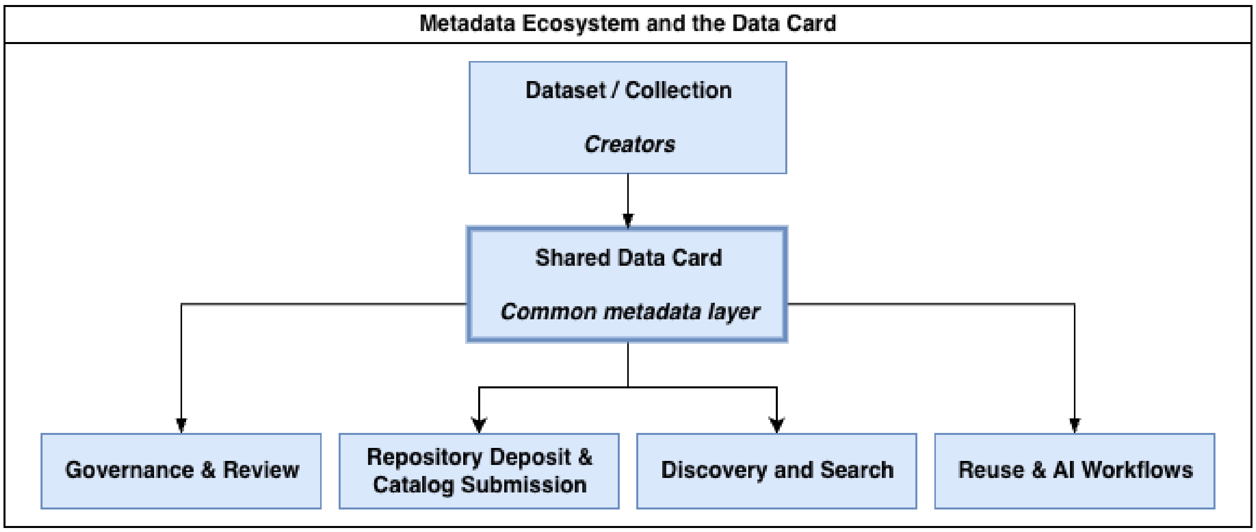

Figure 1. Illustration of the conceptual role of the shared data card in the metadata ecosystem. The data card is designed to facilitate FAIR and AI-Ready data goals.

4. Template Development and Merging Process

4.1 Origins of the Shared Template

The shared template did not begin as a single standard. Rather, it grew out of multiple overlapping efforts, each with legitimate but different priorities. Some groups were focused on discovery and publication. Others emphasized governance, controlled access, or stewardship. Others wanted support for AI suitability, domain extensions, or repository-managed automation.

4.2 Merging Through Capability and Validation Logic

One of the central decisions reflected in the current development work is to move toward a capability-based structure. The root-level `intended_capabilities` block, inspired by and extending findable, accessible, interoperability, and reusable (FAIR) principles [39], signals what the card aims to support. These intended capabilities are summarized in Table 2.

Table 2. Intended Capability Vocabulary

Intended Capability	Purpose
discoverability	Supports cataloging, identification, and search
accessibility	Supports understanding data access and retrieval
interoperability	Supports exchange, and technical and semantic interoperability
reusability	Supports licensing, provenance, and downstream use
governed_use	Supports reviews, review provenance, compliance, and controlled/guided use conditions
ai_usability	Supports AI/ML-specific usage metadata
repository	Recommended repository-managed metadata blocks; updated by the repository

4.3 Simplification of Overlapping Structures

The merging discussions identified that the earlier versions of the template had too many overlapping organizing concepts. The current direction favors:

- FAIR principles, retaining sectioning, primarily for readability,
- using annotations for validation,
- and minimizing extra structural concepts unless they clearly add value.

4.4 Selection of Data Card Format

The approach of using a Markdown (.md) file containing both a YAML Ain't Markup Language (YAML) [40] front matter (for structured metadata) and a Markdown section (for narrative content) was adopted by the multi-community, multi-laboratory group. The Markdown format with YAML front matter is shared across broader Genesis Mission efforts to document datasets, AI models, agents, and other digital artifacts. The combination of structured metadata and narrative documentation provides a flexible mechanism for capturing both machine-actionable information and the contextual details needed for users' understanding and reuse.

The choice of YAML for the frontmatter was a deliberate selection of a data serialization language that is both machine- and human-readable and widely adopted. For instance, Docker Compose files and GitHub Action workflows utilize YAML to declaratively specify system configurations, automation workflows, and execution pipelines. Additionally, YAML supports inline comments within structured metadata, making it easier to create an annotated, self-documenting data card template.

The selection of Markdown for the narrative section of the data card was made to support a lightweight, human-readable, and machine-processable format for documenting contextual information that cannot be easily represented as structured metadata alone. This approach aligns with established community practices, such as READMEs for Research Data [16], which use narrative documentation to communicate the dataset's purpose, methodology, limitations, quality considerations, and guidance on reuse. Markdown is widely supported across software development, scientific computing, and data management ecosystems, supports basic hierarchical document structure, is easily parsable, can be rendered into multiple publication formats, and integrates naturally with version control systems. By combining structured YAML metadata with narrative Markdown documentation, the data card intends to support both automated discovery and validation workflows while remaining accessible and informative to human users.

4.5 Field Annotation Strategy

Fields are annotated using the annotations described in Table 3 to enable validation and to appropriately indicate contextual requiredness. The decision regarding whether fields are designated as required or optional was made by carefully weighing their relevance to imminent use cases against the potential burden on users of completing mandatory fields.

Table 3. Field Annotation Categories

Annotation	Meaning
`required`	Required for the data card
`discoverability_required`	Required for discovery capability; required discoverability fields are also required for all data cards.
`discoverability_if_applicable`	Required when the discovery context applies
`interoperability_required`	Required for interoperability capability
`interoperability_if_applicable`	Required when the interoperability context applies
`reusability_required`	Required for reuse capability
`reusability_if_applicable`	Required when the reuse context applies
`governed_use_required`	Required for governed use/governance
`governed_use_if_applicable`	Required when a governed-use context applies
`ai_usability_required`	Required for AI-related use
`ai_usability_if_applicable`	Required when the AI-use context applies
`reference_only_do_not_include`	Reference fields for repository-aligned metadata; not included in the data card itself

The template's field annotations are also represented within the LinkML schema. In LinkML, each annotation corresponds to a subset that is applied to the relevant fields, indicating which fields are associated with a specific data card capability. While LinkML subset tags themselves do not carry semantic meaning, they provide human-readable context in the schema, and the generated schema documentation uses these tags to group and list the data card metadata fields, as shown in Figure 2.

