Semantic Container for Scientific Narrative Of WorkFLow for Learning, Analytics and Knowledge Engineering

Javed I. Khan[1] and Philip Thomas[2]
e-mail: Javed@ kent.edu | plthomas@ kent.edu

[1] Internetworking and Media Communications Research Laboratories
Department of Computer Science

[2] Information Systems Division
Kent State University
233 MSB, Kent, OH 44242
February 2025

1. Introduction

Modern scientific research is increasingly **collaborative, data-intensive, and computationally distributed**. A single study may involve multiple laboratories, instruments, software environments, and analytical workflows operating across continents. Yet, despite this interconnected nature, there remains **no universal, machine-readable language** capable of describing *how* scientific work actually happens—how people, computation, and instruments interact to produce reproducible knowledge.

SNOWFLAKE — Scientific Narrative of WorkFlow for Learning, Analytics, and Knowledge Engineering — has been developed as that language. It is a **structured, descriptive, and partially executable schema** for representing, analyzing, and operationalizing the complete lifecycle of a scientific project. SNOWFLAKE integrates human, computational, machine, and informational components into a single, semantically unified model. Its framework formally defines five core entities:

- **Human Work Elements** (researchers, analysts, field scientists)
- **Computational Work Elements** (software, models, algorithms)
- **Machine Work Elements** (instruments, sensors, hardware systems)
- **Information Product Elements** (data, models, hypothesis, results, and derived insights)
- **Link Elements** (connects and contextualizes the relationships among the other element).

How it is different from other major workflow languages proposed? Classically, a *workflow* has been defined as a "*formal specification of a process that determines the ordered*

sequence of computational or data manipulation tasks, their dependencies, and the flow of control and information among them" (Taylor *et al.*, 2007; van der Aalst *et al.*, 2003; Deelman *et al.*, 2005). Such definitions—adopted by systems like CWL, WDL, Pegasus, and Galaxy (Gil *et al.*, 2007; Curcin & Ghanem, 2008)—emphasize automation and computational reproducibility. These representations, remain **execution-focused and semantically shallow**, describing *how* computation run but oblivious to the *human, epistemic, intention, scientific content dimensions* that define why these computational steps are performed, by whom, and many other untold details.

In contrast, the **SNOWFLAKE Workflow** is formally defined as a **semantically unified schema** that captures **scientific intent, methodology, execution, accountability, openness,** across human, computational, machine, and information layers. It extends the classical notion of a workflow beyond process automation, reframing it as a **knowledge representation framework** that embeds epistemic, organizational, and ethical context directly into the structure of scientific work.

Each SNOWFLAKE-defined workflow becomes a **scientific object**—uniquely identifiable, interoperable, and semantically interpretable. It encodes not only the sequence of tasks but also their **conditions, roles, dependencies, and governing principles**, bridging human understanding with computational precision. Like its namesake, every SNOWFLAKE document crystallizes a unique yet symmetrical structure, capturing the individuality of a scientific process while preserving universal form.

This document [5] presents the structural design and semantic container – as defined by the **attribute specifications** of the SNOWFLAKE schema, defining each entity and its associated descriptors in a **machine-interpretable format**. Each attribute is assigned a unique identifier, structure, and value constraint to ensure consistency, interoperability, and semantic traceability across diverse research environments. It serves as a technical reference for **encoders, system designers, and developers** implementing SNOWFLAKE-compliant registries, workflow capture systems, or data-integration pipelines. Another companion Technical Report [6] discusses the background, rationale, and innovative aspects of SNOWFLAKE with example use-cases.

2. Design of SNOWFLAKE

The **SNOWFLAKE schema** (Scientific Narrative Of WorkFLow for Learning, Analytics and Knowledge Engineering) is designed as a *meta-representation framework* for describing every dimension of a research project — not just its computations, but also the **human, institutional, informational, and ethical contexts** that make science understandable, communicable, reproducible and interpretable. A scientific workflow is not just a series of computing- it must be understood in the context of scientific knowledge it adds to, and the larger collaborative human knowledge processes involved.

2.1. Formal Definition and Distinction of the SNOWFLAKE

2.2. Architecture of SNOWFLAKE

Each SNOWFLAKE workflow is instantiated as a graph of interlinked object classes—**WorkFlow (W)**, **HumanWorkElement (H)**, **ComputationalWorkElement (C)**, **MachineWorkElement (M)**, **InformationProductElement (I)**, and **Link (L)**—providing a **multi-actor ontology** that models the full lifecycle of a scientific process. Unlike traditional workflow description languages that focus on execution provenance and data lineage, SNOWFLAKE introduces attributes for research hypothesis, theoretical framework, innovation, FAIR indices, and ethical-environmental review, thereby achieving both **computational reproducibility** and **scientific interpretability**.

