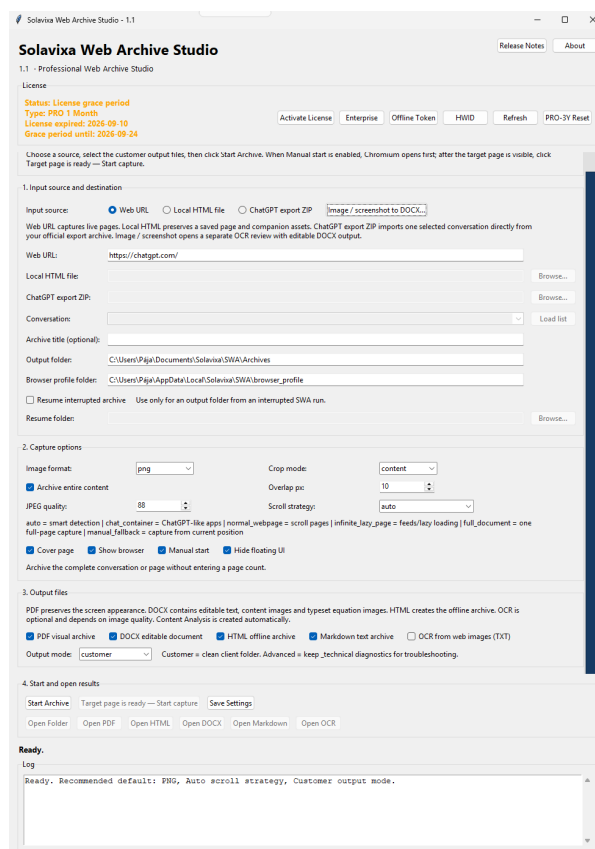


Solavixa Web Archive Studio (SWA) - Professional User Guide

Version 1.1

A practical guide for archiving web pages, local HTML files, official ChatGPT export packages, and images or screenshots, with structured PDF, DOCX, HTML, Markdown, OCR and analysis outputs.



Main application window - Solavixa Web Archive Studio 1.1

Document purpose: This guide explains the current SWA 1.1 workflow: choosing an input source, archiving entire or partial web content, creating customer outputs, preserving mathematical content, converting a standalone image or screenshot to an editable DOCX, using Auto OCR classification, reviewing OCR regions, resuming interrupted work, and checking completed results.

Product	Solavixa Web Archive Studio (SWA)
Primary use	Local archiving and structured document conversion
Supported inputs	Web URL, Local HTML file, ChatGPT export ZIP, Image / screenshot to DOCX
Key outputs	PDF, DOCX, HTML, Markdown, optional OCR TXT, content analysis, archive assets

1. Main Window Overview

The main window contains the complete archive workflow: license information, input selection, archive location, capture settings, output selection, start controls, result shortcuts, status information and a detailed run log.