Slots in subset

Slot	Description
ai_usability	Metadata fields that support the AI usability of the dataset, which is essent...
ai_usage	Describes whether and how this dataset may be used in AI/ML workflows using f...
bias_risks	Any known or potential bias risks associated with using this dataset in AI/ML...
evaluation_use_allowed	Whether this dataset can be used for evaluation of AI models, based on factor...
human_review_required	"Yes"
inference_use_allowed	Whether this dataset can be used for inference with AI models, based on facto...
restrictions	Any specific restrictions or conditions related to using this dataset in AI/M...
safety_considerations	Any known or potential safety considerations associated with using this datas...
training_use_allowed	Whether this dataset can be used for training AI models, based on factors suc...

Figure 2. The data card schema documentation for the slots (fields) tagged with the ai_usability_required LinkML subset.

4.6 Domain-Specific Extension

An optional domain_metadata block is being added to support discipline-specific metadata without forcing all specialized fields into the common base. Multiple domain_metadata blocks are allowed to support multiple domains or schemas associated with a single dataset. Each domain_metadata block includes fields to link the field schema or ontology to reference documentation. The data card template domain_metadata block is shown in Figure 3.

4.7 Alignment Inputs

Figure 4 illustrates the development inputs to the merged data card schema and corresponding template. These inputs include:

- merged shared data card review cycles,
- governance and stewardship comments,
- sensitivity schema redesign,
- repository implementation constraints,
- DCAT/DCAT-US and DataCite alignment work,
- and Alexandria comparison.

```yaml
domain_metadata:                    # [interoperability_if_applicable] e.g., "Accelerator Operations Metadata" |
"Climate Data Variables"
  - name: __NAME__
    description: __DESCRIPTION__        # [interoperability_if_applicable] Free-text description of the domain-specific
    metadata, its purpose, and any important details.
    science_domain: __DOMAIN__          # [interoperability_if_applicable]
                                        # Extends OSTI's supported Subject Areas: "Biology and Medicine" | "Chemistry" |
                                        "Energy Storage, Conversion, and Utilization" | "Engineering" | "Environmental
                                        Sciences" | "Fission and Nuclear Technologies" | "Fossil Fuels" | "Geosciences" |
                                        "Materials" | "Mathematics and Computing" | "National Defense" | "Physics" |
                                        "Power Generation and Distribution" | "Renewable Energy" | "Other"
    schema_reference:                   # [interoperability_if_applicable]
      type: __TYPE__                    # doi | url | ark | handle | local | other
      value: __VALUE__
    version: __VERSION__
    fields:                             # [interoperability_if_applicable] Key-value pairs of domain-specific metadata
    fields. The specific fields will depend on the domain and schema referenced above.
      FIELD_NAME:                       # ${FIELD_NAME} [interoperability_if_applicable] The name of the domain-specific
      metadata field.
        field_value: __VALUE__          # [interoperability_if_applicable] The value for this domain-specific metadata
        field. The specific value will depend on the field and schema referenced above.
        data_type: __DATATYPE__         # [interoperability_if_applicable] The data type of the value (e.g., string,
        integer, float, boolean).
        unit: __UNIT__                  # [interoperability_if_applicable] The unit of measurement for the value, if
        applicable.
        description: __DESCRIPTION__    # [interoperability_if_applicable] A description of the field and its significance.
```

Figure 3. The domain_metadata block from the data card template. Multiple domain metadata extensions are allowed. Linking to a schema/ontology reference is supported.

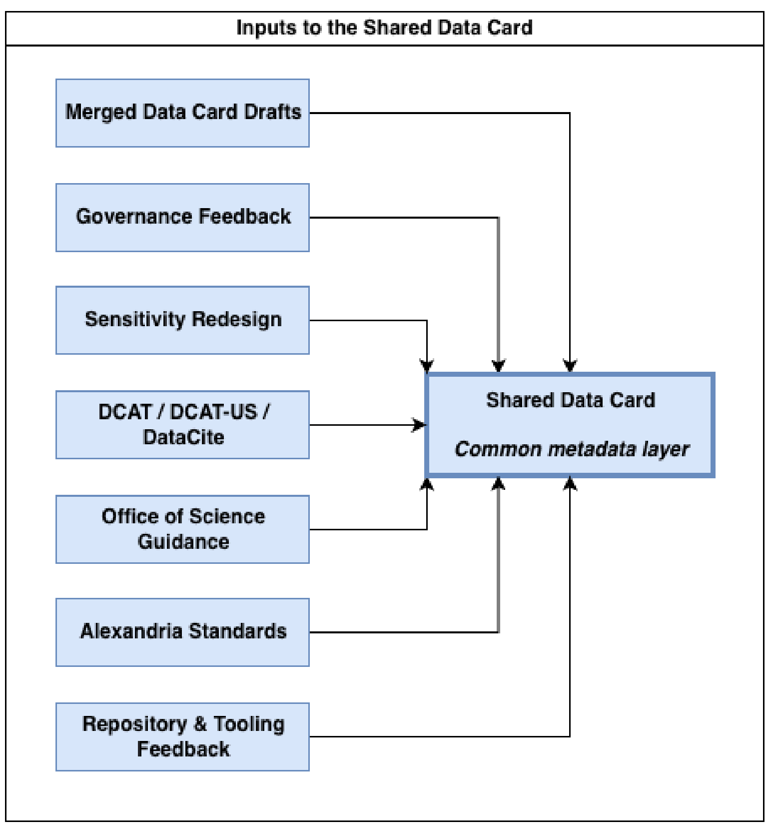

Figure 4. Development Inputs to the Shared Data Card

5. Intended Use

The current direction of the template is to organize metadata according to intended capabilities rather than around a single sequential maturity model. While the exact YAML and LinkML schema continue to evolve, the following high-level blocks, outlined in Table 4, represent the current conceptual organization**.**