In formal terms, while the classical workflow may be represented as a directed acyclic graph $G = (T, E)$, where T denotes tasks and E the control or data dependencies, the SNOWFLAKE workflow extends this to a heterogeneous, typed graph

$$G_{SNOW} = (V, E, \Sigma_V, \Sigma_E)$$

where $\Sigma_V = \{W, H, C, M, I\}$ defines vertex classes (human, computational, machine, informational) and $\Sigma_E = \{L\}$ defines link semantics capturing execution, data-flow, accountability, and ethical relations. This model allows not only automation and reproducibility but also traceable understanding— enabling cross-domain reuse, meta-analysis, and Open Science compliance at the institutional level.

In SNOWFLAKE, a **Workflow** serves as the *top-level narrative container* that encapsulates:

- The **scientific purpose** (research question, hypothesis, and theoretical framework),
- The **sequence and interrelation** of human, computational, and machine tasks,
- The **information products** created, consumed, or transformed, and
- The **ethical, environmental, and institutional contexts** surrounding the research.

A Workflow thus provides both **procedural structure** (the order and dependency of actions) and **epistemic context** (the intent, innovation, and accountability) — bridging the gap between *automation schema* and *scientific narrative*.

3. Elements of SNOWFLAKE

3.1. WorkFlow (W) — Project-Level Schema

The **WorkFlow (W)** element represents the root scientific process — a complete research activity or experiment encompassing its **intent, methodology, execution, and evaluation**. It functions as the meta-node that unifies all other process components (human, computational, machine, informational, and linking). Each workflow instance encapsulates descriptive attributes for the overall project such as its scientific abstract, research question, theoretical framework, innovation vector, ethical and environmental reviews, data stewardship, indices. Collectively, these metadata form a comprehensive declaration of scientific intent, design, implementation, and societal responsibility. Formally,

$$W = \langle ID, D, H, C, M, I, L \rangle$$

where each component defines a substructure of human, computational, machine, informational, and linking entities that together constitute a **self-describing and semantically linked workflow**.

3.2. HumanWorkElement (H) — Human Task or Cognitive Role

The HumanWorkElement (H) models *human participation and accountability* within a workflow. It captures the intellectual, manual, or supervisory actions performed by individuals or teams—such as designing protocols, verifying outputs, annotating data, or approving results. Unlike traditional workflows that abstract human input as opaque or external, SNOWFLAKE explicitly encodes each actor's role, competency level, accountability entity, task time, tools, environment, and verification method. Formally:

$$H_i = \langle Role, Action, Context, Time, Competency, Verification \rangle$$

Where every human instance contributes both process and epistemic traceability to the workflow.

3.3. ComputationalWorkElement (C) — Algorithmic or Software Process

The ComputationalWorkElement (C) represents any *algorithmic, code-based, or software-driven operation*—from data preprocessing to machine learning training or simulation. It defines the computational function, parameters, algorithms, resource usage, and reproducibility metadata ensure both performance accountability and code-level reproducibility. Formally:

$$C_j = \langle Function, Parameters, Resources, Environment, Verification, FAIR_C \rangle$$

Extends classical machine-task nodes into self-describing computational entities.

3.4. MachineWorkElement (M) — Physical or Instrumental Operation

The MachineWorkElement (M) describes *hardware-dependent or physical system operations*—for instance, a GPU cluster performing inference, a lab instrument capturing spectra, or an IoT node collecting sensor data. Each instance models the instrument type, capacity, location, operating environment, calibration, consumables, responsible entity, and machine-level FAIR index (M-FAIR). Formally:

$$M_k = \langle Device, Capacity, Environment, Calibration, Accountability, FAIR_M \rangle,$$

It ensures that physical instrumentation is not only operationally, but also semantically integrated into the reproducibility framework.

3.5. InformationProductElement (I) — Data or Knowledge Artifact

The InformationProductElement (I) captures the *data entities, models, or knowledge artifacts* produced, consumed, or transformed during workflow execution. Each instance may represent raw datasets, derived models, simulation outputs, documents, or analytical results. Attributes such as SchematicDescription, FileSize, CurationExpertise, ValidationMethod, and I-FAIR Index ensure that data are findable, accessible, interoperable, and reusable. Formally,

$$I_l = \langle Content, Schema, Size, Validation, Custodian, FAIR_I \rangle,$$

It make every data element a traceable and reusable unit of scientific evidence.

