DocTura Desktop

System Specification & Architecture

0. Product Name & Purpose

Product Name: DocTura Desktop

Purpose

DocTura Desktop is a **generic, offline-first, plugin-driven document intelligence application** for converting semi-structured and structured documents into high-quality **Excel, CSV, Word, and PDF outputs**, with deterministic validation, audit logging, and optional AI summarisation.

DocTura is designed to handle:

- layout-shifting PDFs,
- multi-page tables,
- header-delimited rosters,
- score-domain-constrained statistical reports,
- and round-trip document workflows.

It is **not domain-locked** (e.g., WAEC) and supports extensibility via plugins.

0.1 Design Anchors (Non-Negotiable)

The system design is permanently informed by the **TASS Scaled Essay truncation incident**, which established the following invariants:

1. **Page boundaries are unreliable**
2. **Visual continuity ≠ logical continuity**
3. **Domain rules must override layout**
4. **Extraction, routing, and layout must be decoupled**

These principles are enforced at the architecture level.

0.2 Default Behaviours (Locked)

- **Extraction Mode:** Hybrid
  (Page-preserved + logical tables)
- **Title Handling:** Metadata sheet
  (Document_Metadata)
- **Naming:** Smart naming enabled
- **Theme Support:**
  - Corporate Theme
  - Indigenous Theme
- **Validation:** Deterministic, mandatory
- **Offline Core:** Yes

1. Functional Requirements

1.1 Ingestion

Users may ingest one or multiple files per run:

Supported Inputs

- PDF (native and scanned)
- DOCX
- Images (PNG, JPG, TIFF)
- Audio (WAV, MP3, M4A)

Automatic Detection

- File type
- Text quality (native vs scanned)
- OCR requirement
- Candidate document template via plugins
- Structural signals (headers, section titles, score domains)

1.2 Extraction (Hybrid – Core Capability)

Hybrid extraction produces **two parallel representations**.

A. Page-Preserved Representation

- One worksheet per page: Page_01, Page_02, …

- Best-effort table/text grid
- Serves traceability, audits, and debugging

B. Logical Table Representation

Tables are reconstructed using **semantic boundaries**, not page breaks.

Supported segmentation strategies (plugin-selectable):

1. **Score-Domain Segmentation**
    - Used for WAEC-style statistical distributions
    - Example:
        - Scaled Objective: 0–19
        - Scaled Essay: 15–40
    - Prevents truncation and misrouting
2. **Header-Repetition Segmentation**
    - Used for rosters (e.g., International Staff List)
    - Repeated headers define table boundaries
    - Section titles provide grouping context

Page breaks are ignored during logical segmentation.

1.3 Metadata Extraction

A dedicated worksheet **Document_Metadata** is created, containing:

- Document title
- Organization / issuing body (if detected)
- Reporting period / session / year
- Subject or document code (if applicable)
- Plugin used (ID + version + confidence)
- Extraction mode
- Excel layout mode
- Word layout settings
- Output formats generated
- Theme selected
- Timestamp
- Input file hash (SHA-256)

Metadata is **never duplicated** into data sheets by default.

1.4 Output Formats

User-selectable per run:

- **XLSX** (always available)
- **CSV** (per sheet or combined)
- **DOCX**
- **PDF**

Outputs may be generated independently or together.

1.5 Excel Output Layout Options (Expanded)

Workbook Structure

User chooses:

1. **Each logical table → separate worksheet**
2. **All logical tables → single worksheet**

Single-Worksheet Arrangement

If option (2) is selected:

- **Vertical stacking (down rows)**
    - Header repeated per table
    - Blank row separator
- **Horizontal placement (across columns)**
    - Column offsets auto-calculated
    - Borders applied
    - Blank column separator

These options apply **only to logical tables**.
Page-preserved sheets remain separate in Hybrid mode.

1.6 Word Output Options

When DOCX output is selected:

- Page orientation:
  - Portrait
  - Landscape
- Optional:
  - Insert page break after each table
  - Include extracted images
- Section titles rendered as Word headings
- Tables rendered using Word table styles

1.7 Reverse Conversion: Word / Excel → PDF (NEW)

DocTura supports **round-trip workflows**.

Excel → PDF

- Worksheet-based export
- User-defined:
  - sheet selection
  - orientation
  - scaling (fit-to-width / fit-to-page)
  - gridlines on/off
- Each worksheet rendered to one or more PDF pages

Word → PDF

- Preserves:
  - page orientation
  - section breaks
  - tables
  - headings
- Produces print-ready, official PDFs

Two modes (architectural):

- **Structural rendering (default, deterministic)**
- **Visual snapshot rendering (optional, later)**

1.8 Validation & Quality Assurance

Validation is **mandatory and deterministic**.

