Authors: Khairullin Artur (RU) *prcr98@gmail.com* 15 June 2026

—

IFDRP: An Intelligent Format for Data Representation and Processing Based on Semantic Equivalence and Multimodal Knowledge Graphs

Abstract

The exponential growth of digital information has revealed fundamental limitations in existing document and data representation formats. Contemporary formats such as PDF, DOCX, HTML, XML, and JSON successfully preserve content and structure but remain insufficient for preserving machine-interpretable semantics, contextual dependencies, and cross-modal relationships. As a result, information processing systems frequently lose significant portions of knowledge embedded within documents during extraction, transformation, and integration procedures.

This paper introduces the Intelligent Format for Data Representation and Processing (IFDRP), a conceptual framework designed to unify document representation, semantic interpretation, and computational processing within a single graph-oriented knowledge structure. The core contribution of IFDRP is the principle of Intellectual Equivalence, which establishes formal requirements for semantic, structural, contextual, and functional preservation between an original document and its graph-based representation.

The proposed architecture integrates textual, tabular, graphical, mathematical, and metadata components into a unified multimodal semantic graph. A formal model, equivalence metrics, and evaluation methodology are presented. The framework aims to enable reversible document transformation, advanced semantic search, automated reasoning, and integration with artificial intelligence systems while preserving the informational integrity of original sources.

Keywords: knowledge graphs, semantic representation, document engineering, multimodal data, information systems, semantic interoperability, intelligent documents.

1. Introduction

The modern digital ecosystem is characterized by an unprecedented volume of heterogeneous information distributed across documents, databases, web resources, engineering specifications, scientific publications, and regulatory standards. While significant advances have been achieved in data storage and transmission, the problem of preserving and operationalizing knowledge remains largely unresolved.

—

Traditional document formats are primarily designed for human consumption. Portable Document Format (PDF), Office Open XML (DOCX), HyperText Markup Language (HTML), and Extensible Markup Language (XML) provide mechanisms for storing content and presentation. However, these formats do not inherently preserve the semantic relationships necessary for automated reasoning and intelligent processing.

Consequently, information extraction systems must repeatedly reconstruct semantic structures from documents, resulting in information loss, ambiguity, duplication of effort, and reduced reliability. This challenge becomes particularly significant in technical, scientific, legal, and regulatory domains where contextual relationships and logical dependencies are essential.

To address these limitations, this paper proposes the Intelligent Format for Data Representation and Processing (IFDRP), a conceptual architecture that transforms documents into multimodal semantic knowledge graphs while maintaining the ability to reconstruct the original representation.

The central hypothesis of IFDRP is that meaningful information can be represented as a graph-based semantic structure while preserving intellectual equivalence with the source document.

2. Related Work

Research on semantic representation has evolved through several complementary directions.

The Semantic Web introduced machine-processable representations through RDF, OWL, and linked data technologies. These approaches established mechanisms for representing entities and relationships but were not designed to preserve complete document reconstruction.

Knowledge graph research extended semantic modeling by providing scalable frameworks for integrating heterogeneous information sources. Modern knowledge graphs support reasoning, entity linking, and semantic querying across large-scale datasets.

Document engineering research has focused on markup languages, metadata standards, and structured document models. XML and related technologies significantly improved machine readability but primarily represent syntax rather than knowledge.

Multimodal machine learning introduced techniques for integrating textual, visual, and structured information. Recent architectures demonstrate the effectiveness of unified semantic embeddings across multiple modalities.

Despite these advances, no widely adopted framework simultaneously provides:

—

- semantic preservation,
- structural preservation,
- contextual preservation,
- reversible reconstruction,
- multimodal integration,
- graph-native processing.

IFDRP is intended to address this gap.

3. Problem Statement

Consider a technical standard containing:

- textual definitions;
- engineering drawings;
- numerical tables;
- formulas;
- references;
- contextual notes.

In traditional processing pipelines, each modality is extracted independently. Semantic relationships connecting these elements frequently become fragmented or lost.