3.6. Link (L) — Relational and Causal Connector

The Link (L) element formalizes the relationships and dependencies among all other object classes in SNOWFLAKE. Links define not just execution order (as in classical DAGs) but also semantic, ethical, and accountability relations between workflow components—such as "produces," "verifies," "executes-on," "resides-on," or "supervises." Each link is typed and labeled ($LinkType, LinkLabel$) and connects two instances ($src, sink$), maintaining metadata on cardinality, validation, and ownership. Formally,

$$L_m = \langle src, sink, type, label, verification \rangle,$$

It provides a relational fabric that enables multi-level reasoning, provenance tracing, and process validation.

4. Elements and Their Attributes

4.1. Workflow Attributes of SNOWFLAKE

Table-3 SNOWFLAKE Main Work Flow (W) Attributes

AID	Attribute	ValueExample	Description
103.1	WorkFlowId	"SNOW:10002"	a unique id starting wih SNOW:
103.2	SNOWTypology	"SNOW:W:xxxxx	An organizational classification of the type of workflow.
103.3	NestedCardinality	"This is details of (SNOW:1019.104.1* 'AccountSetup' needed only first time".	Useful if you break down a large workflow into multiple sub-workflows by nesting. Repeat the Cardinality Condition for parent node.
103.4	WorkFlowTitle	"I.am.MAG.Analyser"	a name representing the domain function of the workflow.
103.5	WorkFlowAbstract	"This project is about …."	A paragraph that summarizes what this workflow achieves, its main inputs and outputs, key steps, and significance for a savvy audience, using the scientific domain's vocabulary.
103.6	ScientificDescription	"Develops quantum interferometry techniques to detect biomarkers at ultra-low concentrations."	A concise narrative explaining the project's scientific purpose.
103.7	ScientificSignificance	"Addresses a major gap in early cancer diagnostics using quantum photonics."	Explains why the project is important to science.
103.8	ResearchHypothesis	"Can quantum entanglement improve biosensor sensitivity beyond classical limits?"	States the main research question or hypothesis.
103.9	Innovations&Novelty	"First integration of NV-diamond sensors in biological fluids."	Specifies what is fundamentally new or original.
103.10	TheoreticalFramework	"Quantum measurement theory, photonics, molecular recognition."	Identifies conceptual or mathematical basis.
103.11	ScienceMethodology	"Cryogenic optical setup, fluorescence readout, ML-based signal processing."	Describes the methodology. The design of experiment how research will be conducted.
103.12	ExpectedOutcomes	"Prototype biosensor demonstrating 10× sensitivity improvement."	Defines measurable deliverables or milestones.
103.13	ValidationPlan	"Cross-compare results with standard ELISA assays and quantum simulations."	How results will be tested or verified.
103.14	SocietalRelevance	"Potential to revolutionize medical diagnostics and reduce healthcare costs."	Explain broader impact of the reseatch. Describes benefits beyond scientific community.

Table-3 SNOWFLAKE Main Work Flow (W) Attributes

AID	Attribute	ValueExample	Description
103.15	EnvironmentalImpacts	"Minimal environmental impact; patient sample data anonymized."	Notes compliance with ethics and sustainability.
103.16	InterdisciplinaryScope	"Combines quantum physics, biochemistry, and data science."	Lists fields and institutions that intersect in the project.
103.17	DataStewardshipPlan	"All datasets and code published under CC BY 4.0 on Zenodo."	Describes data management, licensing, and reproducibility.
103.18	OpenScienceStatement	"the result and data will be distributed in open access publications"	Concise statement describing how the project adheres to Open Science principles (openness in data, code, publications, collaboration, and reproducibility).
103.19	Milestones and Deliverables	"M1: Sensor design (Month 6), M2: Prototype test (Month 12)."	Lists key checkpoints and outputs.
103.20	CollaborationNetwork	"MIT (Physics), NIH (Biotech), ETH Zurich (Data Science)."	Participating entities and roles.
103.21	Infrastructure&Resources	"Cryogenic setup, GPU cluster, optical benches."	Specifies hardware, data, or other needs.
103.22	Sponsorship	"NSF Quantum Initiative Grant #QIS-231045 supported $1M"	Sentences- each explaining one funding source, amount, and type of expenditure
103.23	ScientificDomainTags	"Quantum Biology; Medical Physics; Photonics."	Controlled vocabulary or ontology tags.
103.24	ReproducibilityReview	"8/10 – Full methods and data shared."	Assessment of openness and repeatability of the experiments- use EOSC2023 method with dimensions (D, C, E, P, M, Doc, V)- See Table-11
103.25	CumulativeFAIRIndex	"Collected data is accesible and very well catalogued, but the code is not, we could interview remote collaborator"	Cumulative FAIR assessment for each project elements class- see Table-10(a-c)
103.26	Publications&Artifacts	"Cites prior NSF projects #QBIO-2018 and #QMED-2022."	References related works or datasets.
103.27	Risk&UncertaintyFactors	"Cryogenic system reliability; biocompatibility of sensors."	Known limitations or dependencies.
103.28	EthicalReview	"There is no ethical concern in this project".	Text about the ethical issues in the project
103.29	SecurityReview	"There is possibility of patient personal data leak if not handled correctly"	Text about the safety and security issues in the project

