HOW TO: preparing one RAG document

A practical, copy-ready standard for developers and non-developers

What we are creating

We need one consolidated, upload-ready RAG document for one brand or page. Markdown (`.md`) is the knowledge format our AI uses. You may upload a searchable PDF (`.pdf`) if that is what you have; we extract its text and prepare Markdown for review. The PDF we generate from Markdown is for internal preview and future reference only. Some characters may render incorrectly in that copy, so check the Markdown when reviewing facts. We do not need a multi-file package, a ZIP, or a separate manifest.

The easiest workflow is:

1. Collect the URLs for the approved brand pages.
2. Copy the prompt below into ChatGPT or another LLM that can open web pages.
3. Replace the single `SOURCE URLS` placeholder with every approved page URL. Nothing else needs to be filled in.
4. Ask the LLM to produce exactly one Markdown document.
5. Review it, save it as one UTF-8 `.md` file, then attach it in Configuration > Knowledge. If you only have a searchable PDF, you may upload that instead; we will screen it through the same review process.

Only include public or otherwise approved material. Do not paste credentials, secrets, private customer information, unpublished claims, or personal data into an LLM.

Copy-and-paste prompt

```
We are preparing one consolidated RAG knowledge document for our content automation system.

SOURCE URLS:
- [Paste the first approved page URL]
- [Paste the second approved page URL]
- [Continue until every approved page is listed]

YOUR TASK

Open and read every URL listed above. Infer the brand/page name, markets and languages only from the
supplied pages, and mark them as unclear when the pages do not establish them. Use the actual page content,
not only the page title, search snippet, navigation label, or link anchor. Capture all useful approved
information from every supplied page, including important headings, body copy, product or service details,
audiences, claims, calls to action, safety boundaries, visual guidance, trust information, locations, and
full destination URLs.

Return exactly ONE complete Markdown document. Do not split the response into multiple files. Do not create
a ZIP, JSON manifest, companion file, or separate source package. Do not place commentary before or after
the document.

CRAWLING AND COVERAGE RULES

1. Process every distinct supplied URL. Remove tracking parameters and skip repeated links or pages with the
same canonical address or content. If two URLs point to different sections with different material, keep
both.
2. Preserve each supplied URL in full. Never replace a URL with vague text such as “Read the article”,
“Learn more”, or “Source”.
3. Before writing, create an internal URL inventory from each supplied page. Inspect the navigation, body
links, accordions, tabs, FAQs, contact links, footer, resource/article links, relevant downloadable
```

```
documents, canonical links and any accessible sitemap. Follow all relevant same-domain pages needed to
understand the brand, services, claims, CTAs, safety, trust, markets, fonts or official assets. Do not
silently stop at the supplied pages.
4. Preserve every discovered relevant URL in the source register, including pages that are blocked,
duplicated, excluded or unavailable. Mark duplicates clearly without fetching the same content twice. Do
not wander to unrelated third-party pages.
5. Expand accordions, tabs, FAQs, footnotes, and linked sections when the browsing tool can access them.
6. If a page is unavailable, blocked, empty, duplicated, or conflicts with another source, record that
clearly in `Gaps and conflicts`. Never guess missing content.
7. Keep exact numbers, dates, qualifications, eligibility rules, legal or safety wording, and geographic
limits attached to their source URLs.
8. Do not invent facts, testimonials, customers, metrics, prices, certifications, clinical claims,
features, or asset URLs.
9. Distinguish approved facts from inferred guidance. Omit unsupported inferences from the approved-facts
sections.
10. Write brand-owned guidance with first-person plural pronouns such as “we”, “our”, and “us”; do not
describe the brand as an unrelated third party.
11. Remove cookie banners, repeated menus, footer boilerplate, tracking parameters, and unrelated
navigation text from the knowledge prose, while retaining relevant footer/contact destination URLs in the
source register.

REQUIRED DOCUMENT STRUCTURE

# [Brand or page name] - Consolidated RAG Knowledge Base

## Document control
- Brand or page:
- Primary market:
- Primary language:
- Prepared date:
- Review status: DRAFT - HUMAN APPROVAL REQUIRED
- Source count:

## Brand identity and positioning
State who we are, what we do, where we operate, and the positioning supported by the sources.

## Products, services, and approved claims
Capture each offering, feature, benefit, eligibility rule, limitation, proof point, and approved factual
claim. Put the supporting URL directly beside each consequential claim.

## Audiences and use cases
Describe the approved audiences, their needs, and the situations the sources explicitly address.

## Voice and messaging guidance
Record tone, preferred terminology, first-person “we/us” phrasing, wording to avoid, and representative
source-grounded examples.

## Calls to action and destination URLs
List each approved CTA with its exact visible wording, purpose, market or audience, and full destination
URL. A label without its URL is incomplete.

## Visual identity and official assets
Record approved fonts, logo rules, image style, accessibility guidance, and every official asset URL found
in the sources. Do not invent a colour palette when none is provided. Preserve likely primary-brand and
sub-brand ownership when the evidence distinguishes them.

## Safety, trust, legal, and prohibited claims
Capture privacy, security, clinical or wellbeing boundaries, escalation language, disclaimers, regulated
wording, and claims we must not make.

## Geographic and language guidance
Separate any market-specific services, spelling, terminology, eligibility, URLs, contacts, or restrictions.

## Page-by-page knowledge
Create one subsection for every supplied URL. Include:
- Page title
- Full URL
- Page purpose
```

```
- Detailed, non-duplicative knowledge extracted from that page
- Claims or CTAs unique to that page
- Relevant same-domain links opened from that page

## Source register
List every supplied or directly relevant followed URL, its page title, what it supports, and whether it was
successfully accessed.

## Gaps and conflicts
List unavailable pages, missing facts, ambiguous wording, contradictions, expired information, unsupported
claims, and decisions requiring human approval. Write “None found” only if the review genuinely found none.

FINAL QUALITY CHECK

Before returning the document, verify that every supplied URL and every relevant followed URL appears in the
source register; every successfully opened knowledge page appears in the page-by-page section; every
important claim and CTA retains a full URL; no supplied page was silently skipped; the result is one
document; and no unsupported content was added.
```