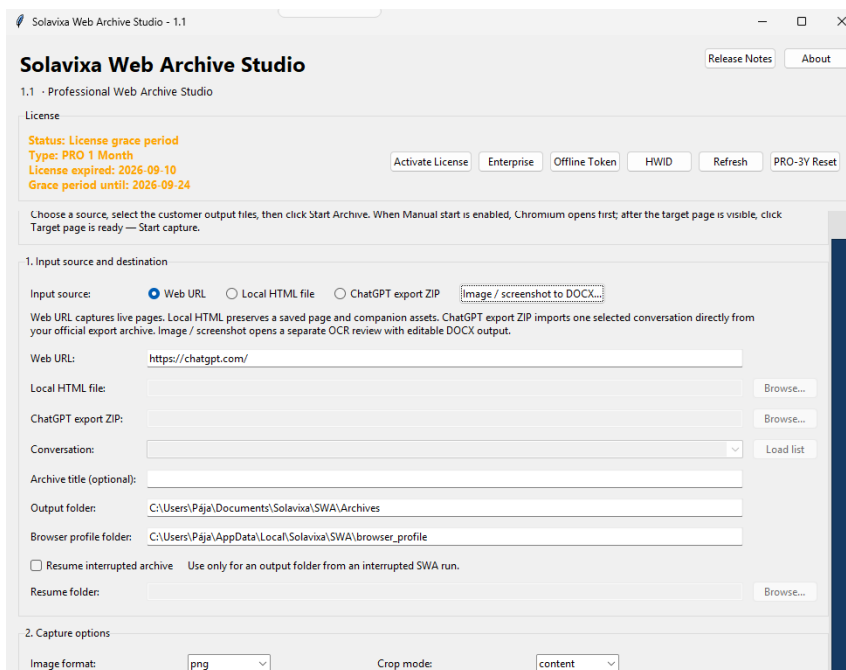


Figure 1. Current SWA 1.1 main window and input controls.

Top information and utility controls

- Release Notes. Opens information about the installed version.
- About. Shows product and publisher information.
- License panel. Shows the current license state, license type and validity information. The available buttons provide activation, Enterprise, offline-token, HWID, refresh and supported reset workflows.
- Status line. Shows the current state such as Ready, browser preparation, capture progress, OCR work or completion.
- Log. Records the main actions and messages from the current run. Check it first when an operation does not finish as expected.

Separate Image / screenshot to DOCX workflow

The Image / screenshot to DOCX... button opens a dedicated conversion workflow for a single image. This workflow is independent of the normal web-archive output checkboxes and is explained in Sections 8-11.

2. Input Sources and Archive Location

Use one primary archive input source for each normal archive run. The separate image-to-DOCX workflow is launched with its own button.

Web URL

Enter the full address of the page or web application to archive. SWA opens the page in a Chromium-based browser and captures it according to the selected scroll strategy.

Recommended for: long articles, documentation, public webpages, web applications and supported conversation pages.

Local HTML file

Click Browse and select an HTML file stored on the computer. SWA reads the local document and creates the selected archive formats. Local CSS, JavaScript, fonts and images are preserved when they are available and can be resolved from the source document.

ChatGPT export ZIP

Click Browse and select an official ChatGPT export ZIP package. Click Load list, then choose the required conversation. For this input type, SWA creates structured document outputs directly from the selected conversation; browser scrolling controls do not apply to ZIP import.

Archive and folder fields

- Archive title (optional). Adds a descriptive archive title. Leave it empty to use an automatically derived title.
- Output folder. Defines where SWA creates customer outputs and, in Advanced mode, technical diagnostics.
- Browser profile folder. Stores the browser profile used by SWA. Keeping the same profile can preserve compatible browser state between runs.
- Resume interrupted archive. Enable only when continuing an incomplete supported run.
- Resume folder. Select the output folder belonging to that interrupted run.

Privacy note: Local HTML, ChatGPT ZIP and image-to-DOCX operations use the selected local source. A Web URL archive may download page assets from the original website.

3. Capture Scope and Capture Options

For normal web capture, the recommended defaults are PNG, content crop mode, 10 px overlap, auto scroll strategy, Archive entire content enabled and Customer output mode.

Figure 2. Current capture and output controls in SWA 1.1.

Archive entire content

Enable Archive entire content when you want the complete conversation or page. In this mode, SWA removes the screen-capture count limit, so you do not need to estimate the length in A4 pages.

Important: A screen capture is the visible content recorded at one scroll position. It is not an A4 page. The final PDF, DOCX and HTML can therefore have different printed page counts.

Limited capture

Disable Archive entire content when you need only a section. SWA then uses a screen-capture limit and a capture direction. Limited capture begins at the current browser position and does not preload the entire history before recording.

Control	How to use it
Stop scrolling after	Maximum number of screen captures. The starting visible screen counts as capture 1. SWA stops without an extra scroll after the limit is reached.
Capture direction: auto	Captures upward when starting at the bottom; captures downward when starting elsewhere, including the middle of a conversation.
Capture direction: up	Captures toward older or preceding content from the selected starting position.
Capture direction: down	Captures toward newer or following content from the selected starting position.