Table-3 SNOWFLAKE Main Work Flow (W) Attributes			
AID	Attribute	ValueExample	Description
103.30	EnvironmentalReview	"The chemicals will be disposed off safely".	Text about the environmental impact of the research activities.
103.31	IRBClassification	"It is exempt from IRB review"	Project facts on the IRB classification of the methodology.
103.32	ExpectedDuration	"36 months (Jan 2025–Dec 2027)."	Text describing project start date, end data, and duration.
103.33	KnowledgeAreas	"Biology+Computer Science"	The list of subject area/decipline of the workflow
103.34	InnovationVector	"Methodological + Applied."	Facts about the Direction, magnitude, and domain of innovation such as theoretical, methodological, applied, integrative- see Table-9
103.35	InnovationReadinessLevel	"TRL 3 – Experimental proof of concept."	A text Indicating development stage using recognized readiness scales (TRL/SRL)
103.36	ActorRoleSet	[SystemAdministrator, Field Scientists, Programmer, Data Analyst, GraduateStudent]	Each human worker in a workflow plays a domain-specific role. List all role types represented.
103.37	NodeRoleset	[AutomatedDeamon, DatabaseConnector,Sorter, Gateway Firewall]	Computing,Machine Elements, Information Product all may have unique roles in the overall architecture- so list application specific roles of various non-human elements.
103.38	#HumanWorkerInstances	0	count of HWI in this worflow
103.39	#ComputeWorkerInstances	0	count of CWI in this workflow
103.40	#DataElementInstances	0	count of DEI in this workflow
103.41	TotalNumberOfInstances	0	Total of the HWI+CWI+DEI
103.42	#TotalLinks	0	count of link elements
103.43	PrincipalDirector	"Dr. A. Kumar (orcid.org/0000-0003-0180-0913], Department of Quantum Sciences, Kent State University [grid.17661.33]"	Name and affiliation, of Principal Investigator/Director/ Leader responsible for the researcher or entity. Include Affiliation
104.43	PrincipalInstitution	Kent State University [grid.17661.33], USA"	Lead institution responsible for the researcher or entity. Locality
103.45	ProjectWebLinks	http://medianet.kent.edu	Link to projects website
103.46	WorkflowNarrator	"Professor Linda Walker and Post Doc John Miller"	The person who narrated the workflow to the scribe
103.47	NarrationPeriod	"Recorded between 2025-10-1 and 2025-12-31 over 6 weeks.	A date range reflecting the period with start date, end date, duration-during which the workflow description was recorded.
103.48	WorkflowScribe	"John Candy"	name of the person(s) who encoded this SNOW

