Sciences du Nord Nature Science and Technology
Vol. 1, No. 02, December 2024: pp. 113-125
E-ISSN: 3046-6504
P-ISSN: 3047-4922
DOI: https://doi.org/10.71238/snnst.v1i02.225

Policy-Bound Semantic Data Meshes for AI-Ready Enterprise Knowledge Foundations: A Systematic Review

Shivakrishna Vangala [1]
[1] Principal Architect, T-Mobile USA Inc., Seattle, United States of America
Corresponding Author: shivakrishna.vangala@gmail.com

ARTICLE HISTORY
Received Oct, 2024 **Revised Nov, 2024** **Accepted Dec, 2024**

ABSTRACT
Semantically rich, policy-aware, and organizationally distributed data forms an essential foundation for enterprise artificial intelligence (AI) initiatives. To define policy-bound semantic data meshes, a systematic review of 28 sources covering fields such as ontology engineering, linked data, knowledge graphs, data governance, access control, data-centric AI and responsible AI documentation was carried out. The corpus was themed into six themes, coded to five governance decision domains and mapped to the four data mesh principles. The scale of the industry knowledge graph, data cascades and the mention of ethical principles in guideline documents were compared quantitatively across sources. The results showed that 92% of the practitioners who were interviewed experienced at least one data cascade, transparency was found in 73 out of 84 ethics guidelines, and 15 FAIR principles were found to be testable to demonstrate interoperability. With industry knowledge graphs, it's possible to integrate up to 1 to 2 billion entities, making integration at scale possible. A layered reference architecture is derived, where the ontologies, machine-readable policies and documentation artefacts are bound to each domain data product; a formal decision rule with deny-overrides semantics and a composite readiness index is proposed. The novelty is that three distinct research streams dealing with knowledge at the enterprise level are merged together in one formulation of governance by design for AI-ready enterprise knowledge.

Keywords: *Access Control; Data Governance; Data Mesh; Data-Centric AI; FAIR Principles; Knowledge Graph; Linked Data; Responsible AI; Semantic Web*

INTRODUCTION

Model performance is the most common metric used to evaluate AI systems, but most of the errors in AI systems come from the models' assumptions about the data they rely on. Ninety-two percent of the interviewees from India, East and West Africa and the United States experienced at least one data cascade (an escalating downstream failure that began with an upstream data issue), and nearly half of them (49.2%) experienced two or more cascades [1]. Such cascades were described as "widespread, late and avoidable", suggesting that preventive actions are better than downstream corrections. In high-stakes AI, data work is thus defined as the main source of unnecessary risks.

While many such operations are considered preparatory steps in a deep learning pipeline, data collection, validation, cleaning, and labelling are known as persistent bottlenecks, and are now seen as factors that affect the quality of models, rather than preliminary operations [2]. Scale is not enough, there is a danger of noise, bias and inconsistency in poorly described and heterogeneous data which larger models can't correct for. These tasks are further compounded by factors in enterprise settings, such as the integration of formats, semantics and quality of information sources, and the governance of information acquisition and validation becomes as crucial as the choice of modelling.

Data-centric approach to AI moves focus away from designing algorithms towards systematic engineering of data. This perspective is mapped on to three goals: training data development, inference data development and data maintenance [3]. Each goal assumes to be data discoverable, well described and governed – which is rarely supported in enterprises with multiple data owners. As such, the quality and documentation of data assets are regarded as engineering issues which can be measured, versioned, and improved, and this is the principle used in the readiness measures proposed in the following.

Data mesh is a proposed architecture solution to centralized data platforms. These principles form the basis for the paradigm: domain-oriented ownership, data as a product, self-serve infrastructure, and federated computational governance [4]. Without shared meaning and enforceable policies, however, without centralization it can be easy to create disparate data products that can't be merged, trusted, or legally reused. Shared infrastructure and common governance will therefore ensure consistency across the organisation and domain teams become accountable for published products.

Data governance has been defined as the practice of authority and control over data management and a structured review of 145 publications has resulted in the development of the six dimensions and 15 research questions [5]. Notwithstanding this width, semantic modelling, machine-readable policy and AI documentation are rarely considered to be one design problem. The aim of the present review is therefore to synthesize 28 sources into a policy bound semantic data mesh formulation and compare the evidence quantitatively and to establish a reference architecture, a decision rule and readiness metrics. The review is then presented as a synthesis and not as an experimental study.

LITERATURE REVIEW

2.1 Semantic Foundations

Semantic data management's conceptual scaffolding is given by ontologies. An ontology explicitly specifies a conceptualization that enables the sharing and reusability of knowledge across systems with a shared vocabulary [6]. With no schema control in a data mesh, a language enables a consistent interpretation of an independently owned data product. Formal vocabularies of this sort also make assumptions explicit, decreasing ambiguity if the data is re-used by machine learning pipelines and automated reasoners.

