Dataset Specification — French Premium Web Corpus v1.2.0

EU AI Act Article 10 / Annex IV documentation bundle

RELEASE IDENTIFIER

`fpwc-v1.2.0-2026-05-15`

PIPELINE CONTENT HASH (SHA-256)

`9071408943537d14e99381c3265a14cce43eeb077942b53ea38344406eb5d98d`

SPECIFICATION VERSION

1.1

DOCUMENT DATE

2026-05-15

DOCUMENT OWNER

FINALEADS LLC (Wyoming, United States) — operating brand *French Corpus LLM*

COMPLIANCE CONTACT

compliance@frenchcorpus.com

Reference framework: Regulation (EU) 2024/1689 of the European Parliament and of the Council (AI Act), Article 10. Complementary methodology: Gebru et al., *Datasheets for Datasets* (2018, arXiv:1803.09010).

Distribution: This document is part of the deliverable bundle for the FPWC release identified above. Any reproduction or redistribution must preserve the bundle integrity and the cryptographic signature embedded in this PDF.

Contents

PART I — MASTER DATASET SPECIFICATION

Dataset Specification Document — French Premium Web Corpus, Finance / Regulation / Economic Edition

Release identifier: `fpwc-v1.2.0-2026-05-15` **Pipeline content hash (SHA-256):** `9071408943537d14e99381c3265a14cce43eeb077942b53ea38344406eb5d98d` **Specification version:** 1.3 **Document date:** 2026-05-15 **Document owner:** FINALEADS LLC (Wyoming, United States) — operating brand *French Corpus LLM* **Contact:** compliance@frenchcorpus.com **Reference framework:** Regulation (EU) 2024/1689 of the European Parliament and of the Council (AI Act), Article 10. Complementary methodology: Gebru et al., *Datasheets for Datasets* (2018, arXiv:1803.09010).

Executive summary

The French Premium Web Corpus (FPWC) is a vertical training corpus assembled exclusively from European Union public sector primary sources, dedicated to French-language finance, prudential regulation, and economic text. Release `fpwc-v1.2.0-2026-05-15` contains **2 148 446 documents** totalling **9.48 billion characters** (approximately **2.37 billion tokens**) sourced from eight institutional providers (DILA, EUR-Lex, DGFiP, ACPR, AMF, Banque de France, DGTrésor, Court of Justice of the European Union through joelniklaus/eurlex_resources). DILA now contributes six distinct bulk datasets — Légifrance JORF, Légifrance LEGI, the *Cour de cassation* jurisprudence stock (CASS), the *Conseil d'État* jurisprudence stock (JADE), the *Conseil constitutionnel* decisions stock (CONSTIT), and the National Collective Agreements stock (KALI). The dataset is delivered in four tiers (Sample / Standard / Premium / Enterprise) and ships with per-document JSON-LD provenance records aligned with the W3C PROV-O vocabulary.

This document satisfies the data governance documentation obligation set out in Article 10 of the EU AI Act for any high-risk AI system trained, validated, or tested with the FPWC. The ten sections that follow are mapped one-to-one against Article 10's required practices (paragraphs 2(a) through 2(h)) plus the supplementary requirements on quality, contextual specificity, and special categories of personal data (paragraphs 3, 4, and 5).

The dataset is suitable for fine-tuning and continued pre-training of large language models intended for deployment in regulated financial and economic services within the European Union. It is **not suitable, without further curation, for any system that processes special categories of personal data, for biometric identification, or for high-risk systems outside the financial-regulatory-economic domain**.

