
Epiphany

TEXT PROJECTION

A companion to the Core Specification

Version 0.2.0 — Canonical-manifest coverage (blobs, extensions, custom profiles, the
derived-reference rule)

Normative for the text form it defines

Contents

1	About This Companion	2
1.1	What This Document Covers	2
1.2	The Subject of the Projection	2
2	The Canonical Text Form	4
2.1	Encoding and Character Set	4
2.2	Atoms	4
2.3	Layout	5
3	What Is Projected	6
3.1	Derive, or Carry — Never Both	6
3.2	Document Structure	7
3.3	Canonical Blobs	7
3.4	Profile Declarations	8
3.5	Extension Declarations	8
3.6	Reduced State	8
3.7	The Canonical Base Snapshot	9
4	Requirements	11
4.1	Canonicity	11
4.2	Round Trip	11
4.3	Strict Parsing	12
4.4	Conformance	12
5	Grammar	13
6	A Worked Example	15
7	Revision History	16

About This Companion

The *Text Projection* is a companion to the Epiphany Core Specification. It fulfils the delegation the core specification makes in Chapter 8, TEXT PROJECTION (`sec:format:textproj`), which declares that “the format admits a deterministic projection to a canonical s-expression text form” and that “the text projection is normative” — while leaving the form itself unwritten. The Binary Format companion likewise excludes it: “the canonical s-expression form — that is the *Text Projection* companion’s”.

This document supplies that form.

1.1 What This Document Covers

- The canonical text syntax: atoms, byte strings, text, and the layout that makes the projection deterministic.
- What is projected, and what is deliberately not.
- The projection and parse requirements, including the bidirectional round-trip the core specification demands.

It does *not* cover the binary encoding of anything — that is the Binary Format companion’s — nor the semantics of any operation, which is the Operation Catalog’s. The projection is a *re-presentation* of the canonical document, never a second definition of it. Wherever the two could disagree, the binary form is normative and the projection is wrong.

1.2 The Subject of the Projection

The projection’s subject is the **canonical document**, not the file. Per the core specification’s Chapter 8 TEXT PROJECTION, the projection **MUST NOT** be required to preserve chunk offsets, compression choices, cache chunks (operation indexes, layout caches, integrity indexes), garbage bytes from prior commits, or superblock generation numbers, slot assignments, and CRCs.

Consequently two bundles that differ only in physical layout project to the same text, and a text re-serializes to *a* bundle rather than to *the* bundle it came from. That is the intent, not a limitation: the physical file is an encoding of the document, and the projection is of the document.

A bundle **MAY** cache its own projection in a `TextProjection` chunk named by `Manifest.text_projection_root`. That chunk is a **non-canonical accelerator** (core specification Chapter 8, SCHEMA VERSIONING): a reader need not understand it, and a writer preserves it verbatim or discards it. A cached projection that disagrees with the operations it claims to project is *stale*, not authoritative.

The Canonical Text Form

2.1 Encoding and Character Set

REQUIREMENT

A text projection **MUST** be UTF-8, with no byte-order mark. Every text field it carries **MUST** be in Unicode NFC, matching the canonical-text rule of the core specification's Appendix D, TEXT AND UNICODE. A parser **MUST** reject non-NFC text rather than normalize it.

2.2 Atoms

An *atom* is a symbol, an integer, a byte string, or a text string.

Symbols are lowercase ASCII words, possibly hyphenated: `envelope`, `insert-event`, `strict-inverse`. They name constructors and enumeration cases. A symbol is never quoted.

Integers are written in base ten, with a leading `-` for negative values, no leading zeros, and no leading `+`. Zero is `0`, never `-0`.

Byte strings carry every identifier, hash, and opaque payload.

REQUIREMENT

A byte string **MUST** be written as `#x` followed by an even number of **lowercase** hexadecimal digits, one pair per byte, in the order the bytes appear in the canonical binary form. The empty byte string is `#x`. A parser **MUST** reject uppercase digits, an odd digit count, and any separator within the digits.

RATIONALE

One rule for every byte string. Hexadecimal has no alphabet variant and no padding to canonicalize, it is greppable, and a corrupted character is locally obvious. Identifiers are 16 or 32 bytes, so its expansion costs nothing where it is read; only an inlined snapshot pays. The core specification permits “base64 or another canonical text form”; a second

encoding would buy a quarter of the bytes of the one body nobody reads, at the price of pinning an alphabet, a padding rule, and a line-wrapping rule, and of choosing which encoding applies where. Ratified at 0.1.0.