Table 4. Intended Uses of the Shared Data Card

Use Area	Primary Contribution of the Data Card
Discovery	Makes datasets findable and identifiable
Accessibility	Describes how datasets may be accessed
Interoperability	Enables structured metadata exchange and transformation
Reuse	Documents reuse conditions, provenance, and limitations
Governed Use	Records review, compliance, sensitivity, and governed-use metadata
AI/ML Use	Supports machine-readable use constraints and suitability information
Interpretability	Provides a human-readable, natural-language guide to help users (humans and machines) understand the dataset and its context.
Contextual Understanding	Captures tacit knowledge and contextual detail in free-text form, supporting human understanding beyond what structured metadata can express

5.1 Discoverability

The card facilitates dataset discovery by utilizing structured metadata that can be indexed, searched, and filtered. Discoverability is the fundamental capability expected of a completed data card, and the *required* discoverability fields are mandatory for all data cards. These required fields include sensitivity metadata for both the dataset and the data card; information about the data card creator (for both human and machine contributors); a designated human contact for the dataset; dataset authors; and sponsor organizations involved in the dataset's development.

5.2 Accessibility

The card describes how a dataset can be requested, accessed, cited, or governed. Accessibility fields may be required or optional. The *required* accessibility fields are required only if the `intended_capabilities` block indicates the data card supports accessibility. The required fields include the dataset access level, which must be selected from a controlled vocabulary, and the location at which to access the dataset.

5.3 Interoperability

The card supports interoperability of the described dataset by providing fields for describing the data collection context, data structure, provenance, related resources, schema, and integrity. This is important for users to understand how to use the dataset and which tools or software they may need to work with it.

Interoperability fields may be required or optional. The *required* interoperability fields are required only if the `intended_capabilities` block indicates the data card supports interoperability. The required fields include those describing the data structure and provenance.

Fields linking the dataset to its related resources (datasets, publications, software, and AI models) are optional. The related resources are linked using a relationship vocabulary shown in Table 5.

Table 5. Relationship Vocabulary for a Dataset's Related Resources

Relationship	Description and DataCite alignment	Applies to Resource Types
`is_derived_from`	Should be used for the dataset that is a derivative of an original resource. Maps to the DataCite IsDerivedFrom relationship.	Datasets, Publications, Software, AI Models
`is_based_on`	Should be used when the dataset continues or is based on a resource. Has a broad mapping to the DataCite Continues relationship.	Datasets, Publications, Software, AI Models
`is_part_of`	Should be used when the dataset is a portion of another resource; may be used for elements of a series. Is an exact map to the DataCite IsPartOf relationship.	Datasets, Publications, Software, AI Models
`has_part`	Should be used when the dataset includes a part of another resource. Is an exact map to the DataCite HasPart relationship.	Datasets, Publications, Software, AI Models
`references`	Should be used when a related resource is used as a source of information for the dataset. Is an exact map to the DataCite References relationship.	Datasets, Publications, Software, AI Models
`other`	Should be used only when the controlled vocabulary does not describe a relationship.	Datasets, Publications, Software, AI Models
`used_to_create`	Should be used when the dataset is used to create software or an AI model. Has a broad mapping to the DataCite IsSourceOf and IsCompiledBy relationships.	Software, AI Models
`used_to_process`	Should be used when the dataset is used to process software resources or AI models.	Software, AI Models
`used_to_analyze`	Should be used when the dataset is used to analyze software resources or AI models.	Software, AI Models
`trained_on`	Should be used when a dataset is used to train software resources or AI models. Has a broad mapping to the DataCite IsRequiredBy relationships.	Software, AI Models
`evaluated_on`	Should be used when the dataset is evaluated on or by software resources or AI models.	Software, AI Models

5.4 Reuse

The card supports the reuse of the described data artifact by providing fields to document the rights, provenance, quality, stewardship, and limitations relevant to reuse. This helps users understand the legal and ethical considerations of using the dataset, who is responsible for maintaining it, and how it has been reviewed for quality and compliance. Required fields describe data completeness, any known issues, and validation methods. Optional fields include the preferred citation (for published data), the data license(s), stewardship information, noise characteristics, and notes documenting data uncertainty.

5.5 Governed Use

Confidently using data under applicable governance rules requires understanding the relevant compliance requirements, restrictions, and governance processes, as well as the history of any formal reviews the data has undergone. Governed use is supported by documentation of the current use of the data, export control status, privacy status, intellectual property status, compliance with a data management plan, OSTI eLink2 metadata compliance, and institutional review board (IRB) approval. Optional fields include explicit descriptions of the data's intended use, permitted use, out-of-scope use, and prohibited use, in addition to a list documenting the reviews conducted on the data.

5.6 AI Use

Safely using the data in AI/ML training, inference, evaluation, and autonomous workflows is supported by requiring the entry of metadata that explicitly documents whether the data may be used for training, inference, and evaluation; any AI use conditions; any biases in the data; AI safety considerations; and whether human review is required in AI workflows.

5.7 Contextual Understanding

Within the Markdown portion of the data card, the information is organized into seven high-level sections aligned with the structured metadata annotations and organization: Machine Usability Snapshot (Table 6), Discoverable, Accessible, Interoperable, Reusable, Governed Use, and AI Usable. Natural language descriptions of the dataset and its context, including documentation of the corresponding data artifact's purpose, collection, methodology, access conditions, data structure, limitations, and AI use considerations, are captured in this section's free-text form.