1. Dataset identifier

Field	Value
Canonical name	French Premium Web Corpus — Finance / Regulation / Economic Edition
Short identifier	FPWC
Release identifier	`fpwc-v1.2.0-2026-05-15`
Release date	2026-05-15
Build wall-time	2026-05-14 19:00 CEST
Pipeline content SHA-256	`9071408943537d14e99381c3265a14cce43eeb077942b53ea38344406eb5d98d`
Source code commit	FINALEADS LLC proprietary repository (access available to commercial licensees under NDA)
Build infrastructure	FINALEADS LLC managed compute on OVH Cloud SBG5 datacenter (Strasbourg, France)
Build environment	Python 3.11.6, PyArrow 17.0.0, Pandas 2.2.3, Transformers 4.57.6
Producer entity	FINALEADS LLC, Wyoming, USA (operating as *French Corpus LLM*)
Persistent identifier	A DOI will be minted via Zenodo upon listing approval on Snowflake Marketplace

The pipeline content hash uniquely identifies the combination of source data state, pipeline code state, and configuration that produced the release. Any modification to source materials, pipeline code, or configuration will change the hash. The hash is embedded in every per-tier `MANIFEST.json` and in every per-document JSON-LD provenance record, enabling cryptographic verification of release integrity.

---

2. Intended purpose

2.1 Primary intended uses

The FPWC is constructed for the following primary uses:

1. **Continued pre-training** of decoder-only large language models intended for deployment in French-language financial, regulatory, or economic contexts.
2. **Supervised fine-tuning (SFT)** of base models, when paired with instruction-response derivations of the corpus content (the present release ships raw text only; SFT derivations are an explicit out-of-scope post-processing step left to the consumer).
3. **Retrieval-augmented generation (RAG)** systems requiring authoritative French-language sources for finance, regulation, and economic policy.
4. **Benchmark construction** for vertical evaluation of French-language LLMs on financial-regulatory tasks.
5. **Auditing and red-teaming** of existing models claiming competence in French finance or regulation.

2.2 Out-of-scope uses

The FPWC is **not intended for**, and the producer disclaims fitness for, the following uses without substantial additional curation:

- Any AI system processing **special categories of personal data** within the meaning of GDPR Article 9. **ACPR documents are pseudonymized at the source level** — natural persons referenced with civil titles (M./Mme/Maître...) have their names replaced with per-document bracketed aliases (e.g., `[P1]`, `[P2]`). EUR-Lex case law may still contain personal names (not pseudonymized). A consumer planning such use must perform its own redaction and Article 9 compliance assessment for sources other than ACPR.
- **Biometric identification, categorisation, or emotion recognition** systems.
- **Critical infrastructure safety** systems where French legal text is not in the operating envelope.
- **Real-time decisioning** systems where the corpus's monthly refresh cadence is insufficient.
- **General-purpose conversational systems** outside the French legal-financial-economic domain. The corpus contains effectively zero general dialogue, narrative fiction, or non-technical text.

2.3 Annex III high-risk mapping

The FPWC is designed primarily for use cases falling under the following AI Act Annex III categories:

- **Annex III, point 5(b)** — AI systems used to evaluate creditworthiness and credit scores.
- **Annex III, point 5(c)** — AI systems used in risk assessment and pricing in life and health insurance.
- **Annex III, point 8** — AI systems intended for use by judicial authorities for researching and interpreting facts and the law. (The corpus contains the underlying legal texts; downstream judicial decision support is the consumer's responsibility.)

A consumer deploying an AI system into any of the above contexts inherits Article 10 obligations on the training data; this Dataset Specification Document is intended to support, but does not by itself discharge, those obligations.

---

3. Composition

3.1 Aggregate composition

Metric	Value
Total documents	2 148 446
Total characters	9 480 632 955
Estimated tokens (chars / 4)	~2 370 158 239
Unique source providers	8 (DILA contributes six distinct bulk datasets)

Metric	Value
Time range covered	1947–2026 (1958–2022 for EUR-Lex; 1990–2026 for most French sources; 1790–2026 for *Cour de cassation* historical stock)
Primary language	French (fr-FR)
Other languages incidentally present	Latin (legal citations), English (foreign treaty references), <0.5 % combined
Document format	Plain UTF-8 text
Storage format on disk	Apache Parquet (zstd-compressed), JSONL for sample tier