Text strings are double-quoted. Inside a string, `"` denotes a quotation mark, `\` a backslash, `\n` a line feed, and `\t` a tab; no other escape exists.

REQUIREMENT

A text string **MUST** escape exactly the characters that require it: the quotation mark, the backslash, `U+000A`, and `U+0009`. Every other character **MUST** appear literally. A parser **MUST** reject an escape sequence outside this set, and **MUST** reject a literal character that the writer was required to escape.

RATIONALE

“Escape exactly” rather than “escape at least”: the text is canonical, so two spellings of one string cannot both be valid. This is the same injectivity the binary form rests on (Binary Format, `req:binfmt:decode-vectors`), stated for text.

2.3 Layout

REQUIREMENT

A projection is a sequence of lines separated by a single `U+000A`, with a final `U+000A` and no other trailing whitespace. Each line is one complete s-expression. Tokens within a line are separated by exactly one space; there is no other whitespace, and no indentation. Each operation envelope **MUST** occupy exactly one line.

RATIONALE

The core specification’s stated use case is that “merge conflicts surface at the operation-envelope level, which is the meaningful level for collaborative editing”. One envelope per line makes a line-based three-way merge conflict *exactly* an envelope conflict — never a conflict inside an envelope, which could otherwise produce a syntactically valid operation that neither side wrote. It also disposes of indentation: there is no whitespace to canonicalize, so the “identical semantics project to identical text” requirement below has nothing to hide in.

The lines are long. Readability is a *tooling* concern, and a pretty-printer is free to reformat for display; what it must not do is write the reformatted text back and call it a projection. Ratified at 0.1.0.

What Is Projected

3.1 Derive, or Carry — Never Both

The manifest’s references are *physical*. A `ChunkRef` carries an offset, a compressed length, and a compression algorithm; a `BlobRef` carries the same. Those are exactly the things the projection **MUST NOT** preserve. But a reference also carries an *identity* — a chunk id, a content hash, a blob id — and every one of those is a function of the content.

One rule resolves both, and it is the rule Requirement 3.6 already applies to reduced state.

REQUIREMENT

A projection **MUST** carry exactly what the document does not determine, and **MUST NOT** carry anything it does.

- **Physical attributes** — a chunk’s or blob’s `offset`, `compressed_length`, and `compression`, and a chunk’s `uncompressed_length` — **MUST NOT** appear. A serializer chooses them freely.
- **Derivable identities** — `ChunkId`, `ContentHash`, `BlobId` — **MUST NOT** appear. They are re-derived from the content by the derivations the Binary Format companion pins (`CONTENT HASHING`, `DOMAIN-SEPARATED PREIMAGES`).
- **Non-derivable identities MUST** appear. In schema major 0 there is exactly one: `SnapshotId`, which the Binary Format companion declares opaque, with readers forbidden from deriving it (`req:binfmt:snapshot-id-opaque`).
- **Content and semantic attributes MUST** appear: a chunk’s `kind` and `schema_version` and its uncompressed payload; a blob’s media type, declared maximum uncompressed length if any, and its payload.

A parser **MUST** reject a projection carrying a value this requirement forbids.

RATIONALE

Carrying a derivable identity would reproduce, at the level of a chunk, the defect Requirement 3.6 rules out at the level of the document: two sources of truth for one fact, with nothing to stop them disagreeing. Carrying a physical attribute would make two encodings of one document project to two texts, breaking Requirement 4.1.

`SnapshotId` is the sole exception, and it is an exception for a stated reason rather than an

oversight: vo has no snapshot producer, so the identity has nothing to derive *from*, and the Binary Format companion accordingly pins it as sixteen opaque bytes that a reader **MUST NOT** attempt to verify. A projection must carry what it cannot recompute.

3.2 Document Structure

A projection is, in order:

1. a `(text-projection <version>)` header line, naming the version of *this companion* the text conforms to;
2. a `(document #x<document-id>)` line, and a `(lineage #x<lineage-id>)` line if the manifest declares one;
3. zero or more `(profile ...)` lines, in canonical order;
4. zero or more `(extension ...)` lines, in canonical order;
5. at most one `(canonical-base ...)` line;
6. zero or more `(blob ...)` lines, in canonical order;
7. zero or more `(envelope ...)` lines, in canonical operation order (core specification Appendix D).

Every sequence is written in the normative order its binary counterpart uses. The projection introduces no ordering of its own.

