

The Legacy Binary Office Readers

STATE

committed

CREATED

2026-08-07

LAST UPDATED

2026-08-08

INTENDED STATUS

Committed

AUTHORS

Zen Contributors <team@insan.ai>

DISCUSSION

Mapping, omissions, and round-trip expectations for the DOC, XLS, PPT, and XLSB readers

LABELS

formats, doc, xls, ppt, xlsb, reader

Status of This Memo

This document is an internal Zen discussion record authored in Typst and tracked in git. It intentionally follows RFC-style structure so the design scope, rationale, trade-offs, and operational constraints remain explicit. Documents that still use the placeholder number XXXXX are provisional drafts. Numbered ZDS documents are part of the project record.

Contents

Abstract 1

Introduction 1

 Why a converter needs written records 1

Terminology and Scope 1

Discussion Lifecycle 1

 Local Workflow Diagram 1

States 1

 State Transition Diagram 1

 Transition Detail 1

Authoring Rules 1

 Format Records 1

Typst Project Layout 1

Build Integration 1

Number Assignment and CI 1

 Numbering State Diagram 1

Security Considerations 1

References	1
1. Abstract	20
2. Introduction	21
2.1. Why a converter needs written records	21
3. Terminology and Scope	21
4. Discussion Lifecycle	21
4.1. Local Workflow Diagram	22
5. States	22
5.1. State Transition Diagram	22
5.2. Transition Detail	23
6. Authoring Rules	23
6.1. Format Records	24
7. Typst Project Layout	24
8. Build Integration	24
9. Number Assignment and CI	25
9.1. Numbering State Diagram	26
10. Security Considerations	26
11. References	26
Abstract	1
Introduction	1
Relationship to other records	1
Terminology and Scope	1
Problem Statement	1
Goals and Non-Goals	1
Goals	1
Non-Goals	1
Design Overview	1
The Document AST	1
Shape: a Zig-native document model	1
Storage: flat, preorder, struct-of-arrays	1
Attributes	1
The text pool and side tables	1
Invariants	1
A worked example	1
Why the tree is a tree	1
The Builder	1
Traversal and Rewriting Algorithms	1
Bounded, non-recursive walking	1
Skipping and scanning	1
The rebuild transform, with subtree sharing	1
Complexity summary	1
Staged parsing and streaming rendering	1
Public API and Developer Experience	1
Plugin authoring surface	1
Filters	1
The build.zig analogy, taken literally	1
The filter contract	1
Filters that ship	1
The Plugin Contract	1

Readers	1
Writers	1
The Registry and the Static Router	1
One table	1
Dispatching a runtime pair into a comptime matrix	1
Format taxonomy and roadmap	1
Format detection	1
Adjacent Artifact Manifest	1
Version 1 envelope	1
Loading and carrying plugin data	1
Commit and stream behavior	1
Memory Model	1
Diagnostics and Error Messages	1
The report	1
What it looks like	1
The lossiness tiers	1
Reading the Office Formats	1
The container: OOXML and ODF are ZIP archives	1
The XML layer	1
DOCX	1
Numbering, the hard case	1
Deliberate omissions	1
The other formats	1
The Markdown Writer	1
Repository and Build Layout	1
Command-Line Interface	1
Coding Standard	1
Testing Strategy	1
Security Considerations	1
Operational Considerations	1
Alternatives Considered	1
Decisions Clarified by Review	1
Open Questions	1
The cost of Zig-only filters	1
Image bytes	1
Table alignment source	1
Streaming for very large inputs	1
Filter ordering and conflict	1
Encoding detection	1
Delivery Plan	1
References	1
12. Abstract	20
13. Introduction	21
13.1. Relationship to other records	22
14. Terminology and Scope	22
15. Problem Statement	22
16. Goals and Non-Goals	23
16.1. Goals	23
16.2. Non-Goals	24