3.2 Per-source breakdown (Premium tier composition)

Source	Provider	Documents	Share	License	Time range
`jade`	DILA — *Conseil d'État* jurisprudence stock	481 446	37.74 %	Licence Ouverte 2.0	1990–2026
`legifrance_jorf`	DILA — *Journal officiel de la République française*	466 242	36.55 %	Licence Ouverte 2.0	1947–2026
`cass`	DILA — *Cour de cassation* jurisprudence stock	121 543	9.53 %	Licence Ouverte 2.0	1790–2026 (historical stock)
`kali`	DILA — National Collective Agreements stock	99 431	7.79 %	Licence Ouverte 2.0	1980–2026
`eurlex_fr`	Publications Office of the European Union (via joelniklaus/eurlex_resources HF)	79 546	6.24 %	Reuse permitted under EU Decision 2011/833/EU; CC-BY-4.0 derivative	1958–2022
`legifrance_legi`	DILA — Consolidated legal codes	18 369	1.44 %	Licence Ouverte 2.0	Current consolidated codes as of 2026-04
`bofip`	DGFiP (Direction générale des Finances publiques)	5 679	0.45 %	Licence Ouverte 2.0	2012–2026
`constit`	DILA — *Conseil constitutionnel* decisions stock	2 788	0.22 %	Licence Ouverte 2.0	1958–2026

Source	Provider	Documents	Share	License	Time range
`acpr`	Autorité de contrôle prudentiel et de résolution	488	0.04 %	Licence Ouverte 2.0	2010–2026
`bdf`	Banque de France	60	< 0.01 %	Licence Ouverte 2.0 (institutional publications)	2018–2026
`dgtresor`	Direction générale du Trésor	36	< 0.01 %	Licence Ouverte 2.0	2020–2026
`amf`	Autorité des marchés financiers	0 (administrative content excluded)	—	Licence Ouverte 2.0	2010–2026

The four DILA jurisprudence and convention-collective datasets (CASS, JADE, CONSTIT, KALI) are new in `fpwc-v1.2.0-2026-05-15`. They are bulk-extracted from DILA's *Freemium* global snapshot archives (the most authoritative stock available outside the API) and re-processed end-to-end through the same MinHash deduplication and quality-scoring pipeline as the rest of the corpus. The AMF source contributes zero rows in the Premium tier of this release; a future release will re-evaluate the topical filter's interaction with AMF's administrative bulletin language.

3.3 Per-tier composition

Tier	Documents	Tokens (estimated)	Source mix policy	Intended use
Sample	500	~620 000	Stratified sample preserving full source proportions	Free preview, lead-generation, evaluation prior to purchase
Limited Trial	50 000	~50 000 000	Stratified sample preserving full Premium-tier proportions across all 11 sources	In-place evaluation on Snowflake Marketplace, login-gated, 30-day window per consumer account, free of charge
Standard	2 148 446	2 370 158 239	Full corpus, no quality filtering beyond stages 1-4b	Continued pre-training of vertical LLMs
Premium	1 275 628	1 891 836 486	Top-quality slice selected by distilled quality classifier composite score	Fine-tuning and RAG production indexing

Tier	Documents	Tokens (estimated)	Source mix policy	Intended use
Enterprise	2 148 446	2 370 158 239	Same documents as Standard, plus quarterly refresh contract, plus optional AI Act audit certification add-on	Long-term contractual deployment in regulated industries

The Premium tier's "top-quality slice" is defined as documents whose composite quality score (a weighted combination of coherence, legal density, finance value, and inverse AI-slop signal, predicted by the distilled DistilCamemBERT classifier trained on Qwen2.5-7B teacher labels) falls in the upper 59 % of the full corpus distribution. The threshold was chosen to maximise tier separation while preserving statistical representativeness of each source.