Table 6. Fields Contained in the Data Card's Machine Usability Snapshot

Intended Capability	Data Card Support
Discoverability	Yes/No
Accessibility	Yes/No
Interoperability	Yes/No
Reusability	Yes/No
Governed Use	Yes/No
AI Usability	Yes/No
License Clarity	Yes/No
Checksum / Fixity	Yes/No
Semantic Context	Yes/No

The sections in the markdown are intended to provide context for a dataset in a human-readable format and to align with the structured metadata. For questions explicitly aligned with structured metadata fields, the associated metadata key(s) are provided with instructions for completing the section. Providing the corresponding metadata keys is intended to support the bi-directional generation of data cards, for example, YAML-to-markdown, or markdown-to-YAML, using an automation tool, such as a large language model (LLM). An example, from the Discoverable section, showing the instructions and metadata keys, is shown in Figure 5.

```
# ---- Discoverable ----

## Description

### Dataset Description [discoverability_required]
[!TODO] <REPLACE: Provide a concise description of the dataset,
including its purpose, scope, and context.><metadata_key:
discoverability.dataset_description.dataset_summary>

### Domain and Purpose [discoverability_if_applicable]
[!TODO] <REPLACE: Describe the domain and the key research areas
involved in collecting the dataset. Can list below>
<metadata_key: discoverability.dataset_description.
purpose><metadata_key: discoverability.dataset_description.
science_domain><metadata_key: interoperability.domain_metadata.
science_domain>
```

Figure 5. The example data card markdown from the Discoverable section shows two questions, their instructions, and the corresponding metadata field keys.

6. Standards and Guidance for Data Card Design

6.1 DCAT 3.0 as the Interoperability Baseline

DCAT 3.0 offers the strongest existing standard foundation for interoperability across catalogs, repositories, and data services. The shared data card is not simply a DCAT serialization, but many of its concepts map directly to DCAT classes and properties, including:

- dcterms:title
- dcterms:description
- dcterms:identifier
- dcterms:creator
- dcterms:publisher
- dcat:contactPoint
- dcterms:issued
- dcterms:modified

- dcterms:language
- dcterms:license
- dcterms:rights
- dcterms:accessRights
- dcat:theme
- dcterms:type
- dcterms:relation
- dcat:distribution
- dcat:DataService
- prov:wasGeneratedBy
- versioning properties and quality-related extensions where needed

This means the shared card should be completed so that its key fields can be transformed into DCAT-aligned records with minimal ambiguity.

6.2 DCAT-US as U.S. Federal Practice Alignment

DCAT-US is particularly relevant because it provides JavaScript Object Notation (JSON) [41] schema and field expectations tailored to U.S. federal catalog environments. The shared card should therefore preserve compatibility, where practical, with:

- identifier structure,
- contact and publisher concepts,
- keyword and theme handling,
- access-rights and license descriptions,
- temporal and spatial metadata,
- and related-resource structures.

6.3 Office of Science and OSTI-Related Expectations

While OSTI does not provide a complete dataset metadata standard in the same sense as DCAT, it relies on DataCite [12] for publishing scientific and technical information and provisioning DOIs. Thus, metadata captured on the shared data card must also eventually satisfy OSTI's DOI requirements. OSTI's workflows, together with DOE data-management guidance, help inform several key design choices for the shared card [6,7,26,42].

In particular, reviewer guidance emphasizes:

- support for validation and replication of results,
- timely and fair access,
- repository selection aligned with desirable repository characteristics,
- data management and sharing resources,
- persistent identifiers,
- metadata and curation quality,
- provenance,
- and explicit treatment of sharing limitations.

These expectations suggest that a good shared data card should include:

- enough dataset description to support validation and replication,
- explicit access and sharing information,

- repository-related metadata,
- provenance and stewardship information,
- and a clear indication of any limitations on sharing or use.

Thus, while DCAT 3.0 [10] provides the strongest metadata vocabulary alignment, the Office of Science / OSTI-related guidance provides strong functional direction for what the data card should capture. The data card employs a controlled vocabulary of relationship terms to connect datasets with related resources such as datasets, publications, software, and AI models. This vocabulary is aligned with DataCite [12] relationships, including:

- datacite:IsDerivedFrom
- datacite:Continues
- datacite:IsPartOf
- datacite:HasPart
- datacite:References
- datacite:IsSourceOf
- datacite:IsCompiledBy
- datacite:IsRequiredBy

6.4 Alignment with ISO/IEC 11179 Metadata Registry Standards

The shared data card is not itself an implementation of an ISO/IEC 11179 metadata registry. However, it is strongly informed by the concepts, structures, and purposes represented across the ISO/IEC 11179 family of standards. In particular, the shared data card can be understood as a dataset-level metadata artifact that draws from ISO/IEC 11179 approaches to identifying, naming, defining, classifying, governing, and documenting data and associated metadata. The shared data card uses many of these ideas in a lightweight, portable form appropriate for cross-program documentation, validation, repository submission, and governed use.

6.4.1 Role of ISO/IEC 11179 in the Shared Data Card Effort

The shared data card aligns with ISO/IEC 11179 in four important ways.

1. Human and machine usability: Metadata should support human readability, machine processing, interoperability, and reuse — per ISO/IEC 11179-1 [27], which frames metadata as structured, standardized, and reusable across systems.
2. Structured attributes: Documentation is treated as structured, identifying, naming, definitional, administrative, relational, and domain-specific attributes, aligned with ISO/IEC 11179-30.
3. Controlled classification: Key fields use controlled classifications or concept-linked metadata rather than free text, per ISO/IEC TR 11179-2 [28], ISO/IEC 11179-3 [29], and ISO/IEC 11179-32 [35].
4. Lifecycle and stewardship: Lifecycle, review, provenance, and stewardship concepts follow ISO/IEC 11179 registration and administration principles, especially ISO/IEC 11179-3 [29] and ISO/IEC 11179-6 [32].

A crosswalk between shared data card concepts and ISO/IEC 11179 parts is included in Appendix B. A crosswalk from shared data card groups to ISO/IEC 11179 concepts is provided in Appendix C.

6.4.2 Where the Shared Data Card Extends Beyond ISO/IEC 11179

The shared data card also includes several areas that are not directly specified by ISO/IEC 11179, or that combine ISO/IEC 11179 ideas with other communities of practice.

These include:

- AI/ML-specific usage suitability and restrictions;
- compact sensitivity and machine-enforcement label hooks;
- repository-ingest operational metadata for implementation workflows;
- profile-based completeness logic tied to intended capabilities;
- blended human-readable and machine-actionable markdown plus YAML authoring patterns.

These extensions should be understood as implementation-oriented additions layered on top of a metadata structure that is broadly compatible with the principles of ISO/IEC 11179, rather than as departures from them.

6.5 Alexandria as an Implementation Reference

Project Alexandria [24] provides a valuable reference for governance-driven research data and metadata for NNSA. It is especially useful for the way it models:

- information sensitivity,
- CUI Basic vs. Specified categories,
- limited dissemination controls,
- Unclassified Controlled Nuclear Information (UCNI) handling,
- steward assignment,
- access, sharing, and usage restrictions,
- and repository/governance distinctions.

The shared data card does not simply duplicate Alexandria's exact schema for its data records, but the Alexandria model offers evidence that these concepts can be implemented in a real governance-aware repository context.

6.6 Sensitivity and Governance Metadata

The sensitivity redesign associated with the current work adds a more structured treatment of:

- overall sensitivity,
- source marking strings,
- source marking scheme,
- classified/CUI/UCNI indicators,
- governance metadata outside the core marking block,
- export control,
- privacy / personally identifiable information (PII) / protected health information (PHI),

- rights / release / records,
- and review provenance.

This redesign follows both DOE- and Information Security Oversight Office (ISOO)-informed governance needs, as well as practical advice from implementation environments such as Alexandria and lab review/release systems.

The shared data card provides sensitivity metadata separately for the data card and the dataset. These are intentionally independent and may differ. For example, a data card may be Public, while the dataset is CUI.

7. Submission, Validation, and Implementation Considerations

7.1 Shared Data Card Template, Schema, and Supporting Resources

The shared data card has been released as a code repository [14] containing a LinkML schema for the structured metadata portion of the card, corresponding schema documentation, data card templates, documentation of the LinkML validation commands, example valid completed data cards, and validation tests illustrating both valid and invalid cases. The structure of the code repository is provided in Appendix D.

7.2 Validation

Ensuring interoperability requires effective data card validation. This motivates one of the primary technical goals of the shared data card: automated validation. This is why the card is paired with:

- a LinkML schema that applies only to the structured metadata section (the YAML frontmatter)
- a template for the complete data card, including both the YAML frontmatter and the markdown narrative sections
- a template for the YAML frontmatter alone, used for an example of the programmatic validation of the structured metadata
- field-level annotations incorporated into the templates and the LinkML schema
- validation tests illustrating both valid and invalid cases.

Validation should ensure:

- required fields are present,
- conditional requirements are respected,
- type enforcement,
- field values conform to expected controlled vocabularies where appropriate,
- and keys remain consistent across template, examples, and schema.

While data card repositories may extend the LinkML-provided validator, the released codebase enables out-of-the-box validation using LinkML's plug-in-based architecture, supporting validation to Pydantic models, JSON [41] schema for structural validation, Shapes Constraint Language (SHACL) [43] for semantic validation, and Structured Query Language (SQL) [44] database constraints to ensure relational data integrity [45]. The validation encompasses completeness and conformance to the schema; however, it does not evaluate the correctness or quality of the information provided.

7.3 Submission

To ensure long-term accessibility, provenance, and discoverability, data cards and, when possible, the data they describe should be deposited in a stable, long-term repository. Ideally, the data card is stored directly with its associated dataset to maintain a strong link between the data's context and its description. Preferably, data cards are submitted to data repositories or catalogs available through the AmSC catalog. An envisioned submission and validation workflow is illustrated in Figure 6.

The AmSC data card code repository [14] is not intended to serve as a long-term archival location for completed data cards. Its contents, outlined in Appendix D, currently include the shared template, schema files, examples, and documentation, and it serves primarily as a venue for collaborative development of the data card framework. Although the repository is not an archive, data cards may be submitted through merge requests, where they are automatically validated by the project's continuous-integration workflow. This validation ensures consistency with the LinkML schema, alignment with template expectations, and compatibility with ongoing schema development. Only public data cards are eligible for inclusion in the AmSC GitLab repository, and other repositories may provide additional or alternative submission mechanisms as needed.

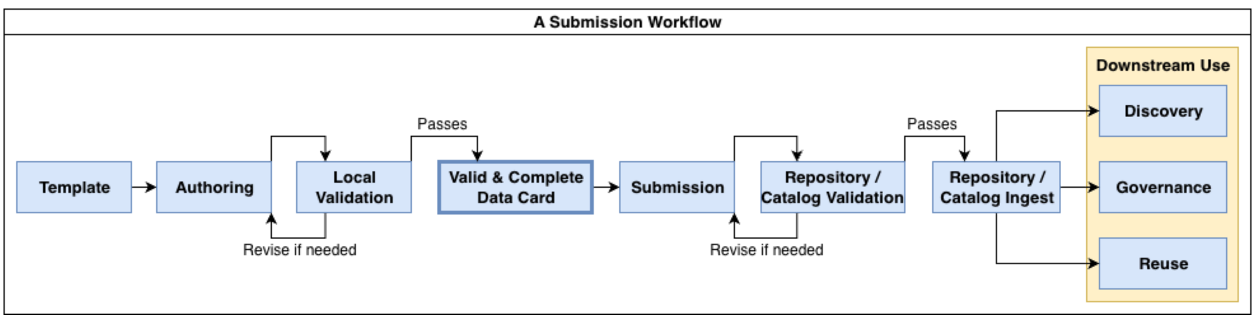

Figure 6. High-Level Submission and Validation Workflow

7.4 Public and Controlled Metadata

An important practical caveat is that repository implementations may not expose all metadata equally. For open/public datasets, the full card may be visible. For controlled or sensitive assets, metadata exposure may be restricted by:

- local policy,
- security and governance controls,
- repository implementation choices,
- or release status.

This is especially important in a DOE-wide context, where not all repositories and catalogs will behave the same way.

7.4 User Experience

The merged data card is designed to support stakeholders of data cards, including upstream AI and human producers and downstream AI and human consumers. Through cross-lab collaboration with corresponding upstream and downstream experts, we have compiled the expected uses of the data cards described in Table 7.

The working-group discussion repeatedly emphasized that the sheer complexity of the full YAML structure can be overwhelming for authors. Therefore, one important implementation principle is that:

- user-facing tools and repository forms may expose only the subset of fields relevant to the current use case or intended capability,
- while still preserving the shared schema beneath the surface.