The Semantic Web evolved the Web from linked documents to machine-interpretable data, and it includes structured markup, ontologies and software agents that interact with published data [7]. This vision was the foundation for the subsequent development of enterprise knowledge graphs and policy-aware reasoning on top of the layered technical stack. One of the core principles was to produce metadata with the data, so that software could integrate data from many sources without writing any extra integration code.

Linked Data was implemented by four publishing rules: URIs are used as names, HTTP URIs are used for lookup, useful information is provided via standards like RDF and SPARQL, and links to other URIs are included [8]. These rules are clearly applicable to mesh interoperability, as they outline the ways in which independent producers make connectable data available. Since identifiers are dereferenceable and typed links are explicit, consumers can traverse from one data product in this domain to another related data product without relying on a central coordinator.

The rules were put into practice in DBpedia to extract structured information from Wikipedia. There were templates that guided the extraction of the semi-structured article content, and the extracted identifiers were used as anchor points for links between the data and other data. The first version contained 1.95 million objects in 103 million RDF triples, and when linked with other open datasets the resulting network of RDF triples was approximately 2 billion [9]. Table 1 provides a correspondence between the six themes in the literature and the concepts that they bring.

Table 1. Literature-to-Concept Map of the Reviewed Corpus

Theme	Sources (n, %)	Contributed Concepts	References
Semantic Web and linked data	7 (25.0%)	Ontology, URI-based publishing, RDF links, open datasets, decentralized publication	[6]–[10], [14], [17]
Knowledge graphs and integration	4 (14.3%)	Graph models, industry-scale graphs, global schema, ontology matching	[11], [12], [15], [16]
Governance and architecture	4 (14.3%)	Data mesh principles, governance dimensions, decision domains, FAIR principles	[4], [5], [13], [22]
Access control and policy	4 (14.3%)	RBAC, ABAC, deontic policy languages, privacy policies	[18]–[21]
Data-centric AI	3 (10.7%)	Data cascades, data quality, training and inference data	[1]–[3]
Responsible AI	6 (21.4%)	Datasheets, model cards, ethics principles, harms, hallucination	[23]–[28]

Note. Percentages are shares of the 28 sources. Source: Synthesized from the reviewed sources.

2.2 Knowledge Graphs and Industry-Scale Integration

Wikidata has added a multilingual, collaboratively edited knowledge base, where statements can have references, and are curated by large communities [10]. Provenance and community curation are important for governance as trust in a statement is dependent on the traceability of the origin. Shared identifiers and labels in multiple languages also enable a single concept to be mentioned in the same way within teams that communicate in different languages.

It has been proven to be viable in the real world at scale. Table 2 shows that the Google knowledge graph consists of 1 billion entities asserted in 70 billion claims, and the Microsoft graph has approximately 2 billion main entities and 55 billion facts [11]. The main requirements of such graphs are always reported as coverage, correctness, structure, and freshness. Five organizations demonstrate this with the experience of building a graph over time, in addition to periodic data loads.

Table 2. Scale Comparison of Open and Industry Knowledge Graphs

System	Entities	Assertions	Assertions per Entity (Derived)
DBpedia (initial release)	1.95 million	103 million triples	≈ 52.8
Google Knowledge Graph	≈ 1 billion	70 billion	70.0
Microsoft graph	≈ 2 billion	≈ 55 billion	≈ 27.5
IBM Watson KG framework	> 100 million	> 5 billion relationships	Not derived
eBay product graph	Not reported	> 1 billion triples	Not derived
Facebook graph	≈ 50 million primary	Not reported	Not derived

Note. The derived ratio equals assertions divided by entities. Source: Synthesized from the reviewed sources.

Systematic analysis of knowledge graphs with respect to graph data models, schema, identity, context, deductive and inductive knowledge, publication and quality have also been conducted, including enterprise and open examples [12]. Graph based representation is thus set up as a common ground for data of different types, and the orders of magnitude disparities between the systems are depicted in Figure 1. The quality, refinement, and publication are considered lifecycle activities that are expected of the mesh data product as well.