3.4 Geographic and contextual coverage

Dimension	Coverage
Legal system	French Republic + European Union law applicable in France
Currency	Euro (EUR); historical references to French Franc (FRF) prior to 2002
Regulatory bodies covered	ACPR, AMF, Banque de France, DGFiP, DGTrésor, plus EU bodies (European Commission, EBA, ESMA, EIOPA, ECB indirectly through EU regulations)
Sub-vertical coverage	Banking, insurance, capital markets, taxation, monetary policy, prudential supervision, anti-money laundering, market abuse, sustainability reporting
Notable inclusions	DORA, MiCA, AI Act, CSRD, ESRS, AMLD6, MiFID II review, CRR3, CRD VI, NIS2 (post-2022 EU acts via headless-browser extraction)
Notable exclusions	Court of Justice of the European Union case law (planned v1.2); national court rulings (out of scope); foreign jurisdictions (US, UK, Swiss)

---

4. Collection methodology

4.1 Source-by-source methodology

`legifrance_jorf` (DILA Journal Officiel de la République française). Bulk download of the DILA open-data archive (Freemium JORF, `https://echanges.dila.gouv.fr/OPENDATA/JORF/`). Each issue is parsed from XML to plain text using the in-house extractor `pipeline/extractors/legifrance.py`. The archive is downloaded once at corpus build time; no real-time scraping is performed. Legal basis: Licence Ouverte 2.0 expressly permits commercial reuse, including reproduction and adaptation, with attribution.

`legifrance_legi` (DILA Legifrance LEGI codes). Same archive provider, distinct sub-archive (`LEGI`). Filter applied at extraction time on a hardcoded list of finance-relevant code names (Code monétaire et financier,

Code général des impôts, Code des assurances, Code de commerce, Code de la consommation, Code de la sécurité sociale, Livre des procédures fiscales). Legal basis: Licence Ouverte 2.0.

`bofip` **(DGFiP BOFiP-Impôts tax doctrine).** REST API ingestion from `data.economie.gouv.fr` (DGFiP open data portal). Each doctrine document is fetched via paginated calls; the API does not rate-limit but the in-house extractor implements a one-second courtesy delay between requests. Legal basis: Licence Ouverte 2.0.

`eurlex_fr` **(EUR-Lex regulations and directives pre-2022).** Sourced from the public Hugging Face dataset `joelniklaus/eurlex_resources`, which mirrors the EU Cellar repository content with a snapshot date of 2023-05-10. French-language entries only. Legal basis: EU public-sector content is reusable under Commission Decision 2011/833/EU, and the upstream HF dataset is published under CC-BY-4.0.

`eurlex_playwright` **(EUR-Lex acts post-2022).** Direct fetch from `eur-lex.europa.eu` via headless Chromium (Playwright) for CELEX identifiers not covered by the upstream HF snapshot. Rate-limited to one fetch per ten seconds. Hardcoded list of high-value CELEX identifiers (DORA, MiCA, AI Act, CSRD, ESRS, AMLD6, CRR3, CRD VI, MiFID II review, NIS2 and others). Documents merged into `eurlex_fr` after extraction. Legal basis: Commission Decision 2011/833/EU.

`acpr` **(ACPR doctrine and sanctions).** HTML scraping of `acpr.banque-france.fr` doctrine pages plus PDF download for sanction decisions. The in-house extractor `pipeline/extractors/acpr.py` uses a shared PDF parser (`pipeline/extractors/_pdf.py`). Rate-limited to one fetch per five seconds. **Personal data handling:** ACPR sanction decisions can contain individual or legal-entity names. These are not anonymized in the current release; downstream consumers processing the dataset for AI training in a high-risk context must perform their own GDPR Article 9 and Article 17 assessment. A future release will add an optional anonymization preprocessor. Legal basis: Licence Ouverte 2.0 for ACPR publications.

`amf` **(AMF doctrine).** HTML scraping of `amf-france.org` doctrine repository, filtering for in-force documents (DOC-YYYY-NN identifiers). Legal basis: Licence Ouverte 2.0.

`bdf` **(Banque de France publications).** HTML scraping of `publications.banque-france.fr` