

Damage endgame database structure guide

Format 10 architecture, capture-threat exclusion, binary and ternary Tunstall coding, adaptive blocks, verification, and KingsRow comparison

Status	Current format-10 database set completed 25 August 2026
Reference set	S:\databases_new2
Generator	databasegen30 3.2; TDB format version 10
Coverage	International draughts WDL databases, 2 through 6 pieces

Executive summary

The current Damage endgame set stores all 155 material configurations from two through six pieces in five canonical TDB packages. Format 10 retains the compact four-byte block descriptor and the shared 192-dictionary catalog while adding explicit opponent-capture-threat exclusion metadata and a per-configuration threat count. Captures remain inherently excluded; threat exclusion is recorded so every compatible probe knows when search must continue instead of accepting a compression placeholder.

Measured result. The five TDB files total 518,044,186 bytes. Including the shared 1,530,862-byte tunstall.dic gives 519,575,048 bytes. Against KingsRow CPR+IDX combined, Damage is 490,978,967 bytes (48.59%) smaller. The validation count file adds only 13,502 bytes.

The compressor makes decisions at two levels. It first tests block sizes from 8,192 through 65,536 positions and stores the smallest complete material database. Inside each candidate block it tests uniform, raw two-bit, ternary Tunstall, and - when only two values remain - binary Tunstall encodings. Capture and enabled opponent-threat positions are replaced by the block's best tested W/D/L placeholder, while retained positions preserve their raw result.

1. Architecture at a glance

- Folder: five canonical dbN.tdb packages, one fingerprinted tunstall.dic, and database_counts.txt. Legacy dbN_<blocksize>.tdb names remain readable only when no canonical package is present.
- Package: 128-byte header, compact array of present 88-byte configuration records, then per-configuration descriptor arrays and payloads.
- Configuration: one material code with its selected power-of-two block size, position count, block count, offsets, capture count, opponent-threat count, and retained W/D/L statistics.
- Block: four-byte descriptor plus zero or more payload bytes; mode is uniform, raw, or Tunstall.
- Dictionary: IDs 0-127 are ternary; IDs 128-191 are binary; IDs 192-254 remain reserved and 255 means no dictionary where applicable.

Compatibility principle. Existing ternary IDs 0-127 were not renumbered. The decoder selects binary behavior from IDs 128-191 without adding a descriptor byte.

2. Published format-10 set

Package	Configs	Blocks	File bytes	Payload bytes	Raw/file
db2.tdb	4	4	1,125	629	1.97:1
db3.tdb	12	64	26,742	25,302	7.54:1
db4.tdb	25	2,567	928,073	915,477	11.31:1
db5.tdb	44	65,808	21,208,129	20,940,897	18.51:1
db6.tdb	70	2,252,805	495,880,117	486,862,609	23.29:1
TDB total	155	2,321,248	518,044,186	508,744,914	23.07:1

Package	Side-to-move captures	Opponent capture threats
db2.tdb	1,056	482
db3.tdb	174,324	77,194
db4.tdb	13,772,294	5,647,179
db5.tdb	697,870,878	253,847,060
db6.tdb	25,636,938,310	7,960,097,247
Total	26,348,756,862	8,219,669,162

Block payload accounts for 98.20% of the five TDB files. The remaining package bytes are headers, compact configuration records, four-byte descriptors, and alignment-free structural overhead. The shared dictionary is stored once for all packages. Threat exclusion reduces the full TDB-plus-dictionary set by 483,115,007 bytes (48.18%) versus the preceding format-9 set.

3. Folder and package organization

File	Purpose
dbN.tdb	Canonical package for total piece count N. Each configuration stores its independently selected 8K, 16K, 32K, or 64K block size.
tuninstall.dic	Shared external catalog of 192 fingerprinted dictionaries. Every TDB header references the same exact content.
database_counts.txt	Reference all-position and non-capture WDL counts used to validate the unchanged raw sources.