4.2. Human Work Attributes of SNOWFLAKE

Table-4 Human Work Element (HWE) Attributes

AID	Attribute	ValueExample	Description
104.1	InstanceID	10002.104.13	Identifier for this human-task instance (workflowID.class.instance).
104.2	Cardinality	"N examiners examines"	Number of individuals perform the identical task.
104.3	EpistemicCondition	"Manually intervene if orange light is seen"	Condition under which this human task executes.
104.4	SNOWTypology	"SNOW:H:xxxxx	An organizational classification of the type of workflow.
104.5	Title	"H1: Account setup"	Short, unique descriptive label for the task or notational identifier like A, B, etc.
104.6	ActorRole	"Publication Data Engineer"	Role of this person in the overall schema.
104.7	FunctionalDescription	"Set up access/password"	What the task does, in domain vocabulary.
104.8	ParametricDescriptors	patient_count, report_type	Inputs/outputs/parameters that impact effort/duration.
104.9	DifficultyDescription	"Time grows linearly with patient_count."	How key parameters affect time/complexity.
104.10	InfoElements	username, password	Information items the actor must possess.
104.11	TaskTime	"2h"	Estimated time for typical scenario.
104.12	ActorLocation	"KSU:Office"	Where the actor performs the task (physical/virtual).
104.13	Persistence	"once"	Frequency/continuity of the task.
104.14	Environment	Windows, Linux	Required operating/working environment.
104.15	Tools	Web Browser, VPN	Software utilities, applications, instruments…
104.16	ProtocolOwner	"OH.KSU.CS.RoboticsLab"	Legal/institutional owner of any SOP/protocol used.
104.17	ProtocolDesigner	"USGS"	Entity that designed the protocol (if any).
104.18	ProtocolType	"Possible to Modify"	Access/licensing/constraints of protocol/SOP.
104.19	HumanFAIRIndex	"Identifiable & accessible"	H-FAIR comments (findable, accessible, interoperable, reproducible).
104.20	**CompetencyRequired**	Expert	Skill level required to perform the task.
104.21	**AccountabilityEntity**	"PI: Dr. A. Kumar"	Person or unit accountable for sign-off.
104.22	**TaskVerificationMethod**	"Supervisor check; audit log"	How completion/quality is verified.

4.3. Computational Work Attributes of SNOWFLAKE

Table-5 Computational Work Element (CWE) Attributes

AID	Attribute	ValueExample	Description
105.1	InstanceID	10002.105.7	Identifier for this computational task.
105.2	Cardinality	"Run n copies for n images"	Number if Identical runs
105.3	EpistemicCondition	"Run when new uncompressed file arrives"	Conditional execution.
105.4	SNOWTypology	"SNOW:C:xxxxx	An organizational classification of the type of workflow.
105.5	Title	"Patient Sorter"	Short, descriptive module name.
105.6	FunctionalRole	"Identity-provider server"	Role of this module in the workflow. Mutiple instances possible for a role.
105.7	FunctionalDescription	"Sort patient records by age"	What the module does in domain terms.
105.8	ParametricDescriptors	n_images, m_iters	Inputs/outputs/parameters affecting cost/output size.
105.9	AlgorithmicDescription	O(n·m)	Complexity w.r.t. key parameters.
105.10	ComputeResource	3×Xeon 28c + P100	Node spec for a typical scenario.
105.11	ExecutionTime	"4h"	Estimated runtime.
105.12	Memory	"64 GB"	Memory requirement per node.
105.13	Persistence	"weekly"	Frequency of execution.
105.14	Environment	Linux; CUDA 12.2	OS/runtime/container constraints.
105.15	Tools	Python, C++, amCharts	Languages/libraries/tools.
105.16	LicenseeOwner	"OH.KSU.CS.RoboticsLab"	Legal owner/licensee of module.
105.17	Manufacturer	"Mindscape Inc."	Original developer/vendor.
105.18	LicenseType	"CC BY-SA"	Usage license type.
105.19	ComputationProcessFAIRIndex	"The program version is available to inspect".	Comments on the foundability, accessbility, integratibility & reproducibility of the experiments plan- use C-FAIR method in Table-10
105.20	SoftwareMaturityLevel	"beta"	Release/stability level.
105.21	ServiceOwner Accountability	"Data Platform Team"	Team accountable for runtime/reliability.
105.22	VerificationMethod	"Unit+integration tests; container digest"	How correctness is established.

