DATA CANON

2026-03-18

Dexter Hadley, MD/PhD

Hadley Lab CANONIC

Abstract

Example

Contents

Data provenance MUST be traceable. Every data element has a documented origin, transformation history, and current state.

Example: Lineage tracking source system, extraction timestamp, transformation logic, load destination. Metadata standards: Dublin Core (15 elements for resource description), DCAT (Data Catalog Vocabulary, W3C) for dataset discovery, Schema.org for structured web data. Data catalogs (enterprise metadata management) provide searchable inventories of organizational data assets. W3C PROV (Provenance) ontology formalizes provenance relationships: Entity (data), Activity (transformation), Agent (actor). Data mesh architecture distributes provenance responsibility to domain teams.

0.1 2. Data Quality

Data quality MUST be validated against defined dimensions and thresholds.

Example: ISO 8000 data quality management standard covering master data, exchange, and reference data. DAMA-DMBOK (Data Management Body of Knowledge) 11 knowledge areas including data quality management. Six dimensions: accuracy (correctness), completeness (no missing values), consistency (no contradictions across systems), timeliness (current enough for use), validity (conforms to business rules), uniqueness (no duplicates). Statistical process control for data quality: control charts, monitoring thresholds, automated anomaly detection. Data profiling discovers quality issues before they propagate downstream.

0.2 3. Privacy Regulation

Privacy MUST comply with applicable regulations governing personal data collection, processing, and transfer.

Example: GDPR (EU) data controllers determine purposes/means of processing; data processors act on controller instructions. Both have obligations. DPIA (Data Protection Impact Assessment) required for high-risk processing (Art. 35). DPO (Data Protection Officer) required for public bodies and large-scale systematic monitoring. CCPA/CPRA (California) applies to businesses with $25M+ revenue, 100K+ consumers data, or 50%+ revenue from selling PI. HIPAA de-identification: Safe Harbor method (18 identifiers removed) or Expert Determination method (statistical verification). Re-identification risk is real studies show 87% of US population identifiable by ZIP + birthdate + gender.

0.3 4. Data Governance

Data governance MUST establish clear ownership, stewardship, and decision rights across the data lifecycle.

Example: DAMA framework defines data governance as the exercise of authority, control, and shared decision-making over data assets. Data stewardship domain experts responsible for data quality within their domain. Data ownership models: enterprise (centralized), federated (domain-owned), hybrid. Master Data Management (MDM) single source of truth for critical business entities (customer, product, provider). Data governance council cross-functional body setting policies, resolving conflicts, prioritizing initiatives. Metadata management business glossary, data dictionary, lineage documentation.

0.4 5. Interoperability

Data systems MUST exchange information using open standards and documented interfaces.

Example: Open data standards: CSV (RFC 4180), JSON (RFC 8259), Parquet (columnar), Avro (row-based). API-first design RESTful APIs with OpenAPI 3.0 specification, GraphQL for flexible queries. Schema registries (Confluent, AWS Glue) enforce data contracts between producers and consumers. FAIR principles: Findable (persistent identifiers, rich metadata), Accessible (retrievable by identifier using open protocol), Interoperable (formal language, qualified references), Reusable (clear license, provenance, community standards). Domain-specific standards: HL7 FHIR (healthcare), FIX/FpML (finance), ACORD (insurance), MISMO (mortgage).

0.5 6. Retention & Destruction

Data MUST be retained according to legal requirements and destroyed when retention obligations expire.

Example: Legal hold (litigation hold) suspends normal retention/destruction when litigation is reasonably anticipated. Failure to preserve = spoliation sanctions under FRCP 37(e). Retention schedules vary by jurisdiction and data type: tax records (7 years IRS), medical records (6-30 years by state), employment records (varies by statute: FLSA 3 years, Title VII 1 year, ADEA 3 years). Right to erasure: GDPR Art. 17, CCPA deletion rights. Secure destruction: NIST SP 800-88 Rev. 1 clear (logical overwrite), purge (degaussing, cryptographic erase), destroy (physical destruction). Certificate of destruction documents compliant disposal.

---

1. Constraints

```
MUST:     Cite specific regulation or standard for data
MUST:     Distinguish between data controller and data
MUST NOT: Present anonymization as absolute  cite re-id
```

DATA | CANON | REGULATORY