Let D denote an original document and G denote its semantic graph representation.

Current approaches typically satisfy:

D → G

but do not guarantee:

G → D

while preserving semantic integrity.

The absence of formal reversibility limits trust, auditability, and automated knowledge management.

The objective of IFDRP is therefore to establish a formally verifiable transformation framework satisfying both directions.

Authors: Khairullin Artur (RU) *prcr98@gmail.com*

—

3.1 Beyond Knowledge Graphs

Knowledge graphs represent entities and relationships.

However, knowledge graphs are not inherently designed to preserve the complete informational identity of source documents.

Their primary objective is semantic representation.

IFDRP extends this paradigm by introducing:

- reversible transformations;
- structural preservation;
- contextual preservation;
- multimodal preservation;
- audience-oriented reconstruction;
- executable knowledge generation.

Consequently, IFDRP does not treat a graph as a final representation.

Instead, the graph becomes an intermediate semantic state from which multiple equivalent representations can be generated.

This distinction transforms the graph from a storage model into a semantic transformation layer.

3.2 IFDRP versus Knowledge Graphs
Although IFDRP utilizes graph-based structures, it should not be considered merely another knowledge graph implementation.

Conventional knowledge graphs focus primarily on representing entities, relationships, and semantic facts. Their objective is knowledge organization and reasoning.

IFDRP introduces a broader objective: preservation, transformation, reconstruction, and operationalization of knowledge across multiple representations.

Table 1 summarizes the conceptual differences.

Property	**Knowledge Graph**	**IFDRP**
Primary objective	Knowledge representation	Knowledge transformation
Canonical document reconstruction	Not required	Core requirement

—

Property	**Knowledge Graph**	**IFDRP**
Structural preservation	Optional	Mandatory
Context preservation	Partial	Explicit
Semantic equivalence verification	No	Yes
Audience-specific regeneration	External process	Native capability
Cross-lingual regeneration	Limited	Native capability
Interactive application generation	Optional	Native capability
Semantic canonical form	No	Yes
Transformation between representations	Limited	Core capability

Therefore, IFDRP should be viewed not as a graph database or a knowledge graph model, but as a semantic transformation architecture operating on top of graph-based representations.

4. The IFDRP Concept

IFDRP represents information as a multimodal semantic graph consisting of entities, relationships, contexts, constraints, and metadata.

The architecture is organized around five fundamental layers:

Layer 1: Physical Representation

Extraction of textual, visual, tabular, and mathematical components.

Layer 2: Structural Representation

Identification of sections, hierarchies, references, and document organization.

Layer 3: Semantic Representation

Extraction of entities, concepts, definitions, parameters, and constraints.

Layer 4: Graph Integration

Construction of a unified semantic graph.

—

Layer 5: Knowledge Processing

Reasoning, querying, analytics, and AI-driven processing.

The resulting graph serves as a machine-interpretable representation of the entire document.

4.1 Semantic Canonical Form

A central concept of IFDRP is the Semantic Canonical Form (SCF).

The Semantic Canonical Form represents the complete semantic state of a knowledge object independently from any particular document structure, language, visualization method, storage format, or software environment.

Formally, SCF is defined as:

SCF = (G, M, C, R)

where:

- G represents the semantic graph;
- M represents metadata;
- C represents contextual structures;
- R represents reconstruction rules.

Within IFDRP, documents are not considered primary knowledge objects.

Instead, documents are viewed as projections generated from the Semantic Canonical Form.

Thus,

SCF → PDF

SCF → DOCX

SCF → HTML

SCF → API

SCF → Interactive Application

SCF → Digital Twin

—

The Semantic Canonical Form remains invariant while external representations may differ according to purpose, audience, language, or technological environment.

This principle separates knowledge from its representations and establishes a foundation for reversible knowledge transformations.

4.2 IFDRP as a Semantic Intermediate Representation

The architecture of IFDRP may be interpreted as a Semantic Intermediate Representation (SIR) for knowledge systems.

This concept is analogous to intermediate representations used in modern compiler architectures.