4.4. Machine Work Attributes of SNOWFLAKE

Table-6 Machine Work Element (CWE) Attributes

AID	Attribute	ValueExample	Description
108.1	InstanceID	10002.108.3	Identifier for this instrument/equipment step.
108.2	Cardinality	"Calibration only once"	number of Identical instances. Default is 1
108.3	EpistemicCondition	"Calibrate if equipment is moved"	Special condition, if needed to create this product.
108.4	SNOWTypology	"SNOW:M:xxxxx	An organizational classification of the type of workflow.
108.5	Title	"Owen Cluster @ OSG"	Short descriptive equipment title.
108.6	FunctionalRole	"GPU cluster for training"	Role of this equipment in the workflow. Mutiple instances possible for a role.
108.7	FunctionalDescription	"Collects hourly temp & humidity"	What the equipment does.
108.8	ParametricDescriptors	sample_frequency	Parameters affecting time/output.
108.9	CapacityDescription	20×Xeon 28c + 160×P100	Capacity spec relevant to task.
108.10	AllocatedConsumables	paper, ink, electricity	Consumables required.
108.11	ExecutionTime	"3h"	Estimated operation time.
108.12	Location	"Medianet Lab"	Equipment location.
108.13	Persistence	"weekly"	Usage frequency.
108.14	Environment	"Indoor, dust-free"	Operating conditions/specs.
108.15	Tools	"Scale, gloves"	Tools/accessories required to operate.
108.16	LicensedOwner	"OH.KSU.CS.RoboticsLab"	Legal owner.
108.17	Manufacturer	"Westinghouse Inc."	Manufacturer/vendor.
108.18	ShareType	"Open with acknowledgment"	Access/license for using equipment.
108.19	MachineProcessFAIRIndex	"The program version is available to inspect".	Comments on the foundability, accessbility, integratibility & reproducibility of the machine-use MFAIR method in Table-10
108.20	OperatorCompetencyLevel	"certified technician"	Minimum operator qualification.
108.21	ResponsibleEntity	"Instrumentation Core Facility"	Unit accountable for uptime/safety.
108.22	CalibrationVerificationMethod	"NIST-traceable standard; weekly QC"	How calibration/accuracy is verified.

4.5. Information Product Attributes of SNOWFLAKE

Table-7 Information Product Element (IPE) Attributes

AID	Attribute	ValueExample	Description
106.1	InstanceID	10002.106.2	Identifier for this information object.
106.2	SNOWTypology	"SNOW.I:xxxxx	An organizational classification of the type of workflow.
106.3	Cardinality	"2 replicas are instantiated"	Number of Identical instances. Default is 1
106.4	EpistemicCondition	"If there is a request for log then create this file"	Special condition, if needed to create this product.
106.5	Title	"De-identified Cohort v1"	Short descriptive title.
106.6	FunctionalRole	"Graduate Student Records"	Work flow role of this element.
106.7	FunctionalDescription	"This is the list of patients; uncompressed; ecrypted"	Explain what is the fuctional purpose of this object, what it object contains, how fills the purpose.
106.8	SchematicDescription	"this file has records for n persons with {name, age, height}.."	Decribe the metadata- include schema, compression methods used, etc.
106.9	SizeComplexityDescription	n*(a*10 + b*5 + c*5)	Approx. size formula vs key parameters.
106.10	FileSize	"1 GB"	Typical size.
106.11	Encoding	"JPEG"	List the encoding protocols in use.
106.12	Sensivity	"There may be FERPA sensitive student grade data "	List the sensitivy of the information
106.13	Persistence	"log is kept for workflow lifetime."	Expected life of object such as process lifetime, workflow lifetime, indefinite, etc.
106.14	Environment	"AES-256 encryption"	Storage/maintenance constraints.
106.15	Tools	"Parquet; Pandas; QGIS"	Formats, viewers, editors.
106.16	LicenseOwner	"OH.KSU.CS.RoboticsLab"	Owner of the data object.
106.17	Source/Manufacturer	"USGS"	Source/original producer (if derivative).
106.18	LicenseType	"CC BY-SA"	License for use/sharing.
106.19	InformationFAIRIndex	"The data collected is available to inspect".	Comments on the foundability, accessbility, integratibility & reproducibility of the experiments plan- use I-FAIR method in table-10
106.20	CurationExpertiseLevel	"advanced"	Skill level for proper curation.
106.21	CustodianAccountability	"Library & Archives / Data Steward: J. Patel"	Custodian responsible for access/retention/licensing.
106.22	ValidationMethod	"Schema checks; checksum; QC sampling"	How integrity/quality is validated.