This approach is consistent with the broader Data Cards literature, which emphasizes structured but user-centered documentation experiences.

Table 7. Uses of Submitted Data Cards

Use	Example Outcome
Catalog registration	Dataset becomes discoverable in a portal or catalog
Governance support	Review/compliance metadata can be surfaced
Repository ingestion	Dataset-level metadata populates repository record
Interoperability transformation	Metadata can be mapped to DCAT-US or related forms; optional domain-specific metadata can be explicitly shared
AI-assisted discovery	Structured metadata supports explicit descriptions of allowed AI uses and risks.

8. Ongoing & Future Work

Near-term work will systematize card authoring using LLM-assisted skills, LinkML validation, and repository pipelines. The approach is to (i) expose only the fields relevant to an author's immediate use case, (ii) draft narrative sections from structured YAML using prompt libraries, and (iii) bi-directional generation from either YAML-to-narrative or narrative-to-YAML to generate schema-compliant submissions. Human-in-the-loop review remains essential: validators will enforce structure and vocabularies while stewards assess correctness, sensitivity, and governed-use constraints. This builds directly on the current template, schema, and validation patterns described in this report and in recent talks.

Efforts to generate data cards *at scale* include using OSTI's expansive catalog of scientific and technical assets, including datasets and corresponding publications, to autonomously generate data, model, and agent cards.

We will integrate completed data cards with discovery pipelines to measure and improve findability (precision/recall, time-to-find, and coverage across catalogs). Milestones already identified include connecting with Genesis Mission teams, including AmSC, validating “discoverability-required” fields across a corpus, and demonstrating improved retrieval using card narratives and relationships.

Future phases will link research artifact cards, datasets, software, models, and agents using persistent identifiers and relationship vocabularies, forming a provenance-aware web of digital assets. Lifecycle cards (data/model/agent) and controlled relationship types will enable agents and humans to traverse lineages, dependencies, and processing chains with confidence.

We intend to leverage curated, provenance-aware workflows as training signals for generative AI. Data cards make dataset context, sensitivity, access constraints, and quality explicit, providing the inputs generative systems need to safely propose, compose, and evaluate reusable workflows. Risks noted in recent briefings (semantic alignment across heterogeneous assets; uneven provenance capture) will be managed through stricter use of persistent identifiers (PIDs), schema crosswalks (e.g., DCAT/DataCite), and governance-metadata normalization.

We will continue to clarify that discoverability and FAIR properties, while foundational, do not alone guarantee AI-readiness. AI-usability requires additional documentation (biases, permitted uses, review provenance, safety hooks) and, where appropriate, human-in-the-loop controls at workflow runtime.

9. Conclusion

The shared data card effort aims to standardize common metadata across DOE-aligned data initiatives while preserving room for local governance, repository implementation, and domain-specific metadata needs. It is not the property of any one project or community. Rather, it is intended as a shared metadata framework that multiple groups can use, validate, extend, and map into their own environments.

Its long-term value depends on balancing:

- common structure and flexibility,
- validation and usability,
- governance needs and open reuse,
- measurable contribution to data discoverability and reuse
- and cross-program interoperability with local autonomy.

As these efforts continue, the shared data card should be understood as an evolving technical object—one that enables conversation and interoperability across communities that may never fully converge on a single metadata worldview, but still need to work together.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Science and U.S. Department of Energy, National Nuclear Security Administration, Office of Advanced Scientific

Computing Research under award number DE-SCL0000139, in alignment with DOE's Genesis Mission.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under Contract No. 89243126CSC000213.

This research used resources of the Oak Ridge Leadership Computing Facility (OLCF; ROR: https://ror.org/05p915b28) at the Oak Ridge National Laboratory, which is supported by the Advanced Scientific Computing Research program in the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

This document and work was also performed at the Idaho National Laboratory and supported by the U.S. Department of Energy (USDOE), Office of Science, Advanced Scientific Computing Research and with the Office of Nuclear Energy under contract no. DE-AC07-05ID14517.

This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes

References

[1] E. M. Bender and B. Friedman, “Data statements for natural language processing: Toward mitigating system bias and enabling better science,” IEEE Trans. Assoc. Comput. Linguistics, vol. 6, pp. 587–604, 2018.

[2] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, “Datasheets for datasets,” Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021.

[3] M. Pushkarna, A. Zaldivar, O. Kjartansson, et al., “Data cards: Purposeful and transparent dataset documentation for responsible AI,” working draft, 2022.

[4] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” in Proc. Conference on Fairness, Accountability, and Transparency (FAT*), 2019.

[5] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, “Towards accountability for machine learning datasets: Practices from software engineering and infrastructure,” in Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2021.