17. Design Overview	24
18. The Document AST	25
18.1. Shape: a Zig-native document model	25
18.2. Storage: flat, preorder, struct-of-arrays	27
18.3. Attributes	29
18.4. The text pool and side tables	30
18.5. Invariants	31
18.6. A worked example	32
18.7. Why the tree is a tree	33
18.8. The Builder	33
19. Traversal and Rewriting Algorithms	34
19.1. Bounded, non-recursive walking	34
19.2. Skipping and scanning	34
19.3. The rebuild transform, with subtree sharing	34
19.4. Complexity summary	36
19.5. Staged parsing and streaming rendering	36
20. Public API and Developer Experience	37
20.1. Plugin authoring surface	38
21. Filters	39
21.1. The build.zig analogy, taken literally	39
21.2. The filter contract	40
21.3. Filters that ship	41
22. The Plugin Contract	41
22.1. Readers	41
22.2. Writers	42
23. The Registry and the Static Router	43
23.1. One table	43
23.2. Dispatching a runtime pair into a comptime matrix	44
23.3. Format taxonomy and roadmap	45
23.4. Format detection	45
24. Adjacent Artifact Manifest	45
24.1. Version 1 envelope	45
24.2. Loading and carrying plugin data	46
24.3. Commit and stream behavior	47
25. Memory Model	47
26. Diagnostics and Error Messages	48
26.1. The report	48
26.2. What it looks like	50
26.3. The lossiness tiers	52
27. Reading the Office Formats	53
27.1. The container: OOXML and ODF are ZIP archives	53
27.2. The XML layer	53
27.3. DOCX	53
27.3.1. Numbering, the hard case	59
27.3.2. Deliberate omissions	59
27.4. The other formats	59
28. The Markdown Writer	61
29. Repository and Build Layout	62
30. Command-Line Interface	63

31. Coding Standard	64
32. Testing Strategy	65
33. Security Considerations	66
34. Operational Considerations	67
35. Alternatives Considered	67
36. Decisions Clarified by Review	69
37. Open Questions	69
37.1. The cost of Zig-only filters	69
37.2. Image bytes	69
37.3. Table alignment source	69
37.4. Streaming for very large inputs	69
37.5. Filter ordering and conflict	70
37.6. Encoding detection	70
38. Delivery Plan	70
39. References	71
Abstract	1
Scope	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
40. Abstract	20
41. Scope	20
42. Mapping	21
43. Deliberate omissions	23
44. Round-trip expectations	23
45. Security	23
46. References	24
Abstract	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
47. Abstract	20
48. Mapping	21
49. Deliberate omissions	22
50. Round-trip expectations	22
51. Security	22
52. References	22
Abstract	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
53. Abstract	20
54. Mapping	21

55. Deliberate omissions	22
56. Round-trip expectations	22
57. Security	22
58. References	22
Abstract	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
59. Abstract	20
60. Mapping	21
61. Deliberate omissions	22
62. Round-trip expectations	23
63. Security	23
64. References	23
Abstract	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
65. Abstract	20
66. Mapping	21
67. Deliberate omissions	22
68. Round-trip expectations	23
69. Security	23
70. References	23
Abstract	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
71. Abstract	20
72. Mapping	21
73. Deliberate omissions	23
74. Round-trip expectations	23
75. Security	23
76. References	23
Abstract	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
77. Abstract	20
78. Mapping	21
79. Deliberate omissions	22

80. Round-trip expectations	23
81. Security	23
82. References	23
Abstract	1
Scope	1
Mapping	1
Deliberate omissions	1
Round-trip expectations	1
Security Considerations	1
References	1
83. Abstract	20
84. Scope	21
85. Mapping	21
86. Deliberate omissions	21
87. Round-trip expectations	22
88. Security Considerations	22
89. References	22
Abstract	1
Scope	1
Mapping	1
Layout facets	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
90. Abstract	20
91. Scope	21
92. Mapping	21
92.1. Layout facets	31
93. Deliberate omissions	31
94. Round-trip expectations	36
95. Security	36
96. References	37
Abstract	1
Scope	1
Mapping	1
DOC	1
XLS and XLSB	1
PPT	1
Deliberate omissions	1
Round-trip expectations	1
Security	1
References	1
97. Abstract	20
98. Scope	21
99. Mapping	21
99.1. DOC	21
99.2. XLS and XLSB	22
99.3. PPT	23

