

01-paperless-ngx.md

kstack: book: Centerpoint Home
Lab chapter: Documents &
Organization page: Paperless-NGX
tags: [paperless, documents, ocr,
archive, productivity]

Overview

Paperless-NGX is the centralised document management and archival system. It ingests scanned documents and PDFs via a watched consume directory and email, applies OCR, tags, correspondents, and document types, then stores the processed results in a structured archive. Gmail OAuth is configured for email ingestion. SSO is provided via Authentik OpenID Connect.

Access

Type	URL	Auth
External	<code>https://paperless.jeeves5454.ddns.net</code>	Authentik OIDC + GeoBlock + CrowdSec
Internal	<code>https://paperless.home.local</code>	Step-CA TLS (internal)

Authentication uses Authentik OIDC (`openid_connect` Django allauth provider). Regular local login is kept enabled (`PAPERLESS_DISABLE_REGULAR_LOGIN=false`). Auto-signup for new OIDC users is disabled — accounts must be pre-created.

Containers

Five containers in this stack:

Container	Image	Role
paperless-webserver-1	ghcr.io/paperless-ngx/paperless-ngx:latest	Web UI + workers
paperless-db-1	postgres:16	Primary database
paperless-broker-1	redis:7	Celery task queue
paperless_gotenberg	gotenberg/gotenberg:8.27	DOCX→PDF conversion
paperless_tika	apache/tika:latest	Content extraction

paperless-webserver-1

Main application container running both the Django web server and Celery workers.

Key environment variables:

Variable	Value / Notes
PAPERLESS_URL	https://paperless.jeeves5454.ddns.net
PAPERLESS_CSRF_TRUSTED_ORIGINS	https://paperless.jeeves5454.ddns.net
PAPERLESS_OAUTH_CALLBACK_BASE_URL	https://paperless.jeeves5454.ddns.net
PAPERLESS_DBHOST	db
PAPERLESS_REDIS	redis://broker:6379
PAPERLESS_TIKA_ENABLED	1
PAPERLESS_TIKA_ENDPOINT	http://tika:9998
PAPERLESS_TIKA_GOTENBERG_ENDPOINT	http://gotenberg:3000
PAPERLESS_TASK_WORKERS	2
PAPERLESS_THREADS_PER_WORKER	2
PAPERLESS_TIME_ZONE	America/Toronto
PAPERLESS_OCR_LANGUAGE	eng
PAPERLESS_CONSUMER_RECURSIVE	true
PAPERLESS_CONSUMER_SUBDIRS_AS_TAGS	true
PAPERLESS_CONSUMER_POLLING	5 (seconds)
PAPERLESS_FILENAME_FORMAT	{{ created_year }}/{{ document_type }}/{{ created_year }}-{{ created_month }}-{{ created_day }}_{{ correspondent }}_{{ title }}

Variable	Value / Notes
PAPERLESS_FILENAME_FORMAT_REMOVE_NONE	true
PAPERLESS_SOCIAL_AUTO_SIGNUP	false
PAPERLESS_ACCOUNT_EMAIL_VERIFICATION	none
PAPERLESS_GMAIL_OAUTH_CLIENT_ID	(visible — treat as semi-public)
PAPERLESS_GMAIL_OAUTH_CLIENT_SECRET	REDACTED
PAPERLESS_APPS	allauth.socialaccount.providers.openid_connect
OIDC provider client secret	REDACTED (inside PAPERLESS_SOCIALACCOUNT_PROVIDERS)

Bind mounts:

Host Path	Container Path	Purpose
/home/jeeves/docker/paperless/data	/usr/src/paperless/data	Index and SQLite state
/mnt/data/paperless/media	/usr/src/paperless/media	Archived document files
/mnt/data/paperless/consume	/usr/src/paperless/consume	Watched consume folder
/mnt/data/paperless/export	/usr/src/paperless/export	Export output folder

“ The media, consume, and export directories live on the NFS-equivalent Ceph volume at /mnt/data/, providing separation from the system drive.

paperless-db-1

PostgreSQL 16 database backend.

Bind mounts:

Host Path	Container Path
/home/jeeves/docker/paperless/pgdata	/var/lib/postgresql/data

paperless-broker-1

Redis 7 for Celery task queue.

Bind mounts:

Host Path	Container Path
-----------	----------------

paperless_gutenberg

Gutenberg 8.27 — converts Office documents (DOCX, XLSX, etc.) to PDF for ingestion. No persistent volumes.

paperless_tika

Apache Tika — extracts content and metadata from complex document formats. No persistent volumes.

Traefik Labels

```
traefik.http.routers.paperless-external.rule: Host(`paperless.jeeves5454.ddns.net`)
traefik.http.routers.paperless-external.entrypoints: websecure
traefik.http.routers.paperless-external.tls.certresolver: letsencrypt
traefik.http.routers.paperless-external.middlewares: plex-geoblock@file,crowdsec-bouncer@file

traefik.http.routers.paperless-internal.rule: Host(`paperless.home.local`)
traefik.http.routers.paperless-internal.entrypoints: websecure
traefik.http.routers.paperless-internal.tls.certresolver: step-ca
```

Notes / Gotchas

- Secrets in `PAPERLESS_SOCIALACCOUNT_PROVIDERS` (OIDC client secret) and `PAPERLESS_GMAIL_OAUTH_CLIENT_SECRET` are in plaintext in the Portainer-managed compose environment — rotate if the Portainer env is ever exposed.
- The consume directory polling interval is 5 seconds. Drop files into `/mnt/data/paperless/consume` (or a subdirectory — subdirectory names become tags automatically via `PAPERLESS_CONSUMER_SUBDIRS_AS_TAGS`).
- Tika and Gutenberg must be running for complex document formats to process; plain-text PDFs work without them.
- The `PAPERLESS_LOGOUT_REDIRECT_URL` is set to the Authentik end-session endpoint — clicking Log Out in Paperless also terminates the Authentik session.

Revision #2

Created 2026-06-17 04:45:51 UTC by Admin

Updated 2026-06-17 04:45:51 UTC by Admin