[6] U.S. Department of Energy, DOE Public Access Plan. Washington, DC, USA: U.S. Department of Energy, Jun. 2023. [Online]. Available: https://www.energy.gov/sites/default/files/2023-07/DOE%20Public%20Access%20Plan%202023%20-%20Final.pdf.
[7] U.S. Department of Energy, “DOE Requirements and Guidance for Digital Research Data Management.” [Online]. Available: https://www.energy.gov/datamanagement/doe-requirements-and-guidance-digital-research-data-management. Accessed: Jun. 18, 2026.
[8] U.S. Department of Energy, DOE O 241.1C, Scientific and Technical Information Management, Washington, DC, USA, Oct. 28, 2024. [Online]. Available: https://www.energy.gov/documents/scientific-and-technical-information-management
[9] A. Nelson, Ensuring Free, Immediate, and Equitable Access to Federally Funded Research, Office of Science and Technology Policy, Washington, DC, USA, Aug. 25, 2022. [Online]. Available: https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf
[10] W3C, Data Catalog Vocabulary (DCAT) – Version 3, W3C Recommendation, 2024.
[11] General Services Administration, "DCAT-US Schema v1.1 (Project Open Data Metadata Schema)," resources.data.gov. [Online]. Available: https://resources.data.gov/standards/catalog/dcat-us/
[12] DataCite Metadata Working Group, *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs, Version 4.5. Hannover, Germany: DataCite e.V., 2024. doi: 10.14454/g8e5-6293.
[13] S. A. T. Moxon, H. Solbrig, N. L. Harris, P. Kalita, M. A. Miller, S. Patil, K. Schaper, C. Bizon, J. H. Caufield, S. Cirujano Cuesta, C. Cox, F. Dekervel, D. M. Dooley, W. D. Duncan, T. Fliss, S. Gehrke, A. S. L. Graefe, H. Hegde, A. J. Ireland, J. O. B. Jacobsen, M. Krishnamurthy, C. Kroll, D. Linke, R. Ly, N. Matentzoglu, J. A. Overton, J. L. Saunders, D. R. Unni, G. Vaidya, W.-M. A. M. Vierdag, LinkML Community Contributors, O. Ruebel, C. G. Chute, M. H. Brush, M. A. Haendel, and C. J. Mungall, “LinkML: An open data modeling framework,” GigaScience, vol. 15, p. giaf152, 2026, doi: 10.1093/gigascience/giaf152.
[14] D. McSpadden, O. Kuchar, V. Walker, and J.L. Bez, “Genesis Mission data-cards,” GitLab, 2026 [Source code]. [Online]. Available: https://gitlab.com/amsc2/modcon/dbs/data-cards
[15] D. McSpadden, V. Walker, O. Kuchar, C. Neher, J. L. Bez, and L. Biven, Genesis Data Card Schema, Template and Supporting Tools [Dataset]. Thomas Jefferson National Accelerator Facility, 2026. doi:10.18141/3376151.
[16] Cornell Data Services, “Writing READMEs for Research Data,” Cornell University. [Online]. Available: https://data.research.cornell.edu/data-management/sharing/readme/. [Accessed: Jun. 22, 2026].
[17] Princeton Research Data Service, “READMEs for Research Data,” Princeton University Library. [Online]. Available: https://researchdata.princeton.edu/research-lifecycle-guide/data-acquisition/data-documentation-metadata/readmes-research-data. [Accessed: Jun. 22, 2026].
[18] B. Gee, “For Authors: Creating a README for Rapid Data Publication,” Dryad Blog, Oct. 18, 2023. [Online]. Available: https://blog.datadryad.org/2023/10/18/for-authors-creating-a-readme-for-rapid-data-publication/. [Accessed: Jun. 22, 2026].

[19] Oak Ridge Leadership Computing Facility (OLCF), Oak Ridge National Laboratory, Oak Ridge, TN, USA. https://ror.org/05p915b28
[20] re3data.org, "Oak Ridge Leadership Computing Facility Constellation Portal," *re3data.org - Registry of Research Data Repositories*, editing status Feb. 27, 2025. [Online]. Available: https://doi.org/10.17616/R31NJN0I
[21] I. Barclay, H. Taylor, A. Preece, I. Taylor, D. Verma, and G. de Mel, “A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions,” Concurrency Computation: Practice and Experience,” 2020.
[22] A. Chander, J. Wang, R. Srinivasan, K. Uchino, and S. Chelian, “Working with beliefs: AI transparency in the enterprise,” in *ExSS ’18*, 2018.
[23] E. Stallman and A. Kesari, “Federal open data as an artificial intelligence resource,” George Washington Journal of Law & Technology, vol. 1, no. 1, pp. 155–192, 2025.
[24] R. L. Rodd, DOE Data Days 2023 Report, Apr. 2024. doi: 10.2172/2345124
[25] Office of Science, U.S. Department of Energy, Guidance for Reviewers of Data Management and Sharing Plans, 2025.
[26] White House Office of Science and Technology Policy, Desirable Characteristics of Data Repositories for Federally Funded Research. Washington, DC, USA: Executive Office of the President of the United States, Apr. 2022. https://hdl.handle.net/10088/113528.
[27] ISO/IEC 11179-1:2023. Information technology — Metadata registries (MDR) — Part 1: Framework.
[28] ISO/IEC TR 11179-2:2019. Information technology — Metadata registries (MDR) — Part 2: Classification.
[29] ISO/IEC 11179-3:2023. Information technology — Metadata registries (MDR) — Part 3: Metamodel for registry common facilities.
[30] ISO/IEC 11179-4. Information technology — Metadata registries (MDR) — Part 4: Formulation of data definitions.
[31] ISO/IEC 11179-5. Information technology — Metadata registries (MDR) — Part 5: Naming principles.
[32] ISO/IEC 11179-6:2023. Information technology — Metadata registries (MDR) — Part 6: Registration.
[33] ISO/IEC 11179-30:2023. Information technology — Metadata registries (MDR) — Part 30: Basic attributes of metadata.
[34] ISO/IEC 11179-31:2023. Information technology — Metadata registries (MDR) — Part 31: Metamodel for data specification registration.
[35] ISO/IEC 11179-32:2023. Information technology — Metadata registries (MDR) — Part 32: Metamodel for concept system registration.
[36] ISO/IEC 11179-33:2023. Information technology — Metadata registries (MDR) — Part 33: Metamodel for data set registration.
[37] ISO/IEC 11179-34:2024. Information technology — Metadata registries (MDR) — Part 34: Metamodel for computable data registration.
[38] ISO/IEC 11179-35:2023. Information technology — Metadata registries (MDR) — Part 35: Metamodel for model registration.
[39] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific Data, vol. 3, Art. no. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
[40] YAML Ain't Markup Language (YAML). [Online]. Available: https://yaml.org/
[41] The JSON Data Interchange Syntax, ECMA Int., Standard ECMA-404, 2nd ed., Dec. 2017. [Online]. Available: https://ecma-international.org/publications-and-standards/standards/ecma-404/

[42] U.S. Department of Energy, “DOE Requirements and Guidance for Digital Research Data Management,” May 14, 2026. [Online]. Available: https://www.energy.gov/datamanagement/doe-requirements-and-guidance-digital-research-data-management. Accessed: Jun. 21, 2026.
[43] H. Knublauch and D. Kontokostas, Eds., Shapes Constraint Language (SHACL), W3C Recommendation, Jul. 20, 2017. [Online]. Available: https://www.w3.org/TR/shacl/
[44] Information Technology — Database Languages — SQL, ISO/IEC Standard 9075-1:2023, 2023. [Online]. Available: https://www.iso.org/standard/76583.html
[45] “LinkML Data Validation,” LinkML Documentation. [Online]. Available: https://linkml.io/linkml/data/validating-data.html. [Accessed: Jun. 23, 2026].