3.3 Canonical Blobs

REQUIREMENT

A blob referenced by a canonical operation or by canonical reduced state is itself canonical (core specification Chapter 8, CANONICAL AND NON-CANONICAL ROOTS). Every such blob **MUST** be projected as a `(blob ...)` line carrying its media type, its declared maximum uncompressed length if it declares one, and its uncompressed payload.

Its `BlobId`, content hash, offset, lengths, and compression are re-derived (Requirement 3.1).

A blob referenced only by acceleration structures is non-canonical and **MUST NOT** be projected.

RATIONALE

An embedded image, font, or audio recording that a canonical operation references is part of the document. Omitting it would make the projection lossy for exactly the documents most in need of archival — and lossy *silently*, since the operations that reference the blob would still be there, pointing at a blob id the text no longer contains. This requirement was absent from version 0.1.0 of this companion, and from the core specification’s own list of what the projection preserves; both are corrected.

3.4 Profile Declarations

A `(profile ...)` line carries the profile’s identity, its semantic version, and its constraints.

REQUIREMENT

A profile identity is a symbol for each closed-vocabulary profile (`full`, `read-only`, `lite`), and `(custom #x<registry-id>)` for `ProfileId::Custom`, whose registry id is sixteen bytes.

RATIONALE

Version 0.1.0 required a symbol for every profile, which made a custom profile unrepresentable and its claim to preserve “all profile declarations” false.

3.5 Extension Declarations

REQUIREMENT

An `(extension ...)` line **MUST** carry every field of the declaration: its identity, its semantic version, whether it is required, its affected object kinds, its edit barriers, and its preserved chunk roots. The affected object kinds and the edit barriers are opaque to the bundle and are projected as byte strings, verbatim.

Each preserved chunk root is projected as its `kind`, its `schema_version`, and its uncompressed payload (Requirement 3.1), never as a `ChunkRef`: a `ChunkRef` is a physical reference, and the projection has no file to point into.

RATIONALE

Version 0.1.0 carried only identity, the required flag, barriers, and a list of chunks, dropping the semantic version and the affected object kinds outright and leaving the chunks’ representation undefined. An extension’s declaration governs how the canonical document is *interpreted*; a projection that loses part of it does not determine the document.

3.6 Reduced State

REQUIREMENT

The projection **MUST** preserve canonical reduced state *by preserving the operations that determine it*. It **MUST NOT** carry a second, literal copy of the reduced state.

RATIONALE

Reduced state is a deterministic function of the operation set and the canonical base (core specification Chapter 6, DESIGN PRINCIPLES). A text that carried both would have two sources of truth for one fact, and nothing could stop them disagreeing — a projection with an internally contradictory document is worse than no projection. This reading is what the core specification’s own round-trip clause already implies: text that “parses to identical canonical document semantics” must re-serialize to bundles with identical semantics, and reduced state is a function of semantics.

Read the core specification’s “all canonical reduced state” as *determines*, not *contains*. Ratified at 0.1.0.

3.7 The Canonical Base Snapshot**REQUIREMENT**

If the manifest declares a canonical base, the projection **MUST** carry a (canonical-base ...) line bearing:

- the `SnapshotId`, verbatim — the one identity that is carried rather than derived (Requirement 3.1);
- the causal frontier it materializes, as opaque bytes;
- its reduction-algorithm version;
- the profile under which it was produced;
- its root chunk’s `schema_version`; and
- *the root chunk’s uncompressed payload, inline*, as a single byte string. That payload is the canonical byte form of the reduced state the snapshot materializes.

The root chunk’s kind is `Snapshot` by role and is not written. The `SnapshotRef`’s `root` and `hash` are **re-derived**: the chunk’s content hash is `hash(Snapshot, schema, payload)` under the Binary Format companion’s chunk-hash preimage, and it is both the root’s `ChunkId` and the `SnapshotRef`’s `hash`. A parser **MUST** perform that derivation rather than read it from the text.

RATIONALE

A canonical base exists precisely so that the operations before its frontier need not be retained. Where they have been pruned, the snapshot is *not* derivable from anything else in the document, and a projection that carried only a reference would be **lossy** — the text would no longer determine the document, which is the one thing it is for. The core specification permits the payload to be “encoded compactly or referenced externally”; inline and compact is the choice that keeps the text self-contained.

The snapshot diffs as one opaque atom. That is acceptable: merges happen among operations, and a base snapshot changes only when the document is compacted, at which point the whole line changes anyway. A later revision **MAY** project the snapshot structurally; doing so does not break the round trip, because the document it denotes is un-

changed. Ratified at 0.1.0.