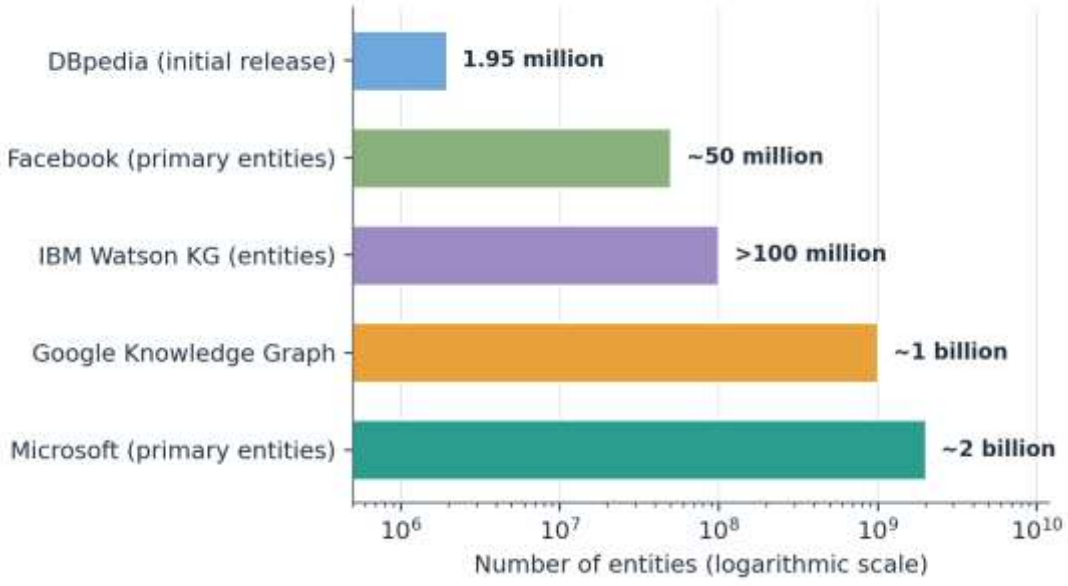

Figure 1. Entity Counts of Open and Industry Knowledge Graphs

Source: Synthesized from the reviewed sources. Values are approximate.

METHODS

A closed corpus systematic review design was used, which consisted of 28 sources provided and excluded any additional external sources. All sources were mapped and coded to the five decision domains of data governance, which are data principles, data quality, metadata, data access, and data lifecycle [13] and then mapped to the mesh capabilities, as shown in Figure 2. For each

theme, the share of sources (n) was calculated, by using Equation (1) with the number of sources in theme k (nk) and the number of sources in the whole corpus (N = 28).

$$p_k = \frac{n_k}{N} \times 100 \quad (1)$$

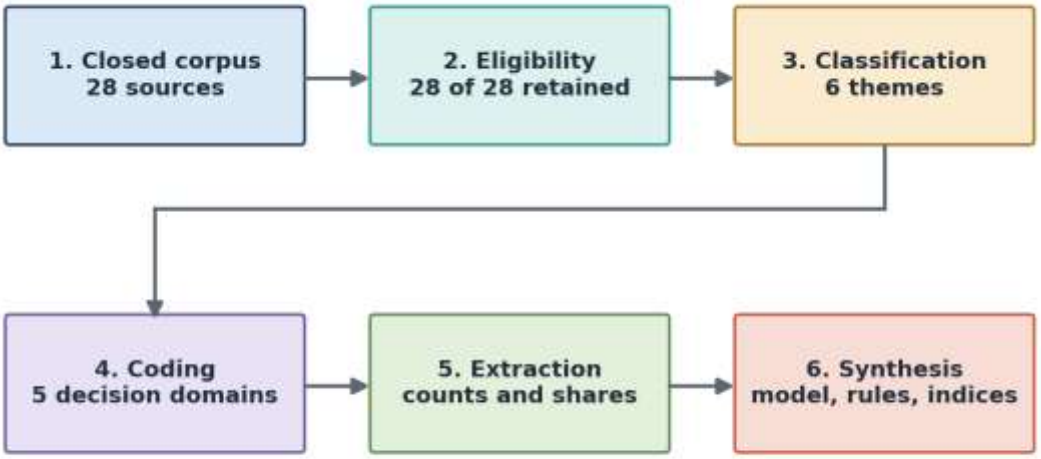

Figure 2. Workflow of the Systematic Synthesis

Source: Synthesized from the reviewed sources.

Thematic organization was provided by previous narrative reviews in the semantic web area [14] and, in the current protocol, explicit inclusion criteria, thematic coding, and quantitative extraction were introduced. Sources qualified where peer reviewed or institutional and were also relevant to semantics, policy or governance, or AI data practice; all 28 sources met the criteria. By period, 7 sources (25.0%) predate 2010, 6 (21.4%) fall between 2010 and 2018, and 15 (53.6%) were published from 2019 onward, as shown in Figure 3. The corpus therefore covers three decades, from 1993 to 2023, which makes it possible to trace the evolution of the field