4. Scroll Strategy and Page Preparation

Scroll strategy controls how SWA moves through long or dynamic pages.

Strategy	Recommended use
auto	Recommended default. SWA detects the most suitable capture method.
chat_container	ChatGPT-like applications where messages are inside a dedicated scrolling container.
normal_webpage	Standard pages that scroll normally from top to bottom.
infinite_lazy_page	Feeds and pages that load more content while scrolling.
full_document	Attempts a single full-document capture when the page structure allows it.
manual_fallback	Captures from the current browser position when automatic detection is not suitable.

Reading order and overlap handling

SWA merges overlapping text fragments and repeated views of the same message when a long message spans multiple screens. When capture proceeds upward, exported conversation content is still arranged in chronological reading order. Screenshots used for the visual archive are also assembled in reading order.

Additional capture controls

- Image format. PNG is recommended for text-heavy and mathematical content. JPEG can reduce size for image-heavy pages.
- JPEG quality. Applies only to JPEG captures; higher values preserve more detail and create larger files.
- Crop mode. Content focuses on the page content and removes unnecessary margins where possible.
- Overlap px. Adds a small overlap between consecutive captures to reduce the chance of missed lines.
- Cover page. Adds archive identification and metadata where supported.
- Show browser. Keep the browser visible when the page requires sign-in, preparation or monitoring.
- Manual start. Pauses after browser preparation. When the page is ready, click Target page is ready - Start capture.
- Hide floating UI. Attempts to suppress sticky controls, floating widgets and overlays that could repeat in the archive.

Tip: For a partial archive, position the browser where you want to begin before starting capture, then select the limit and direction. Capture can finish before the requested limit if the content boundary is reached or scrolling cannot continue.

5. Output Files and Output Mode

Select only the customer outputs you need. SWA follows the selected PDF, DOCX, HTML, Markdown and OCR options.

Output	What it creates
PDF visual archive	For Web URL and Local HTML, preserves the visible appearance using captured page images. For ChatGPT ZIP, creates a structured document-style PDF.
DOCX editable document	Creates editable body text with headings, lists, code blocks, original content images and preserved mathematical content where supported.
HTML offline archive	Creates readable offline HTML with available assets such as images, style sheets, scripts, fonts and KaTeX resources when applicable.
Markdown text archive	Creates a portable text-based version for editing, search, version control or further processing.
OCR from web images (TXT)	Extracts text from captured or embedded web images. This is separate from the Image / screenshot to DOCX workflow.
Content analysis	Creates structured archive analysis automatically when suitable source content is available.

Output mode

- Customer. Recommended for normal use. Creates a clean client-facing archive folder.
- Advanced. Also keeps technical diagnostics, logs, state files and supporting data for troubleshooting or verification.

Result buttons

After a successful run, use Open Folder, Open PDF, Open HTML, Open DOCX, Open Markdown or Open OCR. A result button stays unavailable when its output was not selected or was not created.

Typical customer folder

```
Archive.pdf
Archive.docx
Archive.html
Archive.md
Archive_OCR.txt (when web OCR is selected)
Archive_Analysis.txt / Archive_Analysis.json (when produced)
assets/ (when required)
```

6. Step-by-Step Archive Workflows

A. Archive a Web URL

- 1 Enter the full address in Web URL.
- 2 Choose the Output folder and optionally enter an archive title.
- 3 For a complete page or conversation, keep Archive entire content enabled. For a partial archive, disable it and set the screen-capture limit and direction.
- 4 Keep PNG, content crop, 10 px overlap and auto scroll unless the source requires different settings.
- 5 Select the required output files.
- 6 Enable Manual start when the page needs sign-in, additional loading or manual preparation.
- 7 Click Start Archive. When Manual start is enabled, prepare the page and click Target page is ready - Start capture.
- 8 Wait for completion, then open and review the created files.