3.1 Physical order inside one package

1. 128-byte package header.
2. A compact directory containing only present material configurations; each format-10 record is 88 bytes.
3. For every configuration: a contiguous four-byte descriptor array followed immediately by that configuration's cumulative payload stream.

Derived offsets. The first payload offset is `block_directory_offset + block_count x 4`. Later payload offsets are obtained by cumulatively adding each preceding descriptor's 16-bit payload length.

4. Format-10 TDB header

Offset	Bytes	Field	Meaning
0	8	Magic	DGTDB01\0
8	4	Format version	10
12	4	Header size	128
16	4	Piece count	2...7
20	4	Selected minimum block size	8,192 in the current set
24	4	Bits per raw value	2
28	4	Dictionary count	192
32	4	Logical directory capacity	4,096 material keys
36	4	Stored configurations	Number of compact 88-byte records
40	8	Dictionary fingerprint	Must equal the external dictionary fingerprint
48	8	Directory offset	Normally 128
56	8	Package file size	Must equal actual length
64	8	Source raw bytes	Sum of unchanged two-bit sources
72	8	Payload bytes	Sum of encoded block payloads
80	8	Creation time	Unix timestamp
88	4	Flags	Bit 0: dominant runs; bit 1: opponent capture threats excluded
92	36	Reserved	Zero-filled

5. Material configuration records

The 12-bit key remains $\text{key} = \text{wm} \mid (\text{wk} \ll 3) \mid (\text{bm} \ll 6) \mid (\text{bk} \ll 9)$. Format 10 stores only present records, then reconstructs the 4,096-key in-memory lookup when opening the package.

Offset	Bytes	Field	Meaning
0	4	present	Bit 0 set
4	4	key	Logical 12-bit material key
8	4	positions_per_block	Selected 8K, 16K, 32K, or 64K size
12	4	block_count	$\text{ceil}(\text{position_count} / \text{positions_per_block})$
16	8	position_count	Indexed positions
24	8	block_directory_offset	Start of four-byte descriptors
32	8	raw_bytes	$\text{ceil}(\text{position_count} / 4)$
40	8	compressed_bytes	Payload sum for this configuration
48	8	capture_count	Side-to-move capture placeholders
56	8	retained_win	Stored wins after both exclusions
64	8	retained_draw	Stored draws after both exclusions
72	8	retained_lose	Stored losses after both exclusions
80	8	threat_count	Opponent-capture-threat placeholders

6. Four-byte block descriptor

Offset	Bytes	Field	Meaning
0	2	payload_bytes	Little-endian length, maximum 65,535
2	1	mode + permutation	Bits 0-1: mode; bits 2-7: WDL-to-ABC permutation 0...5
3	1	dictionary/value	Uniform WDL value, Tunstall dictionary ID, or 255 for no dictionary

Mode	Payload	Interpretation
0 - uniform	0 bytes	Dictionary/value byte is the single WDL result.
1 - raw	ceil(positions/4)	Four two-bit values per byte; selected when compressed alternatives are not smaller.
2 - Tunstall	Variable	One-byte phrase codes; dictionary ID selects ternary or binary alphabet.

7. Shared ternary and binary dictionaries

Every dictionary contains 255 Tunstall phrases over abstract symbols. The block permutation maps WDL results to A, B, and C, so a probability profile can be reused regardless of which result is dominant. Binary dictionaries contain only A and B phrases; C has probability zero and is forbidden in a block assigned to that dictionary.

ID range	Class	Alphabet	Purpose
0-127	Ternary	A, B, C	Existing 128 profiles; unchanged IDs
128-191	Binary	A, B	64 dominant probabilities from 99.95% to 50%, spaced evenly in log-odds
192-254	Reserved	-	Available for future dictionary families
255	Special	-	Descriptor no-dictionary value; payload code 255 introduces a dominant run

Selection boundary. A binary dictionary is eligible only when the mapped block contains exactly two distinct values. Eligibility does not force its use: the compressor still compares binary, ternary, and raw payload lengths and stores the smallest.