Before we upload

We should review the generated document as brand owners. Confirm that every URL was processed, claims are accurate, CTA destinations are complete, and unresolved gaps are visible. Remove anything that is not approved.

Save the result as one UTF-8 Markdown file. A searchable PDF is accepted as an input when Markdown is unavailable, but the system will extract text and make Markdown the reviewed knowledge format. A PDF made only from scanned images will not work because the system must be able to extract its text.

Attach the document in Configuration > Knowledge, paste page links, or combine both in one preparation job. We will screen prepared and unformatted files alike and place the seven required brand checks inside one complete editable RAG document. The crawler downloads only assets that the page identifies as a logo, wordmark or reusable brand asset; it does not collect ordinary page photos, blog thumbnails, decorative icons, favicons or font files. You may also attach PNG, JPEG or WebP assets intentionally when needed.

Reviewers do not need to understand Markdown. In non-developer mode, the complete RAG opens as a familiar formatted document with readable headings, paragraphs, lists and links. The system preserves and updates Markdown behind the scenes; raw Markdown is available only in developer mode. Reviewers may also leave general comments beside the document. Comments stay outside the AI knowledge but appear in the designed reference PDF. The PDF itself is a read-only cross-browser reference: use the dashboard document editor for reliable edits and comments.

What success looks like

- One brand or page produces one consolidated RAG document.
- Every supplied page is represented in the page-by-page section.
- Important facts and CTAs retain their full source URLs.
- The document contains searchable text and fits the limits displayed in Configuration.
- Human reviewers can see gaps or conflicts instead of hidden guesses.
- Human reviewers edit one formatted document without handling Markdown syntax and see a compact seven-area completeness check rather than seven separate editing forms.
- Human reviewers can preview only meaningful staged assets, correct labels/tags and approve more than one logo or explicitly proceed without one.
- Knowledge and assets belong only to the selected brand or page.