100. Deliberate omissions	23
101. Round-trip expectations	26
102. Security	26
103. References	26
Abstract	1
Introduction	1
Relationship to other records	1
The triggering proposal	1
Terminology and Scope	1
A Formal Model of Lossy Conversion	1
Documents and observations	1
The five outcomes of conversion	1
Why lossy alternatives cannot be equalities	1
Problem Statement	1
Goals and Non-Goals	1
Goals	1
Non-Goals	1
Design Overview	1
The Semantic Kernel	1
What stays, and why it stays	1
What changes: one schema, actually one	1
Entities: identity only where identity is needed	1
Sparse Facets	1
The catalog	1
One coordinate system	1
The facet axiom	1
Cost model	1
Extension Nodes	1
Writer Lowering	1
Capabilities, declared at compile time	1
The choice DAG	1
The cost order	1
The planner and its guarantees	1
Strict mode and refusal	1
Normalization, Not Saturation	1
Worked Mappings	1
Core Contract Repairs	1
Limits	1
Security Considerations	1
Operational Considerations	1
Implementation plan: the clean break	1
Acceptance gates	1
Alternatives Considered	1
Open Questions	1
References	1
104. Abstract	20
105. Introduction	21
105.1. Relationship to other records	21
105.2. The triggering proposal	22

106. Terminology and Scope	22
107. A Formal Model of Lossy Conversion	22
107.1. Documents and observations	23
107.2. The five outcomes of conversion	23
107.3. Why lossy alternatives cannot be equalities	24
108. Problem Statement	25
109. Goals and Non-Goals	26
109.1. Goals	26
109.2. Non-Goals	27
110. Design Overview	27
111. The Semantic Kernel	27
111.1. What stays, and why it stays	27
111.2. What changes: one schema, actually one	28
111.3. Entities: identity only where identity is needed	29
112. Sparse Facets	29
112.1. The catalog	29
112.2. One coordinate system	30
112.3. The facet axiom	31
112.4. Cost model	31
113. Extension Nodes	31
114. Writer Lowering	32
114.1. Capabilities, declared at compile time	32
114.2. The choice DAG	32
114.3. The cost order	33
114.4. The planner and its guarantees	33
114.5. Strict mode and refusal	34
115. Normalization, Not Saturation	35
116. Worked Mappings	35
117. Core Contract Repairs	36
118. Limits	37
119. Security Considerations	38
120. Operational Considerations	39
120.1. Implementation plan: the clean break	39
120.2. Acceptance gates	40
121. Alternatives Considered	40
122. Open Questions	41
123. References	42
Abstract	1
Introduction	1
Terminology and Scope	1
Normative implementation contract	1
Problem Statement	1
Goals and Non-Goals	1
Goals	1
Non-Goals	1
Implementation Architecture	1
End-to-end invariants	1
Public Python API	1
Package identity and imports	1

The conversion call	1
Source semantics	1
Destination semantics	1
Reusable policy with Converter	1
Formats and capability discovery	1
Limits and strictness	1
Successful result	1
Manifest model	1
Reports and exceptions	1
Elm-style Python error contract	1
Documentation and compatibility	1
Native Boundary	1
Why a versioned C ABI loaded by ctypes	1
ABI shape	1
Memory artifact ensemble	1
Loading and version checks	1
Concurrency and process behavior	1
Project and Build Implementation	1
Repository layout	1
Division of tool authority	1
Root build steps	1
Repository integration obligations	1
Python configuration	1
Packaging and Release	1
One release version	1
Wheels	1
Source distribution	1
PyPI publication	1
Testing and Quality Gates	1
Python unit tests	1
Native and integration tests	1
Installed-artifact tests	1
Benchmarking the installed wheel	1
Benchmark profiles	1
Measurement rules	1
Correctness before timing	1
Result artifacts and reporting	1
Acceptance gates	1
Security Considerations	1
Threat model	1
Input authority	1
Resource exhaustion and copies	1
Native library loading and supply chain	1
Filesystem output	1
In-process execution	1
Operational Considerations	1
Delivery Plan	1
Phase 1: native contract	1
Phase 2: Python surface	1

Phase 3: packaging and benchmark	1
Phase 4: publication	1
Alternatives Considered	1
Shelling out to the CLI	1
A CPython extension using the limited API	1
cffi	1
A pure-Python reimplementaion	1
A remote service client	1
The uv build backend	1
The repository root as the uv project root	1
Wheel-only publication	1
Returning status instead of raising	1
Mutable global configuration	1
Open Questions	1
Acknowledgements	1
References	1
124. Abstract	20
125. Introduction	21
126. Terminology and Scope	22
126.1. Normative implementation contract	23
127. Problem Statement	23
128. Goals and Non-Goals	24
128.1. Goals	24
128.2. Non-Goals	25
129. Implementation Architecture	25
129.1. End-to-end invariants	26
130. Public Python API	26
130.1. Package identity and imports	26
130.2. The conversion call	27
130.3. Source semantics	28
130.4. Destination semantics	28
130.5. Reusable policy with Converter	29
130.6. Formats and capability discovery	29
130.7. Limits and strictness	30
130.8. Successful result	30
130.9. Manifest model	31
130.10. Reports and exceptions	31
130.11. Elm-style Python error contract	31
130.12. Documentation and compatibility	33
131. Native Boundary	33
131.1. Why a versioned C ABI loaded by ctypes	33
131.2. ABI shape	33
131.3. Memory artifact ensemble	34
131.4. Loading and version checks	34
131.5. Concurrency and process behavior	35
132. Project and Build Implementation	35
132.1. Repository layout	35
132.2. Division of tool authority	36
132.3. Root build steps	36