The external DGTUN01 file retains its version-1 container: magic, section offset, fingerprint, dictionary count, then each dictionary's three ppm probabilities and 255 length-prefixed phrases. The zero C probability and dictionary ID range distinguish binary entries without changing the file descriptor shape.

8. Compression decision process

- Validate each raw two-bit database against database_counts.txt; classify captures, optional opponent capture threats, and retained positions.
- Captures are always placeholders. When threat exclusion is enabled, positions where the opponent has an immediate capture are placeholders as well. For every affected block, test loss, win, and draw replacements; retained W/D/L results are never changed.
- For each material configuration, test candidate block sizes 8K, 16K, 32K, and 64K.
- Within each block, start with raw; detect uniform blocks; rank and fully encode the best ternary candidates.

- When only two mapped symbols occur, rank and encode binary candidates as well, testing the necessary WDL mappings.
- Retain the smallest payload per block, then retain the smallest complete material database including four-byte descriptors.
- Round-trip decode selected blocks, verify package structure and retained counts, publish a temporary file atomically, and fingerprint the shared dictionary.

Dominant runs. With the current flag enabled, payload bytes 255, X encode X x 128 repetitions of the selected dictionary's dominant symbol. Phrase codes themselves occupy 0-254.

9. Measured block profile

P	Blocks	Uniform	Raw	Ternary	Binary	Configurations by block size
2	4	0	0	2	2	8K:4
3	64	13	0	25	26	8K:6/16K:3/32K:1/64K:2
4	2,567	766	0	755	1,046	8K:12/16K:2/32K:2/64K:9
5	65,808	16,324	0	24,791	24,693	8K:14/16K:3/32K:3/64K:24
6	2,252,805	496,991	0	1,142,578	613,236	8K:19/16K:7/32K:5/64K:39

Across the format-10 set, 639,003 of 1,807,154 Tunstall blocks (35.36%) use the binary catalog. A further 514,094 blocks are uniform; raw blocks remain available as a strict fallback but none were selected in this set.

10. Probe and verification path

- Open the matching canonical piece-count package and validate magic, format 10, recorded file size, directory count, flags, and dictionary fingerprint.
- Reconstruct the 4,096-key configuration lookup from compact stored records; obtain that configuration's own positions_per_block.
- Compute block = index / positions_per_block and target = index % positions_per_block; derive the payload offset cumulatively from four-byte descriptors.
- Decode uniform or raw directly. For Tunstall, select dictionary ID, reverse the stored WDL permutation, expand phrase codes or dominant runs, and return the target symbol.
- Resolve captures through mandatory capture continuation. If header flag bit 1 is set and the opponent has an immediate capture, continue search rather than treating the stored compression placeholder as a final result.
- Egdbexplorer 1.89 reads format 10, exposes threat-exclusion metadata and counts, and skips excluded threats during probing, navigation, tracing, benchmarking, and verification.

11. KingsRow architecture comparison

Aspect	Damage format 10	KingsRow
Material coverage	All logical WDL configurations	Dominant material databases; both move orientations represented inside KingsRow organization
Data files	Five canonical piece-count TDB packages + one shared dictionary	.cpr1 compressed data plus .idx1 indexes; 6-piece material split across many file pairs
Index granularity	Four-byte descriptor per adaptive block; cumulative payload offsets	.idx1 supplies block boundaries/offset information for .cpr1

Aspect	Damage format 10	KingsRow
Coding	Uniform, raw two-bit, ternary/binary Tunstall, dominant runs, capture/threat placeholders	Tunstall and run-length-oriented compressed blocks
Move orientation	Damage configuration is side-to-move oriented; colour reversal supplies counterpart	Dominant material stored from black perspective with white/black-to-move sections
Dictionary model	One external 192-dictionary catalog shared across 2-6 pieces	Dictionary/block organization embedded in KingsRow's established CPR/IDX format