Schema major 0 has no snapshot producer — pruning and canonical-base creation are deferred (Binary Format, `SNAPSHOTID`) — so this requirement binds whoever writes the first one, and cannot be exercised before then.

Requirements

4.1 Canonicity

REQUIREMENT

Two bundles whose canonical document semantics are identical **MUST** project to **byte-identical** text. A projector **MUST NOT** have any freedom the document does not determine: no optional whitespace, no alternative spelling of an atom, no ordering choice.

4.2 Round Trip

REQUIREMENT

Parsing a projection and re-serializing it to binary **MUST** yield a bundle whose canonical document semantics are identical to the original's. The bundle's physical layout, chunking, and compression **MAY** differ.

Equivalently, and more usefully to an implementer: for every bundle B ,

$$\text{semantics}(\text{parse}(\text{project}(B))) = \text{semantics}(B),$$

and for every valid projection T ,

$$\text{project}(\text{serialize}(\text{parse}(T))) = T.$$

The second equation is the text's own injectivity: it is *stronger* than the first, and it is the one a conformance test can check with byte equality.

4.3 Strict Parsing

REQUIREMENT

A parser **MUST** reject any text that is not the canonical projection of the document it denotes. It **MUST NOT** normalize: not whitespace, not letter case in a byte string, not an escape sequence, not an out-of-order sequence, not a duplicate in a set-typed field. Accepting non-canonical text and normalizing it *is* accepting it, and does not satisfy this requirement.

RATIONALE

This is the same discipline the binary decoders carry, and it exists for the same reason: a lenient parser makes two texts denote one document, and the projection's contract is that a text *determines* its document. The Binary Format companion learned this concretely — a whole-value re-encode guard catches the fields a decoder normalizes and is blind to order-preserving sequences, and a guard on an outer value can *mask* a lenient inner codec rather than fix it (Binary Format, THE DECODE VECTOR CORPUS). A text parser inherits both hazards, and the cheapest total defence is the same one: re-project the parsed document and compare, *and* check per-site the orders that re-projection would restore.

4.4 Conformance

REQUIREMENT

An implementation claiming Text Projection conformance **MUST** implement both directions. A projector alone does not conform: the round trip (Requirement 4.2) is the requirement, and half of it is not checkable.

Grammar

```

projection ::= header document lineage? profile* extension*
              canonical-base? blob* envelope*

header       ::= "(text-projection " version ")" LF
version      ::= integer "." integer "." integer

document     ::= "(document " bytes ")" LF
lineage      ::= "(lineage " bytes ")" LF

profile      ::= "(profile " profile-id " " version " " constraints ")" LF
profile-id   ::= "full" | "read-only" | "lite" | "(custom " bytes ")"
constraints  ::= "(constraints " integer " " retention ")"
retention    ::= "(retention " integer " " (integer|"()") " " bool ")"

extension    ::= "(extension " bytes " " version " " bool
              " (kinds " bytes ") (barriers " bytes ")"
              " (chunks " chunk* "))" LF
chunk        ::= "(chunk " symbol " " schema " " bytes ")" ; kind, schema,
payload
schema       ::= "(schema " integer " " integer ")" ; major, minor

canonical-base
              ::= "(canonical-base " bytes " (frontier " bytes ")"
              " (reduction " integer ") (profile " profile-id ")"
              " " schema " (payload " bytes "))" LF

blob         ::= "(blob " string " " (integer|"()") " " bytes ")" LF
              ; media type, declared-max-uncompressed-length, payload

envelope     ::= "(envelope " bytes " (author " bytes ")"
              " (stamp " integer " " integer " " bytes ")"
              " (causal " vector " " dots ")"
              " " transaction " " payload ")" LF

transaction  ::= "()" | "(transaction " bytes ")"
payload      ::= "(primitive " kind ")"
              | "(resolve-conflict " bytes " " action ")"
              | "(undo " bytes " " policy ")"
              | "(resolve-equivocation " bytes " " bytes ")"

```

```

vector      ::= "(" ("(" bytes " " integer ")")* ")" ; replica, counter
dots        ::= "(" bytes* ")" ; operation ids

bytes       ::= "#x" hexdigit* ; even count, lowercase
integer     ::= "-"? digit+ ; no leading zeros, no "-0"
bool       ::= "true" | "false"
symbol      ::= [a-z] [a-z0-9-]*
string      ::= "'" schar* "'"

```

Observe what does *not* appear: no offset, no compressed length, no compression algorithm, no chunk id, no content hash, no blob id. Every one is either physical or derivable, and Requirement 3.1 forbids both. The lone opaque identity the grammar carries is the `SnapshotId`.