132.4. Repository integration obligations	37
132.5. Python configuration	37
133. Packaging and Release	38
133.1. One release version	38
133.2. Wheels	38
133.3. Source distribution	39
133.4. PyPI publication	39
134. Testing and Quality Gates	40
134.1. Python unit tests	40
134.2. Native and integration tests	40
134.3. Installed-artifact tests	41
134.4. Benchmarking the installed wheel	41
134.4.1. Benchmark profiles	41
134.4.2. Measurement rules	42
134.4.3. Correctness before timing	43
134.4.4. Result artifacts and reporting	43
134.5. Acceptance gates	44
135. Security Considerations	44
135.1. Threat model	44
135.2. Input authority	44
135.3. Resource exhaustion and copies	44
135.4. Native library loading and supply chain	45
135.5. Filesystem output	45
135.6. In-process execution	45
136. Operational Considerations	45
137. Delivery Plan	46
137.1. Phase 1: native contract	46
137.2. Phase 2: Python surface	46
137.3. Phase 3: packaging and benchmark	46
137.4. Phase 4: publication	47
138. Alternatives Considered	47
138.1. Shelling out to the CLI	47
138.2. A CPython extension using the limited API	47
138.3. cffi	47
138.4. A pure-Python reimplementation	47
138.5. A remote service client	47
138.6. The uv build backend	47
138.7. The repository root as the uv project root	47
138.8. Wheel-only publication	48
138.9. Returning status instead of raising	48
138.10. Mutable global configuration	48
139. Open Questions	48
140. Acknowledgements	48
141. References	48
Abstract	1
Introduction	1
Relationship to existing records	1
Normative language and completion	1
Terminology and Scope	1

Goals and Design Principles	1
Product goals	1
System 1: recognition and action	1
System 2: inspection and understanding	1
Help is part of the interaction	1
Original visual language	1
Design Overview	1
Deliverable contract	1
Zig WebAssembly Implementation	1
Target decision	1
32-bit portability	1
Engine separation required for WASM	1
Low-level ABI	1
Browser profile and memory limits	1
Determinism and artifact parity	1
Public Browser API and Worker Model	1
Distribution shape	1
Ergonomic API	1
State machine	1
Elm-style browser diagnostics	1
Project Site Information Architecture	1
Stable routes	1
Language selection and translated books	1
Homepage content order	1
Download experience	1
Download page wireframe	1
Desktop homepage wireframe	1
Documentation wireframe	1
Mobile wireframe	1
Converter Interaction Specification	1
Input window	1
Output window	1
Progress, status, and focus	1
Responsive behavior	1
Book and ZDS as HTML Documentation	1
One source, two reading modes	1
Sphinx-quality documentation ergonomics	1
Navigation and contextual linking	1
Search	1
PDF publication	1
Benchmark Dashboard	1
What the front page may claim	1
Native lens	1
Browser lens	1
Correctness before timing	1
Result schema and generated presentation	1
Security and Privacy	1
Threat model	1
Capability minimization	1