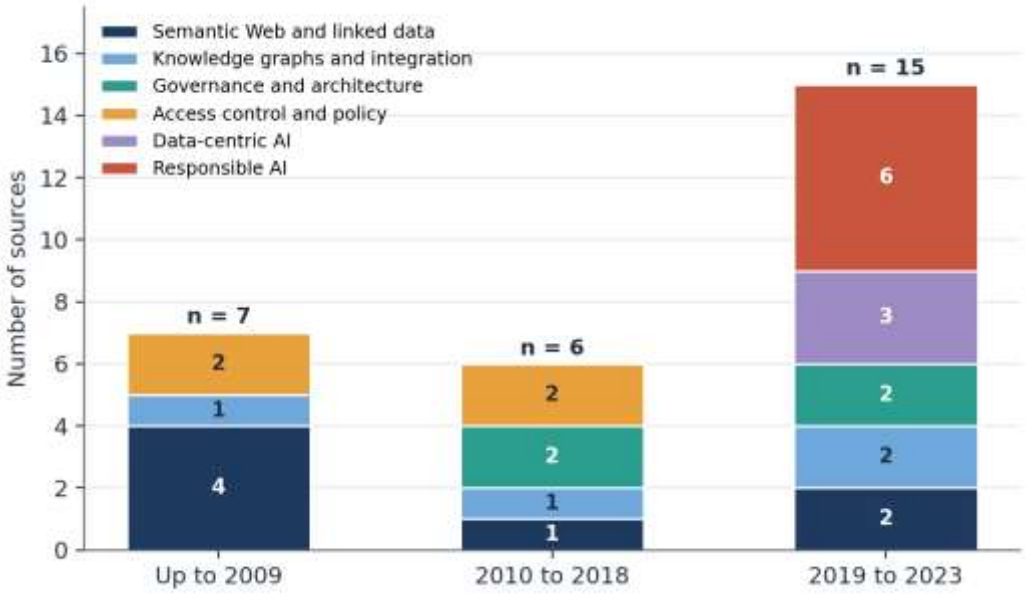

Figure 3. Composition of the Corpus by Theme and Period
Source: Synthesized from the reviewed sources. Counts sum to 28.

RESULTS AND DISCUSSION

4.1 Semantic Integration Layer

The formalization of data integration as a system consisting of a global schema G, source schemas S, and a mapping M between them, with query answers defined as the intersection of results over all global databases that are consistent with the sources, as expressed in Equation (2) [15] was realized. Two kinds of mappings, global-as-view and local-as-view were distinguished, and the type of mapping determines the ease of adding new domain sources.

$$q^{I,D} = \bigcap_{B \in sem(I,D)} q^{B} \quad (2)$$

The biggest challenge to such integration so far is the heterogeneity of independently created ontologies. The open challenges for ontology matching were discussed in [16] which included scalability, user involvement and evaluation. This means that alignment should be considered a managed lifecycle activity in the mesh, and not a one-off migration activity. Automated matchers are then expected to make proposals, and domain experts are proposed to validate them, to allow mappings to be trusted as ontologies evolve.

Later, decentralized publication was suggested for LD, and data was maintained in the hands of its producers and retrieved by federation closely related to domain ownership in data mesh [17]. With such an arrangement, the queries are broken down into pieces and sent to the domain endpoints which prevents duplication and staleness of the data that would otherwise be replicated at a central location. Table 3 maps the four principles of the mesh with the semantic and policy mechanisms that were identified from the literature reviewed.

Table 3. Mesh Principles and Their Semantic and Policy Enablers

Mesh Principle	Semantic Mechanism	Policy-Binding Mechanism	Indicator
Domain-oriented ownership	Domain ontology published by the owning domain	Ownership and stewardship attributes in policies	Domains with a published ontology
Data as a product	Metadata, FAIR scoring, datasheet	Usage terms bound to each product	FS and DC
Self-serve platform	Shared catalog and alignment services	Enforcement points offered as platform services	Share of products registered
Federated computational governance	Global schema and mapping lifecycle	Deny-overrides evaluation of prohibitions and permissions	PC and decision logs

Note. FS = FAIR score; DC = documentation completeness; PC = policy coverage. Source: Synthesized from the reviewed sources.

4.2 Policy Binding Layer

In the case of role-based access control (RBAC), the RBAC96 family of four models was created, from base models to role hierarchies and constraints, to a consolidated model [18]. A user u has permission p if the role r assigned to u also has permission p, as specified in Equation (3) (e.g.,

the user assignment relation UA, the permission assignment relation PA). Constraints enable separation of duty and hierarchies enable inheritance of permissions by senior roles. The method is easy to administer with roles, but not very expressive to share context-dependent data.

$$auth(u,p) \Leftrightarrow \exists\, r \in R : (u,r) \in UA \wedge (p,r) \in PA \quad (3)$$

Attribute based access control (ABAC): The access control is based on the attributes of the subject, attributes of the object, and the environment conditions, in which the access control policies are defined by their attributes [19]. The decision is encoded in Equation (4), where op is the requested operation and As, Ao and Ae are the three sets of attributes. Since decisions are based on attributes and not on predetermined lists of users, new data products can be protected via policy without listing all authorized users. In this review, we explore various approaches to access control, as detailed in Table 4.

$$d(s,o,op,e) = \Pi(A_s, A_o, op, A_e),\ d \in \{Permit, Deny\} \quad (4)$$

Table 4. Comparison of Access Control Approaches

Approach	Authorization Basis	Structural Elements	Strength	Limitation
RBAC (RBAC96)	Roles assigned to users and permissions	Four models, RBAC0 to RBAC3	Administrative simplicity	Limited contextual expressiveness
ABAC	Attributes of subject, object, and environment	Three attribute sources plus operation	Fine-grained, context-aware decisions	Attribute management overhead
Semantic policy language	Deontic rules over ontologies	Four deontic concepts	Obligations and prohibitions with reasoning	Requires ontology and reasoning support

Source: Synthesized from the reviewed sources.