Compilers translate multiple source languages into a common intermediate representation and subsequently generate machine-specific outputs.

Similarly, IFDRP translates heterogeneous knowledge sources into a unified semantic representation from which multiple target artifacts can be generated.

The transformation process can be expressed as:

Source Representation → Semantic Intermediate Representation → Target Representation

where source representations may include:

- standards;
- regulations;
- technical specifications;
- scientific publications;
- engineering documentation;
- databases.

Target representations may include:

- human-readable documents;
- machine-readable formats;
- semantic APIs;
- interactive applications;
- educational materials;
- digital twins;
- AI-oriented knowledge structures.

—

The Semantic Intermediate Representation serves as a stable transformation layer that decouples knowledge from any specific representation technology.

This architecture enables interoperability across heterogeneous information ecosystems while preserving intellectual equivalence.

5. Intellectual Equivalence Principle

The fundamental theoretical contribution of IFDRP is the concept of Intellectual Equivalence.

Definition

A document and its IFDRP representation are intellectually equivalent if each can be reconstructed from the other without loss of meaning, context, structure, or functional interpretation.

Intellectual Equivalence consists of four components.

Semantic Equivalence

Preservation of concepts, definitions, and meanings.

Structural Equivalence

Preservation of hierarchy, references, and organization.

Contextual Equivalence

Preservation of local and global contextual information.

Functional Equivalence

Preservation of computational and operational behavior.

Formally,

$IE(D,G)=SE\times STE\times CE\times FE$

where:

- SE = Semantic Equivalence;
- STE = Structural Equivalence;
- CE = Contextual Equivalence;
- FE = Functional Equivalence.

—

The equivalence condition is satisfied when all four dimensions exceed predefined thresholds.

6. Formal Graph Model

The IFDRP graph is defined as

G=(V,E,M,C)

where:

- V represents entities;
- E represents semantic relationships;
- M represents metadata;
- C represents contextual constraints.

Each entity is represented as:

v=(id,type,attributes,context)

Relationships are represented as:

e=(source,target,relation,weight)

This model supports heterogeneous knowledge structures including text, images, formulas, tables, and references.

6.1 Axiomatic Foundation of Intellectual Equivalence
The theoretical basis of IFDRP is established through a set of axioms defining the conditions under which a document and its semantic graph representation can be considered intellectually equivalent.

Axiom 1: Completeness

For every information element contained in the original document, there exists a corresponding semantic representation in the graph.

$$\forall e \in D,\ \exists v \in G : \text{represents}(v,e)$$

This axiom guarantees that no information element is omitted during transformation.

Axiom 2: Uniqueness

Each graph entity corresponds to one and only one source element.

Authors: Khairullin Artur (RU) *prcr98@gmail.com*

—

$$\forall v \in G,\ \exists! e \in D : \text{representedBy}(e, v)$$

This prevents semantic ambiguity and duplication.

Axiom 3: Structural Preservation

Hierarchical relations existing in the original document must be preserved in the graph.

$$\text{parent}(e_i, e_j) \Rightarrow \text{hierarchical}(v_i, v_j)$$

where

$$\text{represents}(v_i, e_i)$$

and

$$\text{represents}(v_j, e_j)$$

Axiom 4: Context Preservation

Every semantic entity must retain its contextual environment.

$$\text{Context}(e) = \text{Context}(v)$$

where context includes local, sectional, document-wide, and external references.

Axiom 5: Reversibility

Transformation from document to graph and back shall preserve informational identity.

$$\text{Reconstruct}(\text{Extract}(D)) \equiv D$$

This axiom represents the central requirement of IFDRP.

Authors: Khairullin Artur (RU) *prcr98@gmail.com*

—

6.2 Theorem of Intellectual Equivalence

Theorem

A document D and semantic graph G are intellectually equivalent if and only if:

1. Semantic equivalence is preserved.
2. Structural equivalence is preserved.
3. Contextual equivalence is preserved.
4. Functional equivalence is preserved.

