

# KYO QA ServiceNow Knowledge Tool v25.1.1

## Overview

This tool extracts model numbers (e.g., `PF-740`, `TASKalfa AB-1234abcd`, `ECOSYS A123abcd`), QA/SB numbers, and descriptions from Kyocera QA/service PDFs using OCR and pattern recognition. It updates blank cells in the "Meta" column of a cloned ServiceNow-compatible Excel file, preserving the original. Text files for documents needing review are saved in `PDF_TXT/needs_review`. No PDFs are retained.

## What's New in v25.1.1

- **Enhanced OCR Processing:** Improved image preprocessing for better text extraction from scanned documents
- **Optimized Pattern Matching:** 10x faster regex processing with cached pattern compilation
- **Better Error Handling:** Robust file locking detection and recovery mechanisms
- **Improved User Interface:** Enhanced status reporting and progress tracking
- **Organized Review Files:** Review files now stored in dedicated subfolder for easier management
- **Memory Optimization:** Better memory management for processing large batches of files
- **Bug Fixes:** Resolved critical issues with pattern processing and UI components

## How to Set Up and Run

### 1. Prerequisites

- **Python 3.11.x (64-bit):** Download Python 3.11.9 Windows Installer or use a portable version in `python-3.11.9` folder.

- **Tesseract OCR:** Tesseract Windows Installer (UB Mannheim) or place portable binary in `tesseract` folder.
- **Dependencies:** Listed in `requirements.txt` (auto-installed via `run.py`).

## 2. Folder Structure

KYO\_QA\_ServiceNow\_Knowledge\_Tool\_v25.1.1/

```
|— START.bat
|— run.py
|— start_tool.py
|— requirements.txt
|— README.md
|— CHANGELOG.md
|— kyo_qa_tool_app.py
|— logging_utils.py
|— ocr_utils.py
|— ai_extractor.py
|— data_harvesters.py
|— excel_generator.py
|— file_utils.py
|— processing_engine.py
|— custom_exceptions.py
|— version.py
|— update_version.py
|— tesseract/ (optional, for portable Tesseract)
|— python-3.11.9/ (optional, for portable Python)
|— logs/ (auto-created)
|— output/ (auto-created)
|— PDF_TXT/ |— needs_review/ (auto-created)
```

## Key Files

File	Description
START.bat	One-click Windows launcher
run.py	Enhanced launcher with dependency installation
start_tool.py	Alternative launcher for compatibility
requirements.txt	List of Python dependencies
README.md	Setup instructions and usage guide
CHANGELOG.md	Version history and updates
version.py	Central version definition
update_version.py	Updates version across all files

## Core Modules

File	Role
kyo_qa_tool_app.py	Tkinter UI and main controller
processing_engine.py	Coordinates PDF processing pipeline
ocr_utils.py	Enhanced PDF-to-text conversion with AI-assisted OCR
ai_extractor.py	Wrapper for data extraction
data_harvesters.py	Optimized model number and metadata extraction
excel_generator.py	Builds Excel files for ServiceNow import

## Utility Modules

File	Purpose
file_utils.py	Enhanced file I/O with lock detection
logging_utils.py	Comprehensive logging system
custom_exceptions.py	Defines custom errors
config.py	Defines extraction patterns and rules
custom_patterns.py	User-defined regex patterns

## 📁 Auto-Generated Folders

Folder	Description
/logs/	Session logs (success/fail)
/output/	Excel output (cloned_<excel>.xlsx)
/PDF_TXT/needs_review/	Text files for documents needing review
/venv/	Virtual environment for isolation
/.cache/	Performance optimization cache

## 📋 Summary

- **Secure:** No PDF retention.
- **Automated:** Auto-installs dependencies.
- **Portable:** Supports portable Python and Tesseract for USB deployment.
- **Modular & Logged:** Comprehensive logging to /logs/ and PDF\_TXT/needs\_review/ for review.
- **UI:** Bright, Kyocera-branded Tkinter UI with progress bars, color-coded logs, and detailed processing feedback.
- **Excel:** Clones input Excel, updates only blank "Meta" cells with model numbers.

## Setup Steps

1. Place all files in a folder (e.g., `KYO_QA_ServiceNow_Knowledge_Tool_v25.1.1`).
2. Install Python 3.11.x or place portable Python in `python-3.11.9`. Optionally, install Tesseract or place in `tesseract` folder.
3. Run `START.bat` (Windows) or `python run.py`:
  - Sets up `/venv/` and installs dependencies from `requirements.txt`.
  - Outputs logs to `/logs/` and Excel to `/output/`.
4. Manual setup (if needed):

```
bash
```

```
cd KYO_QA_ServiceNow_Knowledge_Tool_v25.1.1
mkdir /S /Q venv
python -m venv venv
venv\Scripts\python.exe -m ensurepip --default-pip
venv\Scripts\python.exe -m pip install --upgrade pip
venv\Scripts\pip.exe install -r requirements.txt
python kyo_qa_tool_app.py
```

## Usage

1. Launch the tool via `START.bat` or `python run.py`.
2. Select an Excel file with a "Meta" column (case-insensitive).
3. Select a folder or PDF files (`.pdf` or `.zip`) containing Kyocera QA/service documents.
4. Click "Start Processing" to:
  - Extract model numbers (e.g., `PF-740`, `TASKalfa AB-1234abcd`), QA numbers, and metadata.
  - Update blank "Meta" cells in a cloned Excel file.
  - Save text files for failed or incomplete extractions in `PDF_TXT/needs_review`.

5. Review output in `/output/cloned_<excel>.xlsx` and logs in `/logs/` or `PDF_TXT/needs_review`.

## Custom Pattern Management

- Click **Patterns** in the main window to edit regex filters stored in `custom_patterns.py`.
- Use **Re-run Flagged** to process files from the `PDF_TXT/needs_review` folder again.
- Both custom and built-in patterns are applied during each run.

## Pause/Resume & Progress Tracking

The tool now features:

- **Pause/Resume** capability for long-running jobs
- **Enhanced progress reporting** with estimated time remaining
- **Color-coded status indicators** showing current processing state
- **Detailed log view** with timestamped entries

## Development and Testing

Run tests with:

```
bash
```

```
pytest -q
```

Requires `pandas`, `PyMuPDF`, `openpyxl`, `pytesseract`, `python-dateutil`, `colorama`, `Pillow` and `opencv-python`. Ensure Tesseract is installed or in `tesseract` folder for OCR tests.

## Versioning

- Current version: **v25.1.1**
- Updates tracked in `CHANGELOG.md`.

- Use `update_version.py` to change versions:

bash

```
python update_version.py v25.1.1 v25.1.2
```

## Logging

- Session logs in `/logs/[YYYY-MM-DD_HH-MM-SS]_session.log`.
- Success/failure logs as `[YYYYMMDD]_SUCCESSlog.md` or `FAILlog.md` in `/logs/`.
- Text files for documents needing review (e.g., failed model extraction) in `/PDF_TXT/needs_review/*.txt`.

## Portable Deployment

For USB deployment:

1. Place portable Python in `python-3.11.9` folder.
2. Place portable Tesseract in `tesseract` folder.
3. Run `START.bat` to auto-detect portable dependencies.
4. No system-wide installation required.

**This is the most robust, efficient, and user-friendly version yet.**