The extension of access control with deontic concepts is what constitutes a semantic policy language. A rights-based policy language, based on rights, prohibitions, obligations and dispensations, was developed for pervasive computing, policies expressed using ontologies and interpreted by reasoning [20]. For a mesh, a deny-overrides rule is proposed, in which a request q is permitted only when no applicable prohibition exists and at least one applicable permission exists, as expressed in Equation (5) and illustrated in Figure 4, where Π^- and Π^+ denote the sets of prohibitions and permissions and $m(\pi, q)$ equals 1 when policy π applies to q.

$$D(q) = Permit \Leftrightarrow [\,\forall\, \pi \in \Pi^- : m(\pi,q) = 0\,] \wedge [\,\exists\, \pi \in \Pi^+ : m(\pi,q) = 1\,] \quad (5)$$

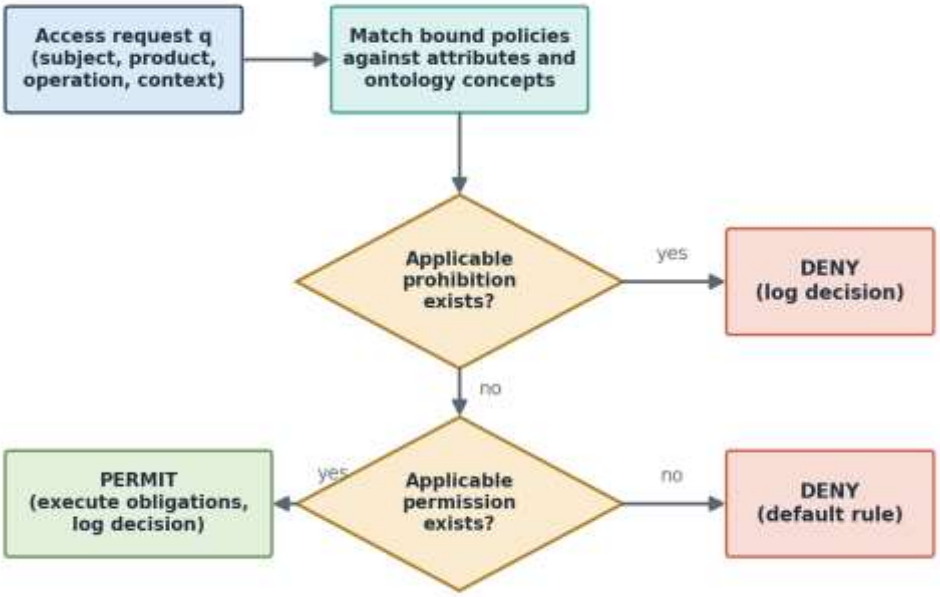

Figure 4. Deny-Overrides Decision Flow for Policy-Bound Requests

Source: Synthesized from the reviewed sources.

Policy issues and solutions using semantic web technologies were reviewed, which has implications for expressing and reasoning about policies via semantic technologies [21]. By combining these observations, a reference architecture was formed, as shown in Fig. 5, in which each domain data product was bundled with data and an ontology, while a policy decision point (PDP) and a policy enforcement point (PEP) were added to every request from an AI consumer. Evaluation of the policy is thus decoupled from data storage, so that the policy can be changed without changing the data.

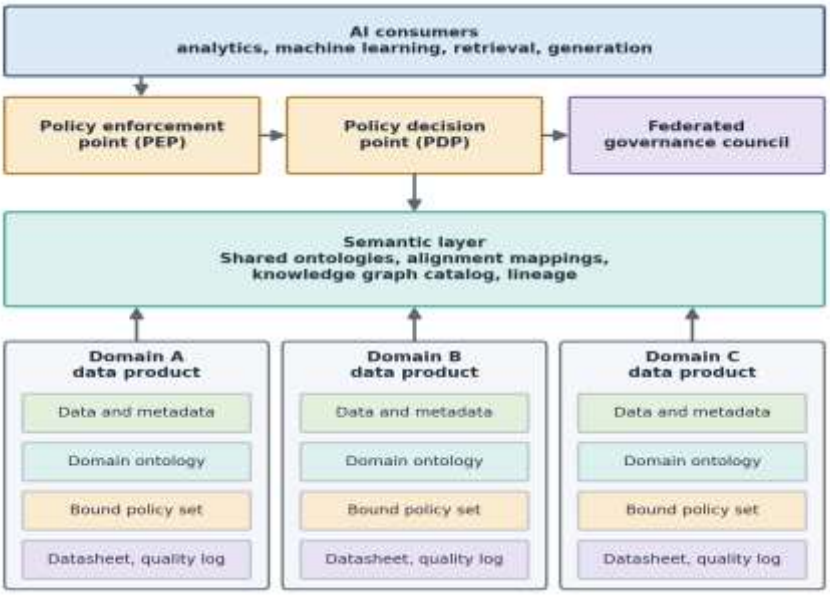

Figure 5. Reference Architecture of a Policy-Bound Semantic Data Mesh

Source: Synthesized from the reviewed sources.

4.3 Stewardship and Documentation Layer

The FAIR principles are divided into four groups: findability, accessibility, interoperability, reusability, of which four, three, and four principles are named respectively [22]. Each principle is a claim that can be tested and therefore a data product can be scored using Equation (6) where the ci are set to 1 if the principle is met and set to 0 if it is not. For instance, findability relies on the use of persistent identifiers and rich metadata, while interoperability relies on shared vocabularies and qualified references. The distribution of principles is shown in Figure 6.

$$FS = \frac{1}{15} \sum_{i=1}^{15} c_i, \; c_i \in \{0,1\} \tag{6}$$

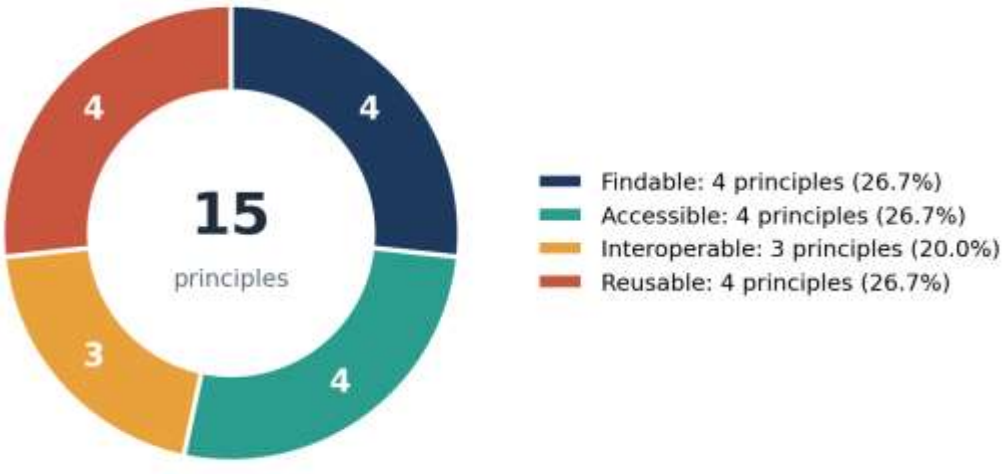

Figure 6. Distribution of the 15 FAIR Principles Across Four Groups

Source: Synthesized from the reviewed sources.

Programs for datasets suggest a series of questions grouped in seven sets that are related to the motivation, composition, collection process, preprocessing, uses, distribution, and maintenance of the data [23]. The datasheet is a human-readable data provenance that can be attached to a data product to fill documentation gaps that were identified as triggers of data cascades, while supporting the machine-readable metadata. Questions that address aspects of collection and preprocessing record decisions that most strongly influence model behavior downstream.

There are nine reporting sections specified in model cards: model details, intended use, factors, metrics, evaluation data, training data, quantitative analyses, ethical considerations, and caveats [24]. Documentation of intended use and evaluation factors should be used to keep models out of hand if they are used outside of the conditions they were evaluated for. The artefacts that address the data and model aspects of the AI supply chain are found in Table 5.

Table 5. Stewardship and Documentation Artefacts Compared

Artefact	Object Documented	Structure	Purpose	Binding in the Mesh
FAIR principles	Data and metadata	15 principles in 4 groups	Findability and reuse by machines	Scored per product (FS)
Datasheets	Datasets	7 question groups	Provenance and appropriate use	Attached to each data product
Model cards	Trained models	9 reporting sections	Transparent model reporting	Linked to consuming models
Governance decision domains	Enterprise data assets	5 decision domains	Organizational accountability	Defined by the governance council

Source: Synthesized from the reviewed sources.

4.4 Ethics, Risk, and Readiness

Of the 84 ethics guideline documents that were reviewed, 11 clusters of ethics principles were identified with transparency present in 73 documents (86.9%), justice and fairness in 68 documents (81.0%), non-maleficence in 60 documents (71.4%), and responsibility in 60 documents (71.4%), while privacy was present in 47 documents (56.0%) [25]. The distribution of this principle is given in Equation (7) with Nd = 84, and is displayed in Figure 7. Transparency is a prominent characteristic, reflecting the fact that the expectations for traceability and explanation are the most widely shared of all guideline authors.

$$f_k = \frac{n_k}{N_d},\ N_d = 84 \tag{7}$$

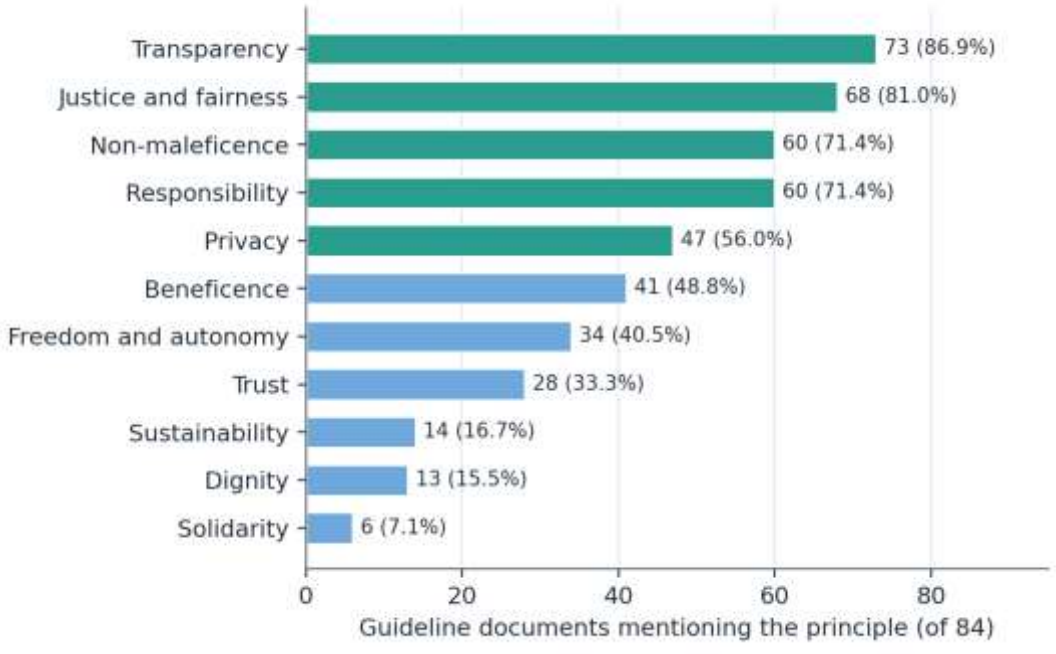

Figure 7. Frequency of Ethical Principles in 84 AI Guideline Documents

Source: Synthesized from the reviewed sources.

The principles were distilled into five key principles, which are: beneficence, non-maleficence, autonomy, justice, and explicability [26]. Four of these principles are based on principles in bioethics, and explicability is introduced to account for intelligibility and accountability of AI systems. Table 6 maps these principles to the guideline themes and to mechanisms that can be used to implement them, including prohibitions for non-maleficence and lineage for explicability.

Table 6. Crosswalk Between Ethical Principles and Mesh Mechanisms

Distilled Principle	Guideline Theme (Documents)	Mesh Mechanism
Beneficence	Beneficence (41; 48.8%)	Declared intended use for each data product
Non-maleficence	Non-maleficence (60; 71.4%)	Prohibitions evaluated before permissions
Autonomy	Freedom and autonomy (34; 40.5%)	Consent and usage-control attributes in policies
Justice	Justice and fairness (68; 81.0%)	Documentation and audit of data composition
Explicability	Transparency (73; 86.9%)	Lineage, decision logs, datasheets, and model cards

Note. Responsibility (60 documents) and privacy (47 documents) are addressed by the governance council and by access policies, respectively. Source: Synthesized from the reviewed sources.

The risks were explored for large language models, such as environmental costs, financial costs, the fact that the data used to train them cannot always be documented, and encoded bias, and it was suggested to be careful in curating and documenting data rather than indiscriminately collecting it [27]. This is the argument for attaching documentation and policy to the product prior to the inclusion of data into model training. If it is not so controlled, then the harm is passed down unnoticed from the large and inscrutable corpora.

The first type of hallucination is when the generated content is unfaithful, while the second type is when there is no evidence of source material [28] that reported the produced content. Using knowledge graphs with governed and provenance information is therefore a plausible mitigation. The composite index in Equation (8) is suggested; it makes readiness measurable, with PC being the proportion of data products that have been 'bound' to machine-readable policy and DC being the proportion of the required documentation items that have been completed, both items being non-negative and the sum of the weights being one. The principal challenges, evidence and mitigations are summarized below in Table 7.

$$R_j = w_F \cdot FS_j + w_P \cdot PC_j + w_D \cdot DC_j,\ w_F + w_P + w_D = 1 \quad (8)$$

Table 7. Challenges, Evidence, and Mitigations in Policy-Bound Meshes

Challenge	Evidence	Mitigation in the Design
Data cascades	92% prevalence among 53 practitioners	Documentation and quality logs bound to products
Schema heterogeneity	Scalability and evaluation of ontology matching remain open	Shared ontologies and a governed alignment lifecycle
Policy fragmentation	Three families of access models coexist	Deny-overrides evaluation over unified policies
Graph scale and freshness	Graphs of 1 to 2 billion entities require coverage, correctness, and freshness	Incremental publication under decentralized ownership
Opacity and hallucination	Intrinsic and extrinsic hallucination categories	Provenance-rich grounding and lineage
Documentation cost	Training data too large to be documented	Documentation at product level before training

Source: Synthesized from the reviewed sources.

CONCLUSION

After systematic synthesis of 28 sources, policy-bound semantic data meshes were created and the evidence is coming together to give a clear answer: decentralized data products need shared ontologies for meaning, machine-readable policies for enforcement, and structured documentation for accountability. Quantitative evidence includes knowledge graphs of 1 to 2 billion entities, data cascades present in 92% of all datasets and 15 testable FAIR principles. A layered architecture was obtained, a deny-overrides decision rule was obtained and a composite readiness index was obtained. The main drawback is that they are based on a closed corpus and the weights of indices have not been validated empirically. Future studies are recommended to assess policy coverage, FAIR scores, and documentation completeness of the operating meshes, as well as to create benchmark for ontology alignment and to create evaluation costs for policy assessment. In summary, the synthesis provides an effective answer to the risks to data quality that prevent enterprises from full AI benefits: governance by design, meaning and policy at the same pace as data.

REFERENCES

[1] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445518.

[2] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: A data-centric AI perspective," *VLDB J.*, vol. 32, no. 4, pp. 791–813, 2023, doi: 10.1007/s00778-022-00775-9.

[3] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, and X. Hu, "Data-centric AI: Perspectives and challenges," in *Proceedings of the 2023 SIAM International Conference on Data Mining*, 2023, pp. 945–948. doi: 10.1137/1.9781611977653.ch106.

[4] I. A. Machado, C. Costa, and M. Y. Santos, "Data mesh: Concepts and principles of a paradigm shift in data architectures," *Procedia Comput. Sci.*, vol. 196, pp. 263–271, 2022, doi: 10.1016/j.procs.2021.12.013.

[5] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," *Int. J. Inf. Manage.*, vol. 49, pp. 424–438, 2019, doi: 10.1016/j.ijinfomgt.2019.07.008.

[6] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993, doi: 10.1006/knac.1993.1008.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001, doi: 10.1038/scientificamerican0501-34.

[8] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," *Int. J. Semant. Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009, doi: 10.4018/jswis.2009081901.

[9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *The Semantic Web (Lecture Notes in Computer Science, vol. 4825)*, vol. 4825, 2007, pp. 722–735. doi: 10.1007/978-3-540-76298-0_52.

[10] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014, doi: 10.1145/2629489.

[11] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges," *Commun. ACM*, vol. 62, no. 8, pp. 36–43, 2019, doi: 10.1145/3331166.

[12] A. Hogan, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, 2021, doi: 10.1145/3447772.

[13] V. Khatri and C. V Brown, "Designing data governance," *Commun. ACM*, vol. 53, no. 1, pp. 148–152, 2010, doi: 10.1145/1629175.1629210.

[14] P. Hitzler, "A review of the semantic web field," *Commun. ACM*, vol. 64, no. 2, pp. 76–83, 2021, doi: 10.1145/3397512.

[15] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2002, pp. 233–246. doi: 10.1145/543613.543644.

[16] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 158–176, 2013, doi: 10.1109/TKDE.2011.253.

[17] A. Polleres, M. R. Kamdar, J. D. Fernández, T. Tudorache, and M. A. Musen, "A more decentralized vision for Linked Data," *Semant. Web*, vol. 11, no. 1, pp. 101–113, 2020, doi: 10.3233/SW-190380.

[18] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," *Computer (Long. Beach. Calif).*, vol. 29, no. 2, pp. 38–47, 1996, doi: 10.1109/2.485845.

[19] V. C. Hu, "Guide to attribute based access control (ABAC) definition and considerations," 2014. doi: 10.6028/NIST.SP.800-162.

[20] L. Kagal, T. Finin, and A. Joshi, "A policy language for a pervasive computing environment," in *Proceedings of the 4th IEEE International Workshop on Policies for Distributed Systems and Networks*, 2003, pp. 63–74. doi: 10.1109/POLICY.2003.1206958.

[21] S. Kirrane, S. Villata, and M. d'Aquin, "Privacy, security and policies: A review of problems and solutions with semantic web technologies," *Semant. Web*, vol. 9, no. 2, pp. 153–161, 2018, doi: 10.3233/SW-180289.

[22] M. D. Wilkinson, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 160018, 2016, doi: 10.1038/sdata.2016.18.

[23] T. Gebru, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[24] M. Mitchell, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229. doi: 10.1145/3287560.3287596.

[25] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019, doi: 10.1038/s42256-019-0088-2.

[26] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," *Harvard Data Sci. Rev.*, vol. 1, no. 1, 2019, doi: 10.1162/99608f92.8cd550d1.

[27] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[28] Z. Ji, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, 2023, doi: 10.1145/3571730.