

Website Image Licensing Audit — Staff How-To (One-Pager)

Purpose:

Quickly scan a website for images/graphics that may require licensing review.

Who should use this:

Any staff member tasked with reviewing site imagery before launch or after content changes.

Steps:

- 1) Open the Streamlit app URL (shared internally).
- 2) Enter the site's Start URL (e.g., <https://example.com>).
- 3) Use defaults (good for most sites): Depth 3 • Max pages 100 • Max images 1000 • Delay ≥ 300 ms • Concurrency 5.
- 4) Ensure "Capture CSS backgrounds" and "Attempt EXIF/IPTC" are ON.
- 5) Click 'Run Audit'. You may stop anytime; results so far will remain.
- 6) Filter results (Stock-only / Risk-only / Search).
- 7) Use the Google Images / TinEye links for quick reverse checks.
- 8) Download CSV/Excel and file it with your project docs.

Flags you'll see:

- Stock source — ensure license: Image is hosted on a known stock/library (Shutterstock, Getty, Adobe Stock, iStock, Pexels, Unsplash, etc.).
- No alt text (check provenance): Missing alt text; verify the image's origin and license.

Passcode (if enabled):

If the app asks for a passcode, enter the staff passcode provided by your administrator.

Limits (why the scan might stop early):

- Hit Max pages/images or Total download cap (raise limits if needed).
- robots.txt blocked some paths (cannot crawl disallowed areas).
- Very large images exceeded the per-image cap (default 5 MB).

Best practices:

- Be respectful: keep base delay ≥ 300 ms.
- Keep scope to the same domain unless you know subdomains are in-scope.
- For heavy JS sites (SPAs), ask admin about a Playwright-enabled version.

Troubleshooting:

- "No images found": check URL is public, raise depth/pages, enable subdomains.
- EXIF not shown: many sites strip EXIF; also check size caps.
- Dynamic content missing: ask admin about a headless-browser add-on.

Contact:

For issues or feature requests, open a GitHub issue or contact the project admin.