12. Complete storage comparison: 2 through 6 pieces

Pieces	Damage TDB	KR CPR	KR IDX	KR combined	Damage - KR
2	1,125	404	111	515	+610
3	26,742	21,628	413	22,041	+4,701
4	928,073	1,176,396	4,729	1,181,125	-253,052
5	21,208,129	33,886,572	126,504	34,013,076	-12,804,947
6	495,880,117	971,393,762	3,943,496	975,337,258	-479,457,141
Total	518,044,186	1,006,478,762	4,075,253	1,010,554,015	-492,509,829

Fair full-file result. The five Damage TDB files are 492,509,829 bytes (48.74%) smaller than KingsRow CPR+IDX. After charging the shared 1,530,862-byte dictionary once, Damage remains 490,978,967 bytes (48.59%) smaller. `database_counts.txt` is validation support rather than probe storage.

The format-10 TDB files are 492,509,829 bytes smaller than KingsRow CPR+IDX. After charging the shared 1,530,862-byte dictionary once, Damage remains 490,978,967 bytes (48.59%) smaller. The 6-piece result supplies nearly all of the advantage at scale.

13. Design trade-offs

Choice	Benefit	Cost / constraint
Binary catalog	Improves exactly-two-value blocks without displacing existing ternary IDs	Adds dictionary bytes and compressor candidate work
Adaptive block size	Balances descriptor overhead against local probability fit per material database	Compression tests four complete alternatives
Four-byte descriptor	Very low index overhead and no separate payload offsets	Random access derives offsets cumulatively; runtime anchors mitigate phrase scanning
Shared dictionary	Catalog cost paid once and fingerprinted across packages	All linked packages must use the matching <code>tunstall.dic</code>
Capture replacement	Large compression gain while preserving quiet WDL semantics	Capture answers always follow mandatory continuation logic
Threat replacement	Removes 8,219,669,162 additional tactical positions from the retained stream	Header flag bit 1 requires every probe consumer to detect opponent captures and continue search

14. Integrity and compatibility

- Reject unsupported format versions, mismatched sizes, invalid block sizes, duplicate configuration keys, out-of-range payloads, invalid modes, and invalid WDL permutations.
- Require the external dictionary magic, section offset, count, and fingerprint to match every package; the current format-10 set uses 192 dictionaries.
- Require binary phrase symbols to remain A/B and assign binary IDs only when the mapped block contains no C value.
- Verify all-position raw counts, retained W/D/L counts, capture and opponent-threat counts, block round trips, final partial blocks, dominant runs, raw fallbacks, and excluded-position handling before publication.
- Readers remain compatible with formats 2-9. Format-10 files require an updated reader because configuration records grow from 80 to 88 bytes and header flag bit 1 changes the probe continuation requirement.

Format	Principal addition	Reader consequence
2-4	Legacy descriptors	Embedded or external dictionaries; historical readers
5	Compact four-byte descriptor	Cumulative payload offsets
6	Compact configuration directory	Only present material records stored
7	Adaptive block size	positions_per_block stored per configuration
8	WDL permutation	Permutation packed into descriptor mode byte
9	Binary dictionaries	IDs 128-191; supported by current readers
10	Threat-exclusion metadata and 88-byte records	Header flag bit 1; read threat_count and continue search for excluded threats

15. Measurements and provenance

- **Damage implementation:** C:\Users\bertt\source\repos\databasegen30\core\endgame_compressor.cpp and compressed_probe.cpp
- **Explorer implementation:** C:\Users\bertt\source\repos\egdbexplorer\src\tdb_database.cpp
- **Generated set:** S:\databases_new2; **format-10 capture-and-threat compression completed 25 August 2026**
- **KingsRow set:** S:\wld_database; .cpr1 and .idx1 files, 2 through 6 pieces
- **Measurement definition:** Damage full TDB file bytes versus KingsRow CPR+IDX bytes; shared Damage dictionary charged once only in the aggregate result