B. Archive a Local HTML File

- 1 Click Browse next to Local HTML file and select the source file.
- 2 Choose the output folder and required formats.
- 3 Use PNG and auto strategy unless the source needs a different setup.
- 4 Click Start Archive.
- 5 Review PDF, DOCX, HTML and Markdown outputs and verify that required local assets are available offline.

C. Archive an Official ChatGPT Export ZIP

- 1 Obtain the official ChatGPT data export ZIP and keep the package unchanged.
- 2 Click Browse next to ChatGPT export ZIP and select the ZIP file.
- 3 Click Load list and choose the required conversation.
- 4 Select PDF, DOCX, HTML, Markdown and optional OCR according to the required deliverables.
- 5 Click Start Archive. SWA parses the selected conversation and copies supported embedded images and files.
- 6 Review the document-style PDF, editable DOCX, offline HTML, Markdown and manifest/assets outputs.

Best practice: Keep the original source file or export ZIP unchanged. Store the generated archive in a separate folder.

7. Mathematical Content and Structured DOCX Output

SWA preserves mathematical content instead of flattening it into plain text whenever the source structure allows it. This is especially important for ChatGPT-like pages containing KaTeX mathematics.

HTML mathematics

- Rendered KaTeX mathematics and the required styles and fonts are preserved in the offline HTML archive when available.
- Subscripts, superscripts, exponents, fractions, Greek symbols, operators and structured expressions are retained.
- Equation boxes, borders, aligned expressions, equation labels and formatted text inside mathematical expressions are preserved where possible.

DOCX mathematics

- Display equations are preserved as typeset equation images when required for faithful appearance.
- Supported inline mathematics can be represented with native Word equation structures while surrounding text remains editable.
- Indices, exponents, frames and text formatting are preserved in the generated document.
- Very wide equation images are scaled to fit the available document width.
- A complete equation may be retained when it crosses a capture boundary so that the mathematical meaning is not broken.

Editing note: Body text remains editable. Display equations preserved as images retain their appearance but are not editable native Word equations.

Compatibility

The generated DOCX structure avoids equation-image metadata and XML property combinations that can cause compatibility problems in older Microsoft Word versions. For best visual fidelity, use PNG for text-heavy and mathematical content.

8. Image / Screenshot to Editable DOCX

Use the Image / screenshot to DOCX... button when the source is a standalone image rather than a webpage or HTML archive. This dedicated workflow combines OCR text recognition with preservation of visual regions that should remain images.

Supported source images

Open a single PNG, JPEG, WebP, BMP or TIFF image. Multi-page TIFFs and animated files must first be separated into individual images.

Local processing

Recognition uses the locally installed Tesseract engine and installed language data. The workflow does not require an online OCR service, API key or additional downloaded AI model.

Language selection

Choose from the OCR languages installed on the computer. For mixed English and Czech content, eng+ces can be used when both language data sets are installed. Missing engines or language data produce an explicit error.

Two OCR approaches

Mode	Recommended use
Auto	Recommended starting point. SWA runs OCR and proposes which regions should stay editable text and which should remain original image crops.
Run OCR with manual region selection	Use when you prefer to mark equations, symbols, icons or other visual regions yourself before final export.

Recommended workflow: Open the image -> select OCR languages -> click Auto -> review text and image selections -> correct any regions -> save DOCX.

9. Auto OCR and Visual Region Classification

Auto is designed to create a practical editable document while protecting source elements that OCR may damage if they are converted to guessed text.

What Auto evaluates

- OCR confidence and recognized text.
- Mathematical notation and symbol-like content.
- Colored visual content and regions not adequately recognized by OCR.
- Display equations, boxed formulas, emoji, graphics and uncertain text.
- Straight horizontal and vertical rules that can be reconstructed as Word borders where supported.

