CFDE Request for Comments

CFDE-Draft-# (RFC Admin Only):			
CFDE-Draft-Title:	The JKG: JSON Knowledge Graph Exchange Format		
CFDE-Draft-type: (Right click the box and select checkmark ✔ CTRL - Z to undo)	Standard Implementation ❏	Design Principle ☑	Policy ❏
Point of Contact Name:	Deanne Taylor and Jonathan Silverstein		
Point of Contact Email Address:	taylordm@chop.edu; j.c.s@pitt.edu		
End of Date Comments *(MM/DD/YY)*:	03/17/2026		
Submitting Team/DCC Name:	U24: Genetics of Childhood Cancer and Birth Defects		
CFDE-Draft-Status:	Pending ☑	Active and Open for Comments ❏	
URL Link to the document:	https://docs.google.com/document/d/1lQFoe8C134gmLBHsqYqPYWpC7n__PYMTiRayZ4Bbw/edit?usp=sharing		
License (Optional):	MIT License		

Purpose

This RFC is presented as a draft to solicit community feedback on the scope, constraints, and design decisions embodied in a newly developed JSON Knowledge Graph (JKG). This RFC would like to get community feedback on whether a validated, self-sufficient, and scalable JSON representation can serve as a unified, shared exchange and construction layer across CFDE-related knowledge graph (KG) efforts.

The JKG format is designed to support both canonical graph construction and portable graph exchange, allowing graph schemas and assertional data to be represented in a unified, validation-ready format. A formal JSON Schema specification for JKG is publicly available at https://github.com/x-atlas-consortia/json-knowledge-graph/blob/main/JKG_Schema.json, along with reference implementations and validation tooling available at the project level at https://github.com/x-atlas-consortia/json-knowledge-graph.

Specific questions suggested for community feedback include: whether the required JKG node and relationship categories are minimal or overly prescriptive; whether JSON Schema provides sufficient validation at the exchange layer for CFDE use cases; how graph-level validation should be handled beyond schema constraints; whether exchange and canonical construction concerns should remain unified or be separated into layered specifications; whether this approach can serve as a method for submitting assertional knowledge to CFDE; whether this approach can serve plans to integrate C2M2 with assertional semantics; and whether this approach can support the use cases proposed by the CFDE Knowledge Graph Working Group for sub-graph extraction, graph exchange and graph ad hoc assembly from various sources.

A validated, scalable, and domain-agnostic exchange format has potential to reduce duplication across NIH-funded projects, lower barriers to reuse and federation, improve transparency and reproducibility, and provide a durable foundation for data integration and AI-driven analysis.

Background

Knowledge graphs have become foundational infrastructure for biomedical research, supporting data integration, discovery, and advanced analytics across genomics, clinical data, cell atlases, electronic health records, and AI-driven modeling. Across the NIH Common Fund Data Ecosystem, substantial investments have been made in knowledge-graph-based resources.

Despite this investment, there is no shared, validated, and scalable exchange and construction format for assertional knowledge graphs. Most graph formats in active use are internal construction schemas optimized for specific execution environments. While effective locally, they do not function well as portable, inspectable, or validation-ready representations.

As a result, exchanging graphs across platforms typically requires bespoke mappings, custom code, or full translation and re-ingestion pipelines. Validation of graph content is often implicit, ad hoc, or tightly bound to a particular system rather than expressed as explicit, reusable constraints. These limitations become particularly acute at scale and are especially important to address for federated analysis, machine learning, and cross-consortium collaboration.

Previous Work

To address this gap, the JSON Knowledge Graph (JKG) has been developed as a candidate specification for representing, validating, exchanging, and constructing knowledge graphs. JKG is not solely a conceptual proposal. It is implemented as an executable format with a formal JSON Schema specification, reference validation notebooks, and scalable performance.

JKG documents can be validated using standard JSON Schema tooling, with example Jupyter notebooks provided in: https://github.com/x-atlas-consortia/json-knowledge-graph. Synthetic and real-world graphs containing millions of nodes and relationships have been successfully validated within practical time bounds using Python-based JSON Schema validation. Reference implementations demonstrate deterministic import into Neo4j using widely deployed and free to the community APOC procedures, export back to JSON, and conversion of database exports into minimized, re-importable JKG representations. Ingest and export performance has been demonstrated at tens-of-millions scale within minutes, and schema-level validation at multimillion scale completes within reasonable, interactive timeframes using parallel processing on modest hardware. This work includes an isomorphic implementation of the Unified

Biomedical Knowledge Graph (UBKG), including UMLS ingestion and the addition of clinical data from electronic health records, as demonstrated at recent CFDE meetings. Documentation for UBKG is available at https://ubkg.docs.xconsortia.org.

JKG is intentionally domain agnostic. The format makes no assumptions about biomedical ontologies or domain-specific semantics. It provides explicit representation of nodes and relationships, source attribution for all assertions, and flexible attribute-value patterns for annotation without embedding domain logic into the specification. As a result, JKG can represent non-biomedical knowledge graphs with equal fidelity.

JKG is complementary to, but distinct from, existing semantic metamodels such as the Biolink Model (see https://biolink.github.io/biolink-model). While Biolink provides rich ontology-aligned semantics well suited for biomedical reasoning and harmonization, it introduces complexity and shared semantic dependencies that make it difficult to use as a lightweight exchange or construction format. JKG instead prioritizes explicit structure, self-sufficiency, and validation over semantic prescriptiveness, allowing it to function as a scalable, domain-agnostic substrate beneath richer semantic models rather than as a replacement for them. In short, Biolink-Model data can be ingested into JKG, while JKG can also simultaneously contain other semantic frameworks such as UMLS, OBO, and non-biomedical semantic models.

JKG's flexibility arises from a simplified model, designed around a small set of explicit type categories including Source, Node_Label, Rel_Label, Concept, and Term. These enable graph self-sufficiency, explicit provenance, and transparent semantic labeling. Relationships support attribute value patterns that allow qualitative or quantitative annotation, including evidence, ranges, and units, without introducing nested structures that degrade performance in property graph systems. This is enabled via explicit attention to cardinality of information, such that no property on edge nor node has deep content (JSON trees) rather than a text or numeric item. This approach, enabling multiple coherent edges among pairs of nodes to the extent needed but typically limiting to one directional relationship among Concepts or Concept to Term enables full specification complex constructs (e.g., EHR transactions). Also, rather than two Concept to Concept relationships in opposite directions, as prior UBKG, reverse relationships are explicitly recognized, again limiting cardinality to what is needed for expressiveness. Validation is explicitly divided into schema level validation, enforced through JSON Schema, and graph level validation, such as global identifier uniqueness or semantic consistency, which cannot be expressed in schema alone and is intentionally left to optional downstream validation tooling.

Next Steps

The next phase of work requires effort beyond the current prototype. This includes development of systematic graph-level validation tooling, broader stress testing across domains and graph backends, refinement of the specification based on structured community feedback, expansion of reference implementations and documentation, creation of reproducible test corpora, and establishing governance and versioning models to support evolution. These activities align directly with the goals of the CFDE RFC process by enabling transparent, inclusive, and evidence-based consensus building around a concrete, tested candidate specification.

RFC approval by Centers and DCCs (metadata submitters/PMs or someone else with an appropriate role)

DCC / Center	Name of the Approver	Initials of the Approver
Metabolomics Workbench	Srini Ramachandran	SR
GlyGen	Jeet Vora	JKV
SPARC	Jeffrey Grethe	JSG
KidsFirst	Deanne Taylor	DT
LINCS	Daniel Clarke	DC
ExRNA	Varduhi Petrosyan	VP
IDG		
HubMAP	Jonathan C. Silverstein	JCS
MoTrPAC	Jimmy Zhen	JZ
SenNet	Jonathan C. Silverstein	JCS
4DN		
GTEx	Kristin Ardlie	KA
SCGE		
Bridge2AI		
HRA		
ICC		
DRC	Daniel Clarke	DC
KC		
TC		
UniProt-CFDE partnership	Chuming Chen	CC

Comments from the CFDE Variant Working Group and CFDE ERCC DCC Responses

Comment #1: [Summary description of comment topic]
[Commenter #1 Name]: [Comment text]
[Commenter #2 Name]: [Comment text]
KidsFirst/HuBMAP/SenNet response: [Response text]