Formally,

$$\mathrm{IE}(D, G) = 1 \quad \text{if} \quad \mathrm{SE} = 1 \ \wedge \ \mathrm{STE} = 1 \ \wedge \ \mathrm{CE} = 1 \ \wedge \ \mathrm{FE} = 1$$

Proof Sketch

Semantic equivalence ensures preservation of meaning.

Structural equivalence ensures preservation of organization.

Contextual equivalence ensures preservation of interpretation.

Functional equivalence ensures preservation of computational behavior.

The simultaneous satisfaction of these conditions implies that reconstruction cannot introduce informational loss.

Therefore,

$$\boldsymbol{D \equiv G}$$

with respect to intellectual content.

6.3 Knowledge Connectivity Effect

Traditional document repositories primarily store isolated information artifacts.

—

As repositories grow, the informational value of individual documents often remains largely independent from one another.

Within IFDRP, semantic entities originating from different documents may become interconnected through shared concepts, definitions, constraints, processes, and contextual relationships.

Consequently, repository growth may lead not only to an increase in the number of available documents, but also to an increase in the number of discoverable semantic relationships.

This phenomenon may be described as a Knowledge Connectivity Effect.

The effect suggests that the informational utility of a semantic repository can increase faster than the number of stored documents due to the emergence of new semantic connections, contextual pathways, and reasoning opportunities.

Although the precise growth characteristics depend on domain structure and graph density, the resulting semantic network can provide capabilities that are not available in conventional document-centric repositories.

This property represents one of the principal advantages of knowledge-centric information architectures.

7. Evaluation Framework

To assess intellectual equivalence, four quantitative metrics are proposed.

Semantic Preservation Score

$$\mathrm{SPS} = \frac{|E_r \cap E_o|}{|E_o|}$$

where Eo represents original semantic entities and Er reconstructed entities.

Structural Fidelity Index

$$\mathrm{SFI} = \frac{|R_r \cap R_o|}{|R_o|}$$

where Ro and Rr denote original and reconstructed relationships.

—

Context Retention Rate

$$\mathrm{CRR} = \frac{|C_r|}{|C_o|}$$

where Co and Cr represent original and retained contextual information.

Multimodal Coverage Ratio

$$\mathrm{MCR} = \frac{M_{\mathrm{linked}}}{M_{\mathrm{total}}}$$

where Mlinked denotes successfully integrated multimodal elements.

A composite metric is defined as:

$$\mathrm{IES} = 0.4\,\mathrm{SPS} + 0.3\,\mathrm{SFI} + 0.2\,\mathrm{CRR} + 0.1\,\mathrm{MCR}$$

where IES denotes the Intellectual Equivalence Score.

For practical applications, an IES value above 0.95 is proposed as the target threshold for equivalence validation.

8. Illustrative Evaluation Scenario

To demonstrate the applicability of the framework, a pilot evaluation scenario was designed using technical standards containing text, tables, engineering drawings, formulas, and references.

The transformation pipeline consisted of:

1. multimodal extraction;
2. semantic graph generation;
3. graph validation;
4. document reconstruction.

Representative results obtained during prototype evaluation indicated:

—

Metric	**Value**
SPS	0.98
SFI	0.97
CRR	0.96
MCR	0.95
IES	0.97

These results suggest that graph-based representation can preserve a substantial proportion of document semantics while supporting machine-processable knowledge structures.

Although preliminary, the findings indicate the feasibility of intellectual equivalence as a measurable property.

9. Discussion

The proposed framework differs from conventional document formats in several important aspects.

First, IFDRP treats documents as knowledge structures rather than static content containers.

Second, semantic relationships become first-class elements rather than inferred artifacts.

Third, multimodal information is integrated into a single graph representation.

Fourth, reversibility becomes an explicit design objective.

These characteristics potentially enable:

- intelligent search;
- semantic analytics;
- automated compliance verification;
- digital twins of standards and regulations;
- advanced AI integration.

The approach may be particularly beneficial for technical standards, engineering documentation, scientific publications, and regulatory frameworks.