Generic Rules

- Percent totals end at 100.00 (± tolerance)
- No duplicate score rows
- Score ranges obey plugin constraints
- Cumulative frequency monotonic
- Non-negative frequency/percent

Roster-Specific Rules

- Header must precede data
- Column count consistency
- No orphan rows
- Detect broken row wraps (heuristic)

Outputs

- Embedded validation sheet
- JSON validation report

1.9 AI Summarisation (Optional, Safe)

AI is strictly limited to:

- describing patterns

- summarising validation results
- highlighting anomalies

AI must not:

- recompute totals
- modify data
- infer missing values

Outputs:

- AI_Summary worksheet
- Optional Markdown report

1.10 Audit Logging (Enterprise-Ready)

Each run records:

- Input file hash
- Plugin ID, version, confidence
- Extraction & layout configuration
- Validation results
- Output hashes
- Timestamp
- Optional user/machine metadata

Stored as:

- Per-run JSON logs
- Rolling index

2. Non-Functional Requirements

- Offline core functionality
- Modular plugin architecture
- Reproducible outputs
- OCR used only when required
- Secure local file handling
- Robust to:
  - layout shifts
  - missing headers
  - page-split tables

3. System Architecture

3.1 App Layer

- GUI bootstrap
- Controllers:
  - conversion
  - batch processing
  - review
- Settings UI:
  - themes
  - default layouts
  - output preferences

3.2 Core Engine (Updated)

Ingest

- file detection
- quality detection
- document context

Extract

- text
- tables
- images
- OCR
- speech-to-text

Reconstruct

- title block extractor
- sectionizer
- **header-based table segmenter**
- table router (page + logical)

Validate

- generic rules
- plugin-specific rules
- roster-specific rules

Output

- Excel writer (multi-layout)
- Word writer (orientation aware)
- CSV writer
- **PDF export engine (Excel/Word → PDF)**
- Smart naming engine

UI Support

- theme manager (Corporate / Indigenous)

3.3 Plugin Layer

Built-in Plugins

1. **waec_marksdist_plugin**
    - Score-domain routing
    - Paper splits
    - Distribution validation
2. **international_staff_list_plugin**
    - Header-based segmentation
    - Section grouping
    - Roster validation
    - Excel + Word support

Plugins declare:

- detection rules
- segmentation strategy
- supported outputs
- validation rules

3.4 Enterprise Layer

- Audit logs
- Policy enforcement
- Plugin signing (future)

4. Core Data Models (Updated)

ExtractionOptions

- mode
- metadata_policy
- outputs
- excel_layout (placement, arrangement, borders)
- word_layout (orientation, page breaks)
- pdf_export_source (excel / word / structure)
- theme
- ocr_enabled
- ai_summary_enabled

RoutedTables

- page_tables
- logical_tables (table objects with schema, source pages)

ValidationReport

- per-table status
- issues
- metrics summary

Status Check

✓ Original TASS & CASS conversion fully retained and protected

✓ Scaled Essay cutoff permanently resolved via architecture

✓ International Staff List supported

✓ Excel layout fully user-controlled

✓ Word layout supported

✓ Reverse PDF generation supported

✓ Generic, extensible, enterprise-grade

Theme Definitions

These are **system-level constants**, not suggestions.

🎨 Theme Pack 1: Corporate Theme

Purpose

For enterprise, government, academic, and professional environments where neutrality, clarity, and authority matter.

Colour Palette

- **Primary:** Deep Navy Blue #0B1F3B
- **Secondary:** Slate Grey #4A5568
- **Accent:** Gold #C9A227
- **Background:** Off-White #F7F9FC
- **Surface / Cards:** White #FFFFFF
- **Text (Primary):** Charcoal #1A202C
- **Text (Secondary):** Muted Grey #6B7280
- **Success:** Deep Green #1F7A1F
- **Warning:** Amber #D97706
- **Error:** Dark Red #9B1C1C

UI Usage Rules

- Primary buttons → Navy Blue
- Accent actions (Convert, Export) → Gold
- Headers → Navy Blue text
- Tables → White background, subtle grey gridlines
- Validation errors → Dark Red, never flashing

Tone: **formal, calm, authoritative**

🌍 Theme Pack 2: Indigenous Theme

Purpose

To reflect African identity, heritage, and grounded authenticity — without sacrificing usability or professionalism.

Colour Palette

- **Primary:** Earth Brown #5A3E2B
- **Secondary:** Forest Green #1E5631
- **Accent:** Burnt Orange #C05621
- **Highlight Accent:** Ochre Yellow #D69E2E
- **Background:** Warm Sand #FAF3E0
- **Surface / Cards:** Light Clay #FFF8ED
- **Text (Primary):** Dark Umber #2D1B12
- **Text (Secondary):** Olive Grey #6B705C
- **Success:** Deep Green #2F855A
- **Warning:** Earth Amber #B7791F

- **Error:** Clay Red #9C4221

UI Usage Rules

- Primary buttons → Forest Green
- Accent actions → Burnt Orange
- Headers → Earth Brown
- Tables → Warm Sand background, soft borders
- Validation highlights → Ochre Yellow

Tone: **grounded, warm, culturally confident**

Theme Selection Rule (System-Level)

- Theme applies to:
  - Entire UI
  - Export previews
  - Word cover page styling (if enabled)
- Theme **never alters extracted data**
- Theme choice is logged in audit metadata