How mixed text and image content is handled

- Detected display equations can be preserved as complete image regions, including surrounding frames.
- Inline mathematical symbols can remain as image crops while neighboring ordinary words stay editable.
- Where a variable prefix is attached to ordinary text, SWA attempts to find a usable boundary between the preserved visual part and editable text.
- Uncertain OCR can remain as original image content instead of being replaced by guessed characters.
- Overlapping or nearby visual fragments can be merged to avoid duplicate overlays and to protect neighboring text.
- Automatic classification reasons are shown in the region list together with counts of editable-text and preserved-image regions.

Important: Auto is a heuristic proposal, not a guarantee of semantic correctness or pixel-identical reconstruction. Always review the result before final use.

Why review is necessary

Mathematical Latin letters and ordinary letters can look identical, and OCR can report high confidence for incorrect recognition. Some symbols may need manual preservation, while some ordinary words preserved as images may be better converted back to editable text.

10. Review and Correct OCR Regions

The OCR editor is the review stage between recognition and final DOCX export. Use it to correct text, formatting and region placement before saving.

Review controls

- Source and layout preview. Review the original image and an approximate reconstructed layout with zooming and scrolling.
- Edit text. Correct recognized words and characters directly.
- Typography. Adjust font family, size, bold and italic formatting.
- Region geometry. Edit a region's position and dimensions when the detected placement is inaccurate.
- Preserve image. Draw or adjust an area that should remain an original image crop, such as a full equation and its frame.
- Convert selected image to text. Retry OCR on a preserved image region. Use Undo if the recognition result is worse than the image.
- Add text. Create a missing editable text region.
- Split text. Divide a text region so different parts can use different formatting or placement.
- Delete region. Remove an unwanted region.
- Undo. Revert recent editing or OCR replacement actions.

Uncertain OCR

Uncertain OCR is highlighted in the region list. When a manually preserved crop intersects part of a recognized word, SWA attempts to recover the remaining text where possible and marks affected fragments for review.

Save and reopen a review

You can save the review as a .swa-ocr.json file and reopen it later with the matching original image. SWA verifies the source-image fingerprint so a review is not accidentally applied to a different image.

Keep the original image. The review file does not embed the source image, so keep the original image alongside the saved review.

Responsiveness

OCR and conversion work runs in the background so the editor remains responsive. Replacing existing OCR regions requires confirmation and can be undone.

11. Export the Image Conversion to DOCX

The image workflow exports a hybrid DOCX containing positioned editable text together with original image crops for equations, symbols, emoji, graphics or other preserved regions.

DOCX layout

- Recognized editable text keeps its source position with suggested typography.
- Preserved visual regions are cropped from the original source pixels rather than redrawn or approximated as text.
- Straight rules can be represented as Word borders where supported.
- Long source images use A4 pagination.
- Preserved equation regions are kept together across page boundaries where possible.
- Oversized preserved regions are scaled to fit the page.
- Sufficient paragraph height is reserved around image crops so borders and framed equations are not clipped.

Optional TXT output

You may also save recognized editable text to a separate TXT file. Text that remains inside image-only regions is not transcribed into this TXT output.

Safe file creation

SWA writes the DOCX through a temporary file and replaces the destination only after successful generation, reducing the risk of leaving a partially written document.

Result buttons and licensing

After a successful image conversion, the main-window result buttons are updated for the generated DOCX and optional TXT output. The normal license or trial check applies before export, and successful image conversion counts as trial use where applicable.

Expected limitations

- Font matching, emphasis and spacing are approximate.
- Word may render positioned text differently from the OCR editor preview or source screenshot.
- Editable text uses positioned frames to retain layout; it does not behave like a normal continuously reflowing document.
- Preserved equations and symbols are images, not editable native Word equations.
- Low-resolution, blurred, low-contrast or visually complex images can require substantial review.
- Content already cropped out of the source image cannot be recovered.

12. Resume an Interrupted Archive

Resume can continue a supported archive run after a crash, browser interruption, system restart or other incomplete operation. It uses state stored in the interrupted output folder.