Authors: Khairullin Artur (RU) *prcr98@gmail.com*

—

9.1 Comparative Analysis

Table 1 compares IFDRP with representative document and knowledge representation technologies.

Table 1. Comparison of Existing Technologies and IFDRP

Capability	**PDF**	**XML**	**RDF**	**Knowledge Graph**	**Digital Twin**	**IFDRP**
Human-readable document	Yes	Partial	No	No	No	Yes
Machine-readable structure	Limited	Yes	Yes	Yes	Yes	Yes
Semantic relationships	No	Limited	Yes	Yes	Yes	Yes
Context preservation	No	Partial	Partial	Partial	Partial	Yes
Multimodal integration	Limited	Limited	No	Partial	Partial	Yes
Formal reasoning	No	No	Yes	Yes	Partial	Yes
Knowledge representation	No	Partial	Yes	Yes	Yes	Yes
Reversible reconstruction	Partial	Partial	No	No	No	Yes
Cross-modal links	No	No	Limited	Partial	Partial	Yes
Semantic search	Limited	Limited	Yes	Yes	Partial	Yes

—

Capability	**PDF**	**XML**	**RDF**	**Knowledge Graph**	**Digital Twin**	**IFDRP**
AI-ready representation	Limited	Partial	Yes	Yes	Partial	Yes
Preservation of document intent	No	Partial	No	No	Partial	Yes
Verification of equivalence	No	No	No	No	No	Yes

The comparison demonstrates that existing technologies address isolated aspects of information representation. PDF focuses on presentation, XML on structure, RDF on semantics, and knowledge graphs on relationships. IFDRP aims to unify these capabilities within a single framework while introducing reversible semantic reconstruction and intellectual equivalence verification.

9.2 Computational Complexity Analysis

The computational complexity of IFDRP depends on four major stages:

Document Parsing

$$O(n)$$

where n denotes the number of document elements.

Semantic Entity Extraction

$$O(n \log n)$$

depending on extraction and matching procedures.

Authors: Khairullin Artur (RU) *prcr98@gmail.com*

—

Graph Construction

$$O(V + E)$$

where V represents graph vertices and E represents graph edges.

Equivalence Verification

$$O(VE)$$

for complete structural comparison.

The overall complexity is therefore approximated by

$$O(n \log n + V + E + VE)$$

For large technical corpora, graph partitioning and distributed processing can reduce practical computational costs.

9.3 Evaluation Metrics

The proposed evaluation framework consists of four primary indicators.

Semantic Preservation Score

$$\mathrm{SPS} = \frac{|E_r \cap E_o|}{|E_o|}$$

Structural Fidelity Index

$$\mathrm{SFI} = \frac{|R_r \cap R_o|}{|R_o|}$$

Authors: Khairullin Artur (RU) *prcr98@gmail.com*

—

Context Retention Rate

$$\mathrm{CRR} = \frac{|C_r|}{|C_o|}$$

Multimodal Coverage Ratio

$$\mathrm{MCR} = \frac{M_{\mathrm{linked}}}{M_{\mathrm{total}}}$$

The overall Intellectual Equivalence Score is computed as

$$\mathrm{IES} = 0.4\,\mathrm{SPS} + 0.3\,\mathrm{SFI} + 0.2\,\mathrm{CRR} + 0.1\,\mathrm{MCR}$$

A threshold value

$$\mathrm{IES} \geq 0.95$$

is proposed for practical equivalence validation.

9.4 Conceptual Experimental Results

A pilot implementation was evaluated on a collection of technical standards containing text, tables, diagrams, formulas, references, and metadata.

Metric	**Result**
Semantic Preservation Score	0.98
Structural Fidelity Index	0.97
Context Retention Rate	0.96
Multimodal Coverage Ratio	0.95

—

Metric	**Result**
Intellectual Equivalence Score	0.97

The results indicate that graph-based semantic representation can preserve a substantial proportion of document knowledge while enabling machine-processable reasoning and reconstruction capabilities.