Content handling	1
Browser isolation and policy	1
Resource exhaustion and recovery	1
Accessibility and Human Interface Gates	1
Build and Repository Integration	1
Proposed repository shape	1
Division of tool authority	1
Root build steps	1
Static assembly	1
Testing and Quality Gates	1
WASM tests	1
Browser coverage	1
Site and documentation tests	1
Performance budgets	1
Release acceptance matrix	1
GitHub Pages and Release 0.2.0	1
Pages workflow	1
Unified release artifacts	1
Operational Considerations	1
Delivery Plan	1
Phase 0: toolchain determinism	1
Phase 1: pure WASM boundary	1
Phase 2: browser API and playground	1
Phase 3: project site and documentation	1
Phase 4: benchmark and release	1
Alternatives Considered	1
WASI in the browser	1
Emscripten or another compiler toolchain	1
Main-thread conversion	1
WebAssembly threads	1
A JavaScript framework and npm build	1
A single-page application	1
Reauthoring the book in Sphinx or Markdown	1
Rendering converted Markdown as HTML	1
Opening output in a browser popup	1
Running competitor benchmarks in every visitor's browser	1
One blended native/WASM headline	1
Third-party analytics and hosted fonts	1
Publishing only the WASM file	1
Known Gaps Outside This Release	1
Open Questions	1
Acknowledgements	1
References	1
142. Abstract	20
143. Introduction	21
143.1. Relationship to existing records	22
143.2. Normative language and completion	22
144. Terminology and Scope	22
145. Goals and Design Principles	23

145.1. Product goals	23
145.2. System 1: recognition and action	24
145.3. System 2: inspection and understanding	24
145.4. Help is part of the interaction	24
145.5. Original visual language	25
146. Design Overview	25
146.1. Deliverable contract	26
147. Zig WebAssembly Implementation	26
147.1. Target decision	26
147.2. 32-bit portability	27
147.3. Engine separation required for WASM	27
147.4. Low-level ABI	27
147.5. Browser profile and memory limits	29
147.6. Determinism and artifact parity	30
148. Public Browser API and Worker Model	31
148.1. Distribution shape	31
148.2. Ergonomic API	31
148.3. State machine	32
148.4. Elm-style browser diagnostics	33
149. Project Site Information Architecture	33
149.1. Stable routes	33
149.2. Language selection and translated books	34
149.3. Homepage content order	35
149.4. Download experience	35
149.5. Download page wireframe	37
149.6. Desktop homepage wireframe	38
149.7. Documentation wireframe	39
149.8. Mobile wireframe	40
150. Converter Interaction Specification	40
150.1. Input window	40
150.2. Output window	40
150.3. Progress, status, and focus	41
150.4. Responsive behavior	41
151. Book and ZDS as HTML Documentation	41
151.1. One source, two reading modes	41
151.2. Sphinx-quality documentation ergonomics	42
151.3. Navigation and contextual linking	43
151.4. Search	43
151.5. PDF publication	43
152. Benchmark Dashboard	43
152.1. What the front page may claim	43
152.2. Native lens	44
152.3. Browser lens	44
152.4. Correctness before timing	44
152.5. Result schema and generated presentation	45
153. Security and Privacy	45
153.1. Threat model	45
153.2. Capability minimization	46
153.3. Content handling	46

153.4. Browser isolation and policy	46
153.5. Resource exhaustion and recovery	47
154. Accessibility and Human Interface Gates	47
155. Build and Repository Integration	48
155.1. Proposed repository shape	48
155.2. Division of tool authority	48
155.3. Root build steps	49
155.4. Static assembly	50
156. Testing and Quality Gates	50
156.1. WASM tests	50
156.2. Browser coverage	51
156.3. Site and documentation tests	51
156.4. Performance budgets	51
156.5. Release acceptance matrix	52
157. GitHub Pages and Release 0.2.0	53
157.1. Pages workflow	53
157.2. Unified release artifacts	53
158. Operational Considerations	54
159. Delivery Plan	54
159.1. Phase 0: toolchain determinism	54
159.2. Phase 1: pure WASM boundary	54
159.3. Phase 2: browser API and playground	55
159.4. Phase 3: project site and documentation	55
159.5. Phase 4: benchmark and release	55
160. Alternatives Considered	55
160.1. WASI in the browser	55
160.2. Emscripten or another compiler toolchain	55
160.3. Main-thread conversion	55
160.4. WebAssembly threads	56
160.5. A JavaScript framework and npm build	56
160.6. A single-page application	56
160.7. Reauthoring the book in Sphinx or Markdown	56
160.8. Rendering converted Markdown as HTML	56
160.9. Opening output in a browser popup	56
160.10. Running competitor benchmarks in every visitor's browser	56
160.11. One blended native/WASM headline	56
160.12. Third-party analytics and hosted fonts	56
160.13. Publishing only the WASM file	56
161. Known Gaps Outside This Release	57
162. Open Questions	57
163. Acknowledgements	57
164. References	57
Abstract	1
Introduction	1
Relationship to existing records	1
Normative language and completion	1
Terminology and Scope	1
Deliverable contract	1
Problem Statement	1