Appendix A. Suggested Core Resource Set for Shared Data Card Support

The American Science Cloud repository supporting the shared data card effort is expected to contain the resources shown in Table 8.

Table 8. Expected American Science Cloud Repository Contents

Resource	Purpose
Template	Canonical YAML template for authoring
Examples	Example completed cards for multiple use cases
LinkML Schema	Machine-readable schema and field definitions
Validation	Validation workflows, tests, and examples
Field Reference Guide	Human-readable explanation of fields
Technical Report	This report and related technical guidance
Author Guidance	Instructions for creating and submitting cards

Appendix B. Crosswalk Between Shared Data Card Concepts and ISO/IEC 11179 Parts

Table 9 summarizes the primary alignment between major shared data card concepts and relevant ISO/IEC 11179 parts.

Table 9. High-Level Crosswalk Between Shared Data Card Concepts and ISO/IEC 11179

Shared Data Card Concept	Primary ISO/IEC 11179 Reference	Alignment
General metadata framework and purpose	ISO/IEC 11179-1:2023	Shared focus on making data understandable, reusable, and interoperable through structured metadata
Basic identifying, naming, definitional, administrative, and relational attributes	ISO/IEC 11179-30:2023	Shared data card uses many of the same attribute categories at the dataset/documentation level
Common metadata registry concepts such as item, designation, definition, classification, registration, stewardship	ISO/IEC 11179-3:2023	Shared data card uses analogous concepts in simplified YAML form rather than registry metamodel form
Data definitions and wording quality	ISO/IEC 11179-4	Shared descriptions, summaries, and field definitions benefit from the same principles of clarity, singularity, and precision
Naming principles and controlled naming conventions	ISO/IEC 11179-5	Relevant for identifier naming, file naming, field naming, and controlled metadata expression
Registration, lifecycle, stewardship, review, and progression status	ISO/IEC 11179-6:2023	Shared release/review/compliance/stewardship sections are conceptually compatible, though not identical in vocabulary

Shared Data Card Concept	Primary ISO/IEC 11179 Reference	Alignment
Data specification metadata (data elements, domains, datatypes, value meanings)	ISO/IEC 11179-31:2023	Relevant where the shared data card describes structured features, schemas, variables, datatypes, and missing value codes
Concept systems and ontology-linked semantics	ISO/IEC 11179-32:2023	Relevant to science domain schemas, task taxonomies, semantic context, and ontology alignment
Dataset registration metadata	ISO/IEC 11179-33:2023	Most directly aligned standard for dataset-level metadata such as access level, rights, provenance, temporal/spatial coverage, distributions, and assessments
Computable data / workflow / reproducibility metadata	ISO/IEC 11179-34:2024	Closely aligned with software environment, pipeline, computational provenance, and AI-ready workflow metadata
Model and schema registration	ISO/IEC 11179-35:2023	Relevant for formal schemas, information models, JSON schemas, and related model artifacts referenced by the card

Appendix C. Crosswalk to Shared Data Card Field Groups

The shared data card can also be mapped at the field-group level to ISO/IEC 11179 concepts. Table 10 provides a practical crosswalk between major sections of the card and the ISO/IEC 11179 family.

Table 10. Field-Group Crosswalk Between the Shared Data Card and ISO/IEC 11179

Shared Data Card Section	ISO/IEC 11179 Conceptual Alignment
`identification`	Identifying attributes; identifiers; version; parent/child relationships
`description`	Naming and definitional attributes; definitions; descriptive metadata
`categorization`	Classification scheme usage; concept systems; controlled vocabulary alignment
`dataset_info`	Data specification metadata; datatype, representation, domain, and feature-related metadata
`dates`	Administrative metadata; lifecycle timing; provenance timing
`access`, `access_policy`, `security`	Administrative and relational metadata; rights; access level; governance-related metadata
`provenance`	Provenance, source relationships, derivation, and generation metadata
`stewardship`	Stewardship organization and maintenance responsibility
`reviews`, `compliance`, `release_status`, `workflow`	Registration lifecycle and administrative status analogs
`citation`, `authors`, `contributors`	Designation, attribution, provenance, and contributor documentation

Shared Data Card Section	ISO/IEC 11179 Conceptual Alignment
`semantic_layer`	Concept system, schema, controlled terminology, ontology, or notation linkage
`ai_usage`, `data_quality`, `integrity`	Extended metadata for trustworthy computational or AI use; partly aligned with dataset assessment and computable data registration concepts
`_repository`	Repository-managed administrative metadata, similar in spirit to registry-managed information

Appendix D. Data Card Code Repository Structure

```
data-cards/

├── datacard_schema/

    └── docs/: documentation for the data card schema

    └── project/: programmatically generated version of the LinkML schema, including
a JSON version

    └── src/

        └── genesis_datacard/

            └── schema/

               └── genesis_datacard.yaml : the LinkML schema

            └── datamodel/: programmatically generated Pydantic data model

    └── tests/

        └── invalid/: for testing, example data card segments that are invalid
against the schema

        └── valid/: for testing, example data card segments that are valid against
the schema

├── examples/: example valid completed data card

├── template/: canonical data card template(s) for use when authoring new data cards

   └── archived/: Directories for each archived versions of the data card template
(these are for reference, but please use the latest version (latest/) for new data
cards)

   └── genesis_sensitivity/: contains the sensitivity markings template; a separate,
reference template for the data card sensitivity and governance markings included in
the data card schema and template

   └── latest/: directory containing the latest version of the data card template

      └── datacard_field_reference_guide.md : Field reference guide for the YAML
front matter

      └── genesis_datacard_merged_shared.md : Genesis Datacard v1.0 including the
YAML and markdown sections; version of the template to use when authoring a data
card; copy this to author a new card

      └── genesis_datacard_merged_shared.yaml : Genesis Datacard v1.0 including the
YAML section

   └── version_X.X/: for reference, previous versions of the data card template
```