Memoir.ai — Data & Schema Reference Book

Canonical Data Structures, Validation Rules, and Storage Schemas

Document Version: 1.0

Status: Engineering Reference

Owner: Data Architecture Team

Last Updated: YYYY-MM-DD

---

1. Purpose

This reference book defines the canonical data structures, schema rules, validation logic, and persistence models used within Memoir.ai. It serves as the authoritative guide for engineers implementing ingestion, storage, search, AI processing, export, and synchronization systems.

---

2. Scope

This reference includes:

• Canonical event model

• Participant and identity schemas

• Narrative and snapshot schemas

• Provenance and citation models

• Import validation rules

• Deduplication and merge logic

• Normalization pipelines

• Export schemas and packaging

• Supabase metadata schemas

• Row-level security enforcement

• Schema migrations and seeds

User interface behavior is out of scope.

---

3. Canonical Data Model Overview

Memoir.ai standardizes all imported sources into unified structures to enable cross-platform timeline reconstruction and AI summarization.

Core entities include:

• Event

• Person

• Attachment

• Thread

• Narrative

• Citation

• Snapshot

• DataSource

These entities enable ingestion from heterogeneous archives while maintaining consistency.

---

4. Event Entity (Core Record)

The Event is the atomic record within the vault.

Fields include:

event_id — UUID primary key

source_id — Source reference identifier

timestamp — ISO8601 event time

platform — Origin platform identifier

content_raw — Raw message or text body

metadata — Platform-specific structured data

Events may represent messages, posts, media, or emails.

---

5. Person Entity

Participants are extracted and unified across imports.

Fields include:

person_id — UUID identifier

display_name — Preferred name

identities — JSON mapping of emails, phones, handles

Events relate to participants via many-to-many relationships.

---

6. Narrative & Snapshot Entities

Narratives represent AI-generated summaries.

Fields include:

snapshot_id — Snapshot identifier

version_number — Revision count

narrative_body — Markdown narrative text

generated_at — Generation timestamp

Each regeneration creates a new version entry.

---

7. Citation & Provenance Model

Citations map narrative fragments to source evidence.

Attributes include:

citation_id — UUID

target_narrative_id — Narrative reference

source_event_ids — Evidence references

anchor_text — Supported fragment

confidence_score — Evidence strength

Deleting source events invalidates citations.

---

8. Import Validation Rules

Records must pass structural validation before ingestion.

Validation includes:

• Required fields present

• Timestamp format correctness

• Source integrity checks

• Media path verification

• Encoding cleansing

Batches failing >10% validation pause ingestion.

---

9. Deduplication & Merge Logic

Exact duplicates require:

• Timestamp match ±1 second

• Participant match

• Identical content hash

Fuzzy merging applies when:

• Timestamps within ±30 seconds

• Text similarity above threshold

Media deduplicated via SHA-256 hashing.

---

10. Normalization Pipeline

Normalization converts raw platform exports into canonical schema.

Processes include:

• Participant resolution

• Timestamp standardization to UTC

• Event type mapping

• Attachment extraction

• Thread reconstruction

Normalization ensures cross-platform compatibility.

---

11. Export Schema Overview

Exports include events, participants, narratives, and media.

Export structure includes:

export_metadata — Export info

events — Canonical event list

narratives — AI snapshot outputs

citations — Evidence mapping

Exports use open formats to ensure portability.

---

12. Export ZIP Layout

Export bundles follow:

Memoir_Export/

  data/

  media/

  docs/

  checksums.txt

Media stored via content hash to prevent duplication.

Narratives mirrored as Markdown files for direct readability.

---

13. Supabase Metadata Schema

Cloud layer tracks metadata only.

Primary tables include:

auth_profiles — User identity data

workspaces — Vault ownership records

data_sources — Import sources

import_jobs — Background ingestion jobs

snapshot_records — Snapshot metadata

citations — Evidence metadata

audit_logs — Activity tracking

No personal message content stored remotely.

---

14. Row-Level Security Policies

All Supabase tables enforce RLS.

Rules include:

• Users access only owned workspaces.

• Import jobs restricted by ownership.

• Snapshot access restricted by workspace owner.

Administrative operations restricted to service roles.

---

15. Schema Migrations

Schema evolution handled through ordered migrations.

Migration rules:

• Backward compatibility maintained

• Migration scripts versioned

• Rollback procedures defined

• Checksums verify migration integrity

Production migrations require validation before rollout.

---

16. Development Seeds

Seed data enables development environments.

Seed actions include:

• Insert test users

• Create default workspaces

• Assign free-tier entitlements

Seeds must never include production data.

---

17. Data Integrity Principles

Integrity is preserved through:

• Hash-based deduplication

• Provenance tracking

• Version immutability

• Validation checkpoints

• Recovery support

All ingestion operations must remain idempotent.

---

18. Operational Best Practices

Engineering guidance:

• Validate schema changes early.

• Maintain canonical schema stability.

• Monitor ingestion error ratios.

• Audit deduplication effectiveness.

• Verify export compatibility regularly.

---

19. Conclusion

The Memoir.ai data and schema architecture ensures unified ingestion, accurate provenance tracking, reliable narrative generation, and long-term portability while maintaining user ownership and system integrity.