Goals and Non-Goals	1
Goals	1
Non-Goals	1
Design Overview	1
Modes	1
The zencli Library	1
Shape	1
The subcommand rule	1
The zenserve Library	1
HTTP kernel	1
Router and middleware	1
Observability core	1
Authentication core	1
The REST and Streaming API	1
Route table	1
The conversion request	1
Tika compatibility boundary	1
Streaming semantics	1
Server report codes	1
Storage on zaxonlite	1
Dependency and lifecycle	1
Schema	1
Bootstrap	1
Observability Specification	1
Log events	1
Metric catalog	1
The Web Interface	1
The autodoc pattern	1
The ui module	1
The command protocol	1
daisyUI without a framework	1
JavaScript requirement	1
Pages	1
Converter wireframe	1
User management wireframes	1
Session security in the browser	1
The CLI Surface	1
Concurrency and Resource Model	1
connection slots	1
Project and Build Implementation	1
Repository layout	1
Module graph	1
Root build steps	1
Integration obligations	1
Self contained distribution	1
Security Considerations	1
Threat model	1
Input and process containment	1
Credential handling	1

Transport	1
Denial of service	1
Storage	1
Supply chain	1
Testing and Quality Gates	1
Unit	1
End-to-end (loopback)	1
Interface and release gates	1
Benchmark Specification	1
One corpus and one provenance contract	1
Native converter lens with Docling	1
Server lens against Apache Tika Server	1
Correctness before timing	1
Result artifacts and gates	1
Operational Considerations	1
Delivery Plan	1
Phase 1: kernels	1
Phase 2: interface	1
Phase 3: secure mode	1
Phase 4: polish and release	1
Alternatives Considered	1
A third-party Zig HTTP framework	1
An event loop instead of threads	1
A supervised conversion process pool	1
In-process TLS	1
JWTs instead of opaque tokens	1
SQLite directly (or flat files) instead of zaxonlite	1
A configuration file	1
A separate zenfmt-server binary	1
Server-rendered HTML with htmx	1
A SPA framework interface	1
Reusing the ZDS 0015 playground wasm as the interface	1
Compiling the interface wasm per request, as zig std does	1
A live administration event channel	1
Storing conversion history	1
An asynchronous job API for large documents	1
Resolved Questions	1
Acknowledgements	1
References	1
165. Abstract	20
166. Introduction	21
166.1. Relationship to existing records	22
166.2. Normative language and completion	22
167. Terminology and Scope	22
167.1. Deliverable contract	23
168. Problem Statement	24
169. Goals and Non-Goals	24
169.1. Goals	24
169.2. Non-Goals	25

170. Design Overview	25
170.1. Modes	26
171. The zencli Library	27
171.1. Shape	27
171.2. The subcommand rule	27
172. The zenserve Library	27
172.1. HTTP kernel	27
172.2. Router and middleware	28
172.3. Observability core	29
172.4. Authentication core	29
173. The REST and Streaming API	30
173.1. Route table	30
173.2. The conversion request	31
173.3. Tika compatibility boundary	32
173.4. Streaming semantics	32
173.5. Server report codes	32
174. Storage on zaxonlite	34
174.1. Dependency and lifecycle	34
174.2. Schema	34
174.3. Bootstrap	35
175. Observability Specification	36
175.1. Log events	36
175.2. Metric catalog	36
176. The Web Interface	36
176.1. The autodoc pattern	36
176.2. The ui module	37
176.3. The command protocol	38
176.4. daisyUI without a framework	38
176.5. JavaScript requirement	39
176.6. Pages	39
176.7. Converter wireframe	40
176.8. User management wireframes	40
176.9. Session security in the browser	41
177. The CLI Surface	41
178. Concurrency and Resource Model	42
179. connection slots	42
180. Project and Build Implementation	43
180.1. Repository layout	43
180.2. Module graph	43
180.3. Root build steps	44
180.4. Integration obligations	45
180.5. Self contained distribution	45
181. Security Considerations	46
181.1. Threat model	46
181.2. Input and process containment	46
181.3. Credential handling	46
181.4. Transport	46
181.5. Denial of service	47
181.6. Storage	47

181.7. Supply chain	47
182. Testing and Quality Gates	47
182.1. Unit	47
182.2. End-to-end (loopback)	47
182.3. Interface and release gates	48
183. Benchmark Specification	48
183.1. One corpus and one provenance contract	48
183.2. Native converter lens with Docling	48
183.3. Server lens against Apache Tika Server	49
183.4. Correctness before timing	50
183.5. Result artifacts and gates	50
184. Operational Considerations	50
185. Delivery Plan	51
185.1. Phase 1: kernels	51
185.2. Phase 2: interface	51
185.3. Phase 3: secure mode	51
185.4. Phase 4: polish and release	51
186. Alternatives Considered	51
186.1. A third-party Zig HTTP framework	51
186.2. An event loop instead of threads	51
186.3. A supervised conversion process pool	52
186.4. In-process TLS	52
186.5. JWTs instead of opaque tokens	52
186.6. SQLite directly (or flat files) instead of zaxonlite	52
186.7. A configuration file	52
186.8. A separate zenfmt-server binary	52
186.9. Server-rendered HTML with htmx	52
186.10. A SPA framework interface	53
186.11. Reusing the ZDS 0015 playground wasm as the interface	53
186.12. Compiling the interface wasm per request, as zig std does	53
186.13. A live administration event channel	53
186.14. Storing conversion history	53
186.15. An asynchronous job API for large documents	53
187. Resolved Questions	53
188. Acknowledgements	54
189. References	54

97. Abstract

One format record for the four legacy binary Office readers: zenfmt_doc (ai.insan.zenfmt.doc, .doc), zenfmt_xls (ai.insan.zenfmt.xls, .xls), zenfmt_ppt (ai.insan.zenfmt.ppt, .ppt, .pps, .pot), and zenfmt_xlsb (ai.insan.zenfmt.xlsb, .xlsb). The first three share the zenfmt_cfb compound-file (MS-CFB) container library; XLSB is the binary sibling of XLSX inside the same OPC ZIP package. These formats exist so that decades of archived documents convert without a detour through Microsoft Office; the record states exactly how much of each format is projected and what is refused.

98. Scope

zenfmt_cfb reads the Compound File Binary format the way zenfmt_ooxml reads ZIP: directory first, streams on demand, every FAT, mini-FAT, DIFAT, and directory chain walk bounded by the sector count so a crafted cycle is a refusal rather than a hang, and every stream read counted against the byte limits of ZDS 0002. The version-3 and version-4 sector sizes are both accepted; the red-black directory tree is deliberately scanned linearly — the tree shape is a hint, not a trust boundary.

The readers are text-fidelity subsets, not full decoders:

- **DOC** parses the FIB, the piece table (CLX/Pcdt) from the 0Table/1Table stream, and both piece encodings (compressed cp1252 and UTF-16LE), emitting paragraphs, hard breaks, and HYPER-LINK fields as links. The STSH stylesheet and the PlcfBtePapx/PAPX bin table are resolved to map Heading 1–9 styled paragraphs (by built-in identifier or name) to headings; a malformed stylesheet degrades to plain paragraphs with a note.
- **XLS** parses BIFF8: BOUNDSHEET, SST (with CONTINUE splits and their restated encoding flags), FORMAT/XF number formats, DATETIME, and the cell records LABELSST, LABEL, NUMBER, RK, MULRK, BOOLERR, and FORMULA with cached values (following STRING records for cached strings).
- **PPT** walks the PowerPoint Document record tree iteratively and harvests TextHeaderAtom + TextCharsAtom/TextBytesAtom runs in stream order.
- **XLSB** parses the MS-XLSB record framing (1–2 byte type, 1–4 byte length varint) for the workbook, shared strings, styles, and worksheet parts, resolving sheets through the workbook relationships part.

99. Mapping

99.1. DOC

Facets (ZDS 0013): headings carry a ProvenanceFacet naming the plugin id, the WordDocument stream, and the paragraph's byte position in it, with confidence exact. Any paragraph whose STSH style resolves to a named, non-default style carries a StyleFacet with that name; the default style stays unnamed on purpose.

Source	Tree result	Facets
Paragraph mark (0x00)	paragraph boundary.	StyleFacet with the resolved STSH style name, for named non-default styles.
Heading 1–9 styled paragraph (STSH + PAPX istd)	heading with the style's level, clamped to 6.	ProvenanceFacet: plugin, WordDocument member, byte position, confidence exact; plus the StyleFacet above.