What this grammar is, and is not. The atom productions (`bytes`, `integer`, `symbol`, `string`) and the line shapes above are **normative**. The productions left unexpanded — `kind`, `action`, `policy`, `constraints`, `barrier` — are *derived* rather than invented here: there is exactly one per corresponding binary discriminant, named by the Operation Catalog’s section name in lowercase hyphenated form (`insert-event`, `transpose-interval`, ...), with its fields in the payload schema’s declaration order. A future revision **SHOULD** spell them out; until it does, the Operation Catalog and the Binary Format companion’s wire table jointly determine them, and an implementation disagreeing with those disagrees with this document.

That is a real gap, stated rather than papered over. It is the difference between a design gate and a finished companion.

A document of one operation — a transposition of two pitches up a perfect fifth, over a compacted base and one embedded image — projects to six lines:

```
(text-projection 0.2.0)
(document #x05050505050505050505050505050505)
(profile full (0 1 0) (constraints 67108864 (retention 1 () true)))
(canonical-base #x1f8b... (frontier #x00) (reduction 1) (profile full) (schema 0
  1) (payload #x0000...))
(blob "image/png" () #x89504e47...)
(envelope #x00000000000000007000000000000001 (author #
  x000000000000000000000000011223344) (stamp 42 7 #
  x00000000000000007000000000000001) (causal ((#x0000000000000001 3)) (#
  x00000000000000002000000000000009)) (transaction #
  x00000000000000007000000000000005) (primitive (transpose-interval (targets #
  x0000000000000000700000000000000100000000000007000000000000002) (interval 4
  7))))
```

The envelope's targets are a *set*: strictly increasing, no duplicates (Operation Catalog, req:opcat:transpose-in Binary Format seq[†]). A parser **MUST** reject a duplicate rather than absorb it, exactly as the binary decoder does.

Revision History

Date	Section	Change
July 9, 2026	All	0.1.0 — Initial companion. Supplies the canonical s-expression form the core specification’s Chapter 8 TEXT PROJECTION declares normative and leaves unwritten, and which the Binary Format companion excludes as “the <i>Text Projection</i> companion’s”. Ratified: reduced state is preserved by <i>determining</i> it, never by a second literal copy (req:textproj:reduced-state-derived); a canonical base snapshot is inlined as one opaque byte string, because a pruned document’s base is derivable from nothing and a reference-only projection would be lossy (req:textproj:base-snapshot-inline); lowercase hex is the single byte-string encoding (req:textproj:hex); one envelope per line, so a line-based merge conflict is exactly an envelope conflict (req:textproj:envelope-per-line). Parsing is strict-canonical (req:textproj:strict-parse) — normalizing non-canonical text is accepting it — and conformance requires <i>both</i> directions (req:textproj:conformance). No implementation yet; this document is the design gate.

Date	Section	Change
July 9, 2026	Chapters 3, 5	0.2.0 — Canonical-manifest coverage. 0.1.0 was <i>lossy for documents that are valid today</i>, and its claim to preserve the manifest’s canonical roots was false in three ways: a canonical blob (an embedded image, font, or recording referenced by a canonical operation) had no representation at all; an <code>ExtensionDeclaration</code> lost its semantic version and its affected object kinds, and left its preserved chunk roots undefined; and <code>ProfileId::Custom</code> was unrepresentable, a symbol being required where a sixteen-byte registry id is carried. The three share one cause, now stated as <code>req:textproj:derive-or-carry</code>: a <code>ChunkRef</code> and a <code>BlobRef</code> are <i>physical</i> references — offset, compressed length, compression — which the projection may not preserve, and they also carry <i>derivable</i> identities, which it may not duplicate. Carry the content and the semantic attributes; re-derive the rest. The sole non-derivable identity in schema major 0 is <code>SnapshotId</code>, which the Binary Format companion pins as opaque and forbids readers to derive. Consequently: <code>req:textproj:canonical-blobs</code> (a canonical blob is projected; a non-canonical one is not), <code>req:textproj:profile-id</code> ((custom #x...)), <code>req:textproj:extension-declaration</code> (every field, chunks as kind + schema + payload), and <code>req:textproj:base-snapshot-inline</code> extended to say what the inlined payload <i>is</i> and how the root chunk id and the snapshot hash are re-derived rather than read. The core specification’s own list of what the projection preserves omitted canonical blobs; it is corrected there too. The 0.1.0 ratifications — reduced state derived, base inlined, hex, one envelope per line, strict parsing — stand unchanged.