Pacific Northwest
NATIONAL LABORATORY

Agentic-ready data for materials science

November 13, 2026

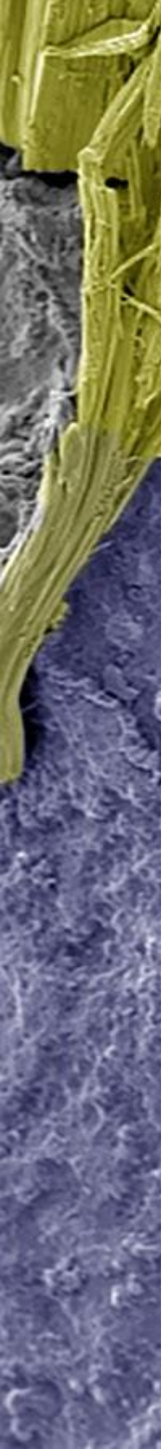
Presented by Nolan Cutler

Vinay Amatya, Kathy New, Alvin Hoang, Sarah Allec, Nathan R. Tallent, Tingkun Liu

U.S. DEPARTMENT of ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy



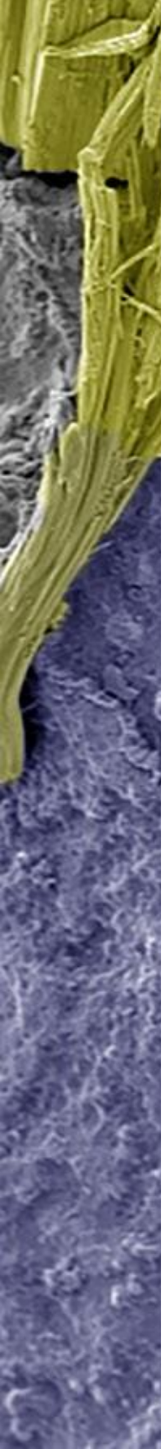


Pacific Northwest NATIONAL LABORATORY

Scientific Integration Problem and LLMs use-case

- Fragmented Observations
  - Instruments – measurements
  - Notebooks – preparation, intent
  - Simulations and publications – hypothesis, assumptions
- Missing scientific context
  - Uncertainty specimen identity blocks joins
  - Missing units obscure comparisons
  - Unrecorded processing breaks traceability

- Why LLM agents?
  - Many data labels, metadata, experiment plans are Natural-language inputs
  - Agents increasingly capable on multimodal data
  - Reduction in manual feature work
  - Can work with few data samples
  - Scientific context via Internal knowledge + expert input