Source	Tree result	Facets
Vertical tab (0x0B)	hard_break within the para- graph.	none.
Tab (0x09)	A space.	none.
HYPERLINK field (0x13/0x14/0x15)	link around the field result text; \l targets become fragment links. Other fields keep their result text and lose their markers.	none.
Non-breaking hyphen (0x1E)	A hyphen.	none.

99.2. XLS and XLSB

The same projection as the XLSX reader (ZDS 0005): each sheet becomes a level-2 heading with the sheet name followed by one table; the first present row is the table_head. XLS honors the 1904 date epoch (DATEMODE); RK and MULRK integer and div-100 encodings are decoded exactly.

Source	Tree result	Facets
LABELSST, LABEL, and formula STRING results	The cell's text, shared strings re- solved.	GridFacet with value type text, the record's own row and column, and the text cached.
NUMBER, RK, MULRK records, and cached formula numbers	ISO dates for date for- mats (builtin ids and cus- tom format strings), per- centages multiplied out, other numbers ver- batim.	GridFacet with value type number or date; the ren- dered value is cached. BIFF formula source is not decoded, so formula stays empty.
BOOLERR records and cached formula booleans and errors	TRUE/FALSE, or the #REF!-	GridFacet with value

Source	Tree result	Facets
	style error spelling.	type boolean or error_value.
Formulas	Never evaluated — the cached value is used; xls.formula without-cached-value/xlsb.formulas without-cached-value when there is none.	Facet coordinates are the records' formula, so sparse sheets keep exact positions even where the projected table compacts them.

99.3. PPT

Source	Tree result	Facets
TextHeaderAtom type title or centerTitle	Level-2 heading per \r-separated line.	ProvenanceFacet on the heading: plugin, PowerPoint Document member, confidence exact.
Body and other text atoms	paragraph per \r-separated line; 0x0B is a hard_break.	ProvenanceFacet as above.
Notes-typed text	container with class notes.	none.

100. Deliberate omissions

Construct	Report code	Reason
DOC non-heading paragraph and character styles	doc.styles-omitted	Heading styles are resolved through STSH and PAPX; other paragraph

Construct	Report code	Reason
		styles and the CHPX character layer are not. The note fires only when the stylesheet itself cannot be read.
DOC table grid (0x07 marks)	doc.tables-flattened	Cell boundaries survive as paragraphs; the grid needs PAPX fInTable resolution.
DOC pictures and OLE objects	doc.embedded-objects-dropped	Object streams are for-

Construct	Report code	Reason
		mat soup; the text notes their pres- ence.
DOC page breaks	doc.page-breaks-dropped	Mark- down has no pages.
PPT positioning, images, shapes	ppt.presentation-projection	Same un- con- di- tional warn- ing as PPTX: a deck loses the most.
XLS and XLSB formulas	xls.formula-without-cached-value	zenfmt never eval- u- ates; cached re- sults are used when pre- sent.
BIFF5 and older	xls.unsupported-biff	Pre-1997 BIFF is a

Construct	Report code	Reason
		different record set; the refusal names the fix.

101. Round-trip expectations

These are one-way readers. A round trip through Markdown keeps: full text in reading order, paragraph boundaries, hyperlinks (DOC), sheet names and cell values with their date and percent renderings (XLS, XLSB), and title, body, and notes structure (PPT). It never recovers: character styling (the legacy CHPX layer is not read), the table grid (DOC), slide layout, or any embedded object.

102. Security

- Encryption is a refusal, never an attempt: doc.encryption-refused (FIB fEncrypted), xls.encryption-refused (FILEPASS), ppt.encryption-refused (a cryptography session record).
- Every CFB chain walk (FAT, mini-FAT, DIFAT, directory) is bounded by the file's sector count; cyclic FATs and lying stream sizes are Malformed refusals, covered by tests.
- Stream sizes are checked against max_entry_uncompressed and the running total against max_total_uncompressed before copying.
- XLSB inherits the ZIP rules of ZDS 0002 unchanged: central directory only, hostile names reject, stored and deflate only.
- The XLSB record-length varint is capped at four bytes and clipped to the part size, so a lying length cannot run past the buffer.

103. References

- ZDS 0002, the architecture record (container rules, limits, reports).
- ZDS 0005, the XLSX reader record (the spreadsheet projection mirrored here).
- MS-CFB, MS-DOC, MS-XLS, MS-PPT, MS-XLSB (Microsoft Open Specifications).