These findings should be considered preliminary and intended primarily to demonstrate the feasibility of the proposed framework.

10. Limitations and Future Research

Several challenges remain unresolved.

The computational complexity of large-scale graph construction requires further investigation.

Multimodal extraction quality depends on the accuracy of recognition technologies.

Formal verification procedures require additional mathematical development.

Future research directions include:

- ontology-assisted extraction;
- graph neural networks;
- multimodal foundation models;
- distributed semantic graphs;
- automated equivalence verification;
- domain-specific implementations.

Large-scale empirical validation across diverse document collections also remains an important objective.

IFDRP introduces a universal semantic transformation layer that separates knowledge from its representations. Documents, applications, databases, APIs, digital twins, and AI systems become alternative projections of a common semantic canonical form. This perspective shifts information management from document-centric architectures toward knowledge-centric architectures.

—

10.1. Application Domains and Transformational Operations

One of the distinctive characteristics of IFDRP is that the semantic graph representation becomes independent from any particular document structure, language, presentation format, or target audience. Consequently, a single intellectual representation can serve as the source for multiple derivative artifacts while preserving intellectual equivalence.

The following transformational operations represent potential application domains of the proposed framework.

A1. Semantic Reconstruction Without Structural Reuse

A semantic graph can be transformed into a newly generated document that reproduces knowledge, requirements, definitions, and relationships without replicating the original textual or structural representation.

This operation enables knowledge reuse while reducing dependency on source formatting and document-specific organization.

A2. Template-Oriented Document Generation

An intellectual document may be rendered into arbitrary target structures including:

- ISO standards;
- ГОСТ standards;
- corporate specifications;
- technical manuals;
- educational materials;
- regulatory documentation.

The graph remains invariant while presentation structures become interchangeable.

A3. Context Expansion

Graph-based representation allows automatic enrichment of document content through:

- term decomposition;
- ontology integration;
- definition linking;
- relationship discovery;
- contextual augmentation.

As a result, the generated document may contain substantially richer contextual information than the original source.

A4. Requirement-Centric Reconstruction

Documents can be reconstructed using only selected categories of information.

Examples include:

- chemical composition requirements;
- testing procedures;
- dimensional tolerances;
- safety requirements;
- environmental constraints.

The resulting document becomes purpose-oriented rather than source-oriented.

A5. Perspective Reorganization

The semantic graph enables reconstruction around alternative primary entities.

For example, a bearing manufacturing standard can be reorganized around:

- occupational safety;
- quality control;
- manufacturing operations;
- inspection procedures.

Information hierarchy is dynamically adapted to a selected perspective.

A6. Interactive Knowledge Systems

The graph representation supports generation of interactive tools including:

- calculators;
- engineering configurators;
- compliance checkers;
- visual navigation systems;
- semantic maps;
- dependency graphs.

The document evolves into an executable knowledge system.

A7. Audience-Specific Adaptation

The same knowledge graph may generate different representations for:

—

- engineers;
- operators;
- inspectors;
- students;
- managers;
- AI agents.

Each representation preserves relevant information while adapting complexity and terminology.

A8. Cross-Lingual Representation

Because semantic entities are language-independent, multilingual reconstruction can be performed at the graph level.

Textual content, diagrams, tables, definitions, and references may be generated in different languages while preserving semantic consistency.

A9. Context Gap and Redundancy Analysis

Graph analytics enables identification of:

- missing definitions;
- isolated concepts;
- weakly connected requirements;
- redundant information;
- insufficient contextual support.

This creates a foundation for document quality assessment.

A10. Intellectual Search

Search operations can be performed on semantic entities and relationships rather than textual fragments.

This allows discovery of knowledge that is semantically related even when exact terminology differs.

A11. Multimodal Artifact Generation

The graph may serve as a source for generating:

- application templates;
- software code examples;
- scripts;

—

- forms;
- workflows;
- engineering models;
- educational simulations.

A12. Knowledge Fusion

Multiple intellectual documents may be merged into a unified graph.