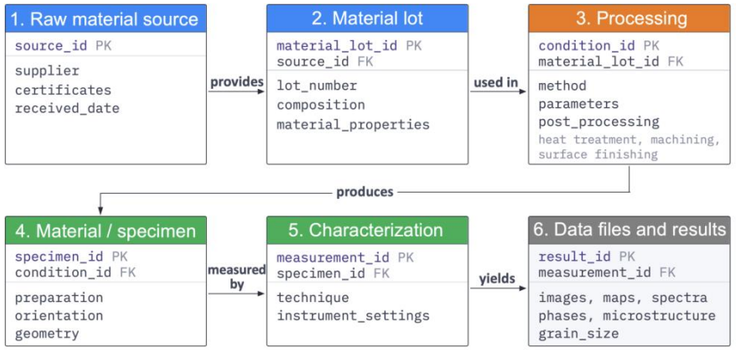


Guiding principles of agentic readiness

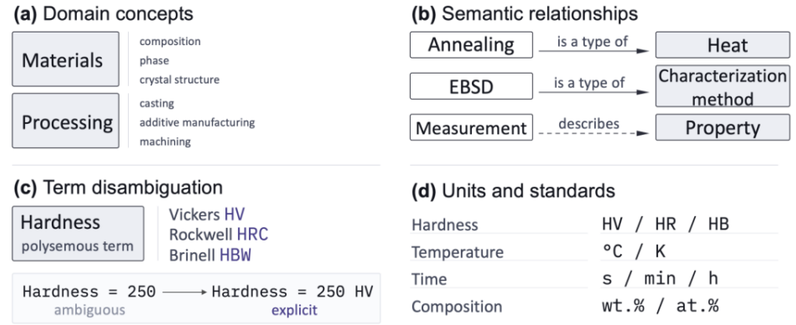
Principle	Operational meaning
FAIR adherence	Discoverable evidence and reusable descriptions
Domain ownership	Producing teams maintain scientific meaning and quality
Lineage Tracking	Outputs record their origin and transformation
Ontology flexibility	Evolving vocabularies coexist with explicit mappings
AI-readiness	Context and relationships support interpretation
Continuum computing	Storage and execution can work across edge, lab workstations and HPC

Data Transformations for AI readiness

- Domain Schematization
  - Defines Raw material source, Material lot, Processing, Material/specimen, Characterization, and Data files and results
  - E.g. composition -> processing -> structure -> properties
- Automated Dataset Enhancement
  - Supplement raw experimental data with metadata or relevant literature information
- Data Provenance and Ontology
  - Provenance aids in reproducibility, error tracing, bias detection; and for compliance and audit.
  - Ontology ensures consistent meaning and use of terms across various data entities
- Source-specific AI Skills
  - Reusable instructions and tools define how each (source) data enters the Lakehouse
  - Skill fusion establishes explicit relationships across resources
- Data Integration and Management
  - Data Lakehouse based data management infrastructure for robust, modular and verifiable data storage and curation.

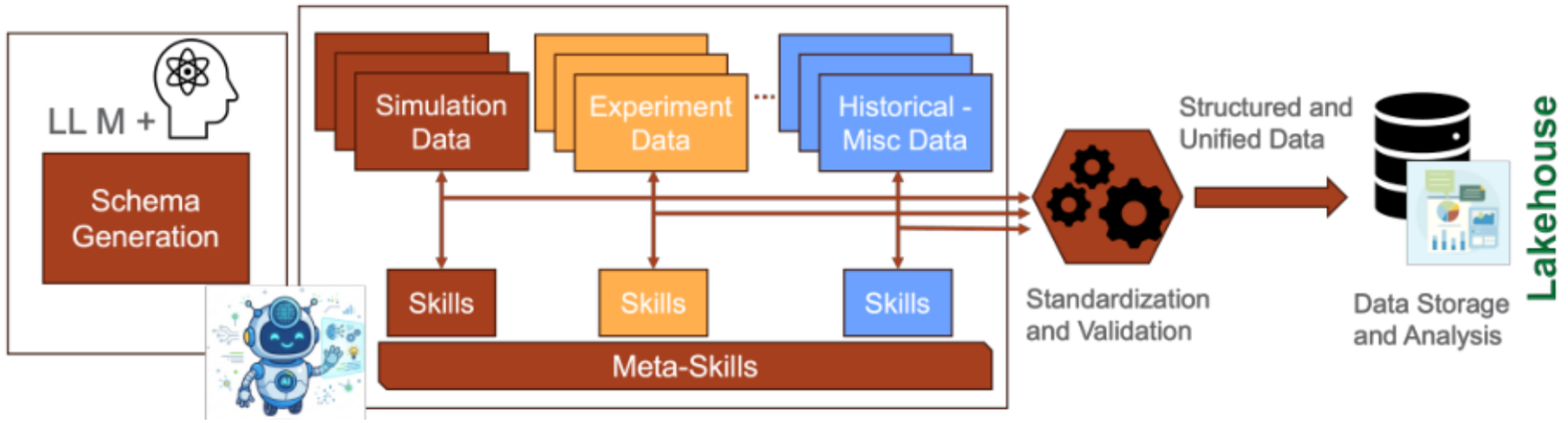


Detail recipe or structured blueprint for complex information



Illustrates concept classes, typed relationships, term disambiguation and units or standards