4.6. Links Attributes of SNOWFLAKE

Table-8 Link Attributes

AID	Attribute	ValueExample	Description
107.1	InstanceID	10002.107.12	Identifier for this link instance.
107.2	Cardinality	"one-to-many"	Relationship cardinality (default 1-1)
107.3	LinkEpistemicCondition	"stop this path- only if disk is full"	Special condition, if needed to deactivate this path.
107.4	SNOWTypology	"SNOW.L:xxxxx	An organizational classification of the type of
107.5	Title	"Dataset for amoeba sequence"	Domain-facing description of the relationship.
107.6	LinkType	"Execute-After"	Semantic type of relation.
107.7	LinkLabel	"produces"	Optional label for the edge.
107.8	SRCInstanceID	10002.105.7	Source node InstanceID.
107.9	SinkInstanceID	10002.106.2	Target node InstanceID.
107.10	Reserved	TBD	TBD
107.11	Reserved	TBD	TBD
107.12	Reserved	TBD	TBD
107.13	Persistence	"continuously active"	state and duration or continuity of the connection.
107.14	Environment	"TCP Connection"	A decription of enviroment needed
107.15	Tools/Protocols	"Ethernet"	A decription of tools/protocols needed
107.16	Reserved	TBD	TBD
107.17	Reserved	TBD	TBD
107.18	Reserved	TBD	TBD
107.19	LinkFAIRIndex	"The relationship is available to inspect".	Comments on the foundability, accessbility, integratibility & reproducibility of the link-
107.20	Reserved	TBD	TBD
107.21	LinkOwnerAccountability	"Data Engineering Team"	Accountable entity for maintaining this relation.
107.22	LinkVerificationMethod	"Contract tests; lineage check"	How the linkage is validated

Table-9 Link Types & Sub-Types (Labels)

LinkType	LinkLabels	Meaning
Execute-After	FreeSchedule| CoSchedule	Connects task A to task B, implying B executes after A; FreeSchedule = any machines; CoSchedule = coordinated
Co-Execute	FreeSchedule| CoSchedule	Connects A and B; must run together (parallel), optionally on the same machine if CoSchedule
Data-Access	Read|Write|Execute|Create|Delete|Etc.	Links a CWE/HWE instance to an IPE instance indicating the former reads/writes/executes/creates/deletes the
Executes-on	mappable| requested| pending | granted |locked	Links an CWE instance to an MWE instance on which it is assigned to run; states: mappable, requested, pending, granted, locked.
Resides-on	mappable| requested| pending | granted |locked	Links an IPE instance to an MWE instance where it is hosted; states: mappable, requested, pending, granted,
Derived-Into	[name of the process]	Links parent IPE to derivative IPE
Interrupt	[Tokens]	A C/M/H Element can send and Interrupt to another

5. Semantics of the Attributes

The SNOWFLAKE schema provides a **unified, semantically rich framework** for describing the complete structure and behavior of scientific projects—integrating human, computational, machine, informational, and relational dimensions into a single interoperable model. Through its structured set of high-level attributes, it acts as a **semantic container system**, ensuring that each component of a workflow—whether a human task, a software module, a machine process, or an information product—is both *syntactically defined* and *semantically interpretable*. These high-level attributes serve as **semantic containers**, capturing the essential descriptors that define the identity, structure, and function of each workflow element. These are both horizontally and vertically extensible.

5.1. Extension of Semantics by Enumeration

It is intended, that the **deeper semantic meaning** of each attribute is further enriched through their **enumeration values**, which provide contextual granularity and operational precision. Enumeration ranges—such as *persistence types*, *competency levels*, *cardinality types*, or *license categories*—extend the expressive power of SNOWFLAKE while maintaining iteroperability across institutions and systems.

The possible range of these enumerations can be **tailored to an organization's specific practices** or aligned with established **community standards**. For example, **FAIRIndex** or **EOSC2023 reproducibility indicators** can be used for encoding attributes that measure **reusability, transparency, and openness** of scientific elements. Similarly, community-adopted ontologies (e.g., Dublin Core, OBI, schema.org) or domain-specific metadata schemes (e.g., GA4GH, CODATA, ISO 19115) may be used to instantiate standardized value sets.

In this way, SNOWFLAKE bridges **local flexibility and global interoperability**—allowing research organizations to describe their unique workflows while adhering to international best practices in open science, data stewardship, and computational reproducibility.

5.2. Homomorphism

We also designed the schema to be **horizontally extensible**, allowing new attributes to be added as scientific practices evolve. An important architectural innovation in SNOWFLAKE is its **homomorphic design**, in which most attributes are **structurally uniform and conceptually parallel** across the five element classes—Human, Computational, Machine, Information, and Link. This means that analogous attributes such as *ExecutionTime*, *Environment*, *Tools*, *AccountabilityEntity*, and *FAIRIndex* recur across entities with consistent semantics. Such uniformity enables the schema to be computationally traversed, compared, and analyzed regardless of whether a process involves a person, a machine, or a software module. This 'beauty' of the design not only makes SNOWFLAKE more amenable computationally- but more importantly- for easier human conceptualization of the narrative.

Where a consistent analogy could not be established, the corresponding attribute spaces were explicitly marked as **reserved**, rather than forcing non-homomorphic or contextually inconsistent mappings. This preserves conceptual symmetry while leaving room for future semantic expansion.

The **homomorphic structure** supports interoperability at scale by allowing automated systems to infer analogies between heterogeneous entities—for example, *TaskVerificationMethod* in human elements and *VerificationMethod* in computational elements, or *CustodianAccountability* in information products and *ResponsibleEntity* in machine elements. Such structural harmony enables **cross-layer reasoning**, **automated provenance validation**, and **semantic transformations**, including mapping human workflows into computational graphs or exchanging metadata with other workflow description languages.

In practice, this design principle allows institutions to integrate **heterogeneous workflow records**—from laboratory instruments to AI pipelines—under a single, semantically coherent representation. By unifying descriptive structure, attribute meaning, and enumerated value semantics, SNOWFLAKE bridges the conceptual divide between documentation and execution, enabling science to describe, verify, and reproduce itself with unprecedented precision.

5.3. Flexibility & Evolution

Organizations adopting SNOWFLAKE are encouraged to define **local enumeration tables** and **controlled vocabularies** to contextualize attributes such as *Environment*, *VerificationMethod*, *Competency Level*, *LicenseType*, and *CardinalityCondition*. These controlled value sets should align, where possible, with **community or domain standards** — for example, EOSC reproducibility indicators for experimental transparency, FAIR assessment metrics for openness and reuse, or ISO/IEC standards for process verification

and instrument calibration, but at the same time – adjust extend to accommodate local situations and needs more faithfully instead of strictly adhering to standards. Collection of these localized schemes can provide pathways to future improvements of the standards.

The SNOWFLAKE specification encourages a **federated implementation model**: institutions can extend or specialize attribute enumerations while maintaining interoperability through the preservation of the global attribute structure and semantic identifiers. In practice, this means a laboratory, data center, or research network can encode its internal procedures and compliance constraints (e.g., IRB classifications, safety protocols, or software certification tiers) within the SNOWFLAKE schema while remaining cross-compatible with external repositories and metadata registries.

By keeping attribute definitions stable and **enumerations adaptable**, organizations can achieve both **semantic precision** and **institutional flexibility**—ensuring that every SNOWFLAKE instance remains interoperable, auditable, and semantically self-describing across projects, platforms, and scientific disciplines.

6. Conclusions

The project has been supported by funding from the National Science Foundation NSF Award#2201558, and NSF Award# 1925678, engineering time contributed by the Division of Information Technology, location and engineering time donated by Department of Computer Science. The ICE cluster has also received supplementary funding from the division of Research and Economic Development.

7. References

[1] van der Aalst, W.M.P., ter Hofstede, A.H.M., Kiepuszewski, B. and Barros, A.P. (2003) '*Workflow patterns*', Distributed and Parallel Databases, 14(1), pp. 5–51., Available at: https://doi.org/10.1023/A:1022883727209.

[2] Deelman, E., Singh, G., Su, M.-H., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Vahi, K., Berriman, B., Good, J., Laity, A., Jacob, J.C. and Katz, D.S. (2005) '*Pegasus: A framework for mapping complex scientific workflows onto distributed systems', Scientific Programming*, 13(3), pp. 219–237., Available at: https://doi.org/10.1155/2005/914734.

[3] Taylor, I.J., Deelman, E., Gannon, D. and Shields, M. (eds.) (2007) *Workflows for e-Science: Scientific Workflows for Grids*. London: Springer., Available at: https://doi.org/10.1007/978-1-84628-757-2.

[4] Curcin, V. and Ghanem, M. (2008) '*Scientific workflow systems—Can one size fit all?*', in Proceedings of the Cairo International Biomedical Engineering Conference (CIBEC 2008). Cairo: IEEE, pp. 1–9., Available at: https://doi.org/10.1109/CIBEC.2008.4786085.

[5] Javed I. Khan and Philip Thomas, *An Exploration into Scientific Narrative Of WorkFLow for Learning, Analytics and Knowledge Engineering*, Technical Report 2025-03-01 Internetworking and Media Communications Research Laboratories, Department of Computer Science, Kent State University, available from: http://medianet.kent.edu/technicalreports.html

[6] Javed I. Khan and Philip Thomas, *Semantic Container for Scientific Narrative Of WorkFLow for Learning, Analytics and Knowledge Engineering*, Technical Report 2025-03-02 Internetworking and Media Communications Research Laboratories, Department of Computer Science, Kent State University, available from: http://medianet.kent.edu/technicalreports.html