- 1 Do not delete or rename files inside the interrupted output folder.
- 2 Restart SWA and enable Resume interrupted archive.
- 3 Click Browse next to Resume folder and select the interrupted archive folder.
- 4 Keep the original input and output selections when possible.
- 5 Click Start Archive and follow any browser preparation message.
- 6 After completion, review the end of the PDF and the final sections of DOCX, HTML or Markdown to confirm that the archive is complete.

Do not use Resume to append unrelated content to a completed archive. Create a new output folder for every new archive task.

Recommended Settings by Content Type

Content type	Recommended settings
Normal article or documentation	PNG, content crop, auto or normal_webpage, 10 px overlap, Archive entire content, Customer mode.
ChatGPT-like web application	PNG, auto or chat_container, Show browser, Manual start when preparation is needed, Hide floating UI.
Partial conversation section	Disable Archive entire content, position the browser, set screen-capture limit and choose auto/up/down direction.
Long feed or lazy-loaded page	PNG, infinite_lazy_page, Archive entire content or a sufficient partial-capture limit, Show browser for monitoring.
Local HTML document	PNG, auto, HTML + DOCX + PDF, verify offline assets after completion.
Official ChatGPT export ZIP	PDF + DOCX + HTML + Markdown; OCR only when embedded images need text extraction.
Image / screenshot to DOCX	Choose OCR languages, Auto, review text/image regions, correct layout, then save DOCX; optional TXT.

13. Practical Tips and Troubleshooting

The browser opens but the target content is not ready

Enable Show browser and Manual start. Wait for loading to finish, close overlays, open the required content and then start capture.

A partial capture archives the wrong section

Disable Archive entire content, move the browser to the intended starting position and check capture direction. In auto mode, starting at the bottom captures upward; starting elsewhere captures downward.

The archive is missing the end of a long page

For full content, enable Archive entire content. For limited capture, increase the screen-capture limit or use a strategy that better matches the page type.

Floating buttons repeat on many PDF pages

Enable Hide floating UI and close cookie banners, menus, chat widgets and sticky overlays before capture.

The PDF is very large

Create only the outputs you need. For image-heavy pages, JPEG can reduce size; PNG remains recommended for text-heavy or mathematical content.

The offline HTML is readable but not fully interactive

Offline HTML preserves readable content and available assets. It does not reproduce live application behavior, API services or every interactive element.

A ChatGPT ZIP does not import

Use the original official export package, keep its structure unchanged and confirm that it contains the exported conversation data.

OCR from web images is empty or inaccurate

Confirm that web-image OCR was selected and that the required OCR engine/language data is available. Recognition quality depends on source quality.

Image Auto OCR preserves too much or too little

Review the region list. Convert a preserved image region back to text when OCR is usable, or use Preserve image when equations, symbols, emoji or uncertain text should remain visual.

A result button is disabled

The corresponding output was not selected, was not created or the run did not complete. Review the Status and Log areas.

14. Licensing, Updates and Usage Notice

Licensing

Available functions depend on the active Trial, PRO or Enterprise license. Use the activation controls in the License panel. The application displays the active license state and relevant validity information.

Saved settings and capture consistency

SWA saves the selected output formats and capture-direction settings. At the beginning of a run, the selected settings are captured and conflicting controls are disabled while background work is in progress so mid-run changes cannot produce inconsistent results.

Review Release Notes after an update

Capture behavior, supported outputs, OCR behavior and document compatibility can improve between versions. Review Release Notes after updating SWA.

Legal and usage notice

Solavixa Web Archive Studio is intended for lawful archiving, documentation and document-conversion use only. Users are responsible for ensuring that they have the right to access, copy, store, process and share archived content. Respect copyright, confidentiality, personal-data protection, website terms, platform rules and applicable law. Do not distribute authentication data, private exports or personal information without authorization.

Image OCR responsibility

OCR and automatic visual-region classification are assistance tools. Always review important output. Recognition and layout reconstruction can be imperfect, especially with mathematical symbols, complex formatting, small text, low contrast or low-resolution images.

Solavixa Web Archive Studio (SWA)
Professional User Guide - Version 1.1
Last updated: September 2026