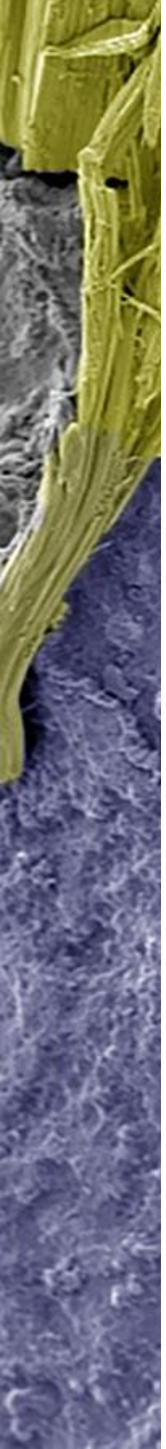
Data Management Workflow



Agentic AI ready Data Management Workflow

- Skills connect schema generation, validation and the Lakehouse

- Phase 1
  - knowledge and rule definition
- Phase 2
  - automated code generation
- Phase 3
  - Validated ingestion
- Phase 4
  - synthesis and cataloging
- Phase 5
  - knowledge discovery



Pacific Northwest NATIONAL LABORATORY

Preliminary Results

Technique	Files	Data volume	Conditions	Skills
EDS	38	79.2 MB	1	1
SEM	6	21.8 MB	1	1
TEM	4	195.0 KB	1	1
TKD	6	702.2 KB	1	1
XRD	29,092	11.2 GB	1	1
Total	29,146	11.3 GB	5	5

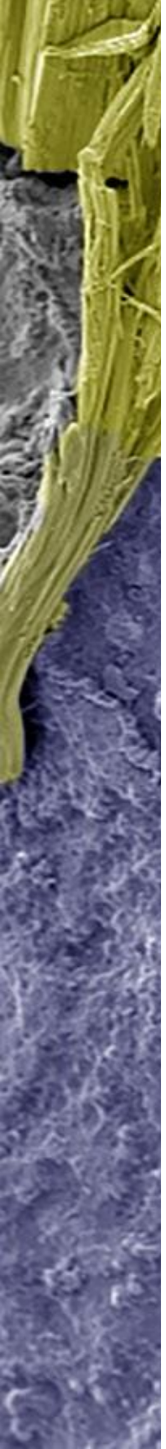
1. AI Agent generated summary of different Raw Data

ID and relation	Endpoints	Type	Context
R1 composition_vs_phase	EDS to TKD	hypothesis	same_field_of_view
R3 dislocation_vs_solute	TEM to EDS	hypothesis	correlated_site
R4 morphology_context	SEM to EDS	hypothesis	same_region

2. Reported hypothesis relationships

ID and relation	Endpoints	Type	Context
R2 bulk_vs_local_phase	XRD to TKD	derivable	shared_sample_condition
R5 phase_evolution_ vs_processing	XRD to XRD	derivable	shared_condition_id
R6 source_to_specimen	__all__ to __specimen__	derivable	shared_sample_condition

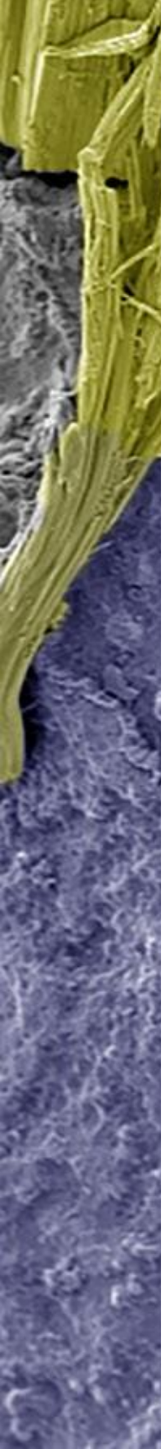
3. Reported derivable relationships



Pacific Northwest NATIONAL LABORATORY

Conclusion

- Detailed workflow for managing complex (experimental) dataset in the materials science domain
- Better tracking provenance, error minimization through standardization and durable data curation
- Demonstrated the value of meta-skills to discover complex relationships between different data sources and types
- Combined the data orchestration workflow with Lakehouse framework, with AI agents for dynamic reasoning, support schema evolution and robust and scalable data management



Pacific Northwest
NATIONAL LABORATORY

Thank you!!

Pacific Northwest
NATIONAL LABORATORY

Thank you