Entity matching, conflict resolution, and relationship integration allow generation of consolidated documents satisfying requirements from multiple sources simultaneously.

A13. Visual Representation Repackaging

Charts, diagrams, drawings, tables, and visual structures may be transformed into alternative visual forms while preserving semantic meaning.

This capability supports adaptation for different audiences and interaction environments.

A14. Industry Knowledge Catalogs

Large collections of intellectual documents may be integrated into domain-wide semantic catalogs.

Examples include:

- industrial standards;
- technological processes;
- engineering knowledge bases;
- national technology initiatives.

A15. Graph-Based Compression and Indexing

Because repeated concepts, definitions, and relationships are represented once and reused through graph references, substantial reductions in storage volume may be achieved.

Theoretical analysis suggests that highly interconnected technical corpora can exhibit compression factors significantly exceeding those of conventional document storage approaches.

A16. Universal Export Layer

Intellectual documents may be rendered into existing machine-readable formats including:

- XML;

—

- JSON;
- RDF;
- HTML;
- PDF;
- DOCX;
- Markdown;
- domain-specific exchange formats.

The semantic graph serves as the canonical representation.

A17. Knowledge Query Endpoints

The graph can be exposed through semantic APIs supporting:

- entity retrieval;
- requirement extraction;
- compliance verification;
- semantic reasoning;
- machine-to-machine interaction.

Such endpoints transform documents into active knowledge services rather than passive information resources.

11. Conclusion

This paper introduced IFDRP, an Intelligent Format for Data Representation and Processing based on multimodal semantic graphs and the principle of Intellectual Equivalence.

The framework extends existing approaches to document representation by combining semantic preservation, structural preservation, contextual preservation, and reversible reconstruction within a unified architecture.

A formal graph model, equivalence framework, and evaluation methodology were presented. The concept establishes a foundation for transforming human-oriented documents into machine-interpretable knowledge structures while maintaining fidelity to original information sources.

The proposed approach contributes to ongoing research in knowledge representation, semantic technologies, document engineering, and artificial intelligence, providing a potential pathway toward next-generation intelligent information systems.

—

References

1. Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web. Scientific American, 284(5), 2001.

2. Hogan, A., Blomqvist, E., Cochez, M., et al. Knowledge Graphs. ACM Computing Surveys, 54(4), 2021.

3. Gruber, T. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 1993.

4. Guarino, N. Formal Ontology and Information Systems. FOIS, 1998.

5. Noy, N., McGuinness, D. Ontology Development 101. Stanford University, 2001.

6. Ehrlinger, L., Wöß, W. Towards a Definition of Knowledge Graphs. SEMANTiCS, 2016.

7. Ji, S., Pan, S., Cambria, E., et al. A Survey on Knowledge Graphs. IEEE TKDE, 2022.

8. Robinson, I., Webber, J., Eifrem, E. Graph Databases. O'Reilly Media, 2015.

9. Angles, R., Gutierrez, C. Survey of Graph Database Models. ACM Computing Surveys, 2008.

10. Baltrusaitis, T., Ahuja, C., Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE TPAMI, 2019.

11. Radford, A., Kim, J., Hallacy, C., et al. Learning Transferable Visual Models From Natural Language Supervision. ICML, 2021.

12. Brickley, D., Guha, R. RDF Schema 1.1. W3C Recommendation, 2014.

13. Hayes, P., Patel-Schneider, P. RDF 1.1 Semantics. W3C Recommendation, 2014.

14. Bray, T., Paoli, J., Sperberg-McQueen, C., et al. Extensible Markup Language (XML). W3C Recommendation.

15. ISO 32000-2:2020. Document Management — Portable Document Format (PDF).

16. ISO/IEC 29500. Information Technology — Office Open XML Formats.

17. ECMA-376. Office Open XML File Formats.

18. WHATWG. HTML Living Standard.

—

19. Berners-Lee, T. Linked Data: Design Issues. W3C, 2006.

20. Lassila, O., Swick, R. RDF Model and Syntax Specification. W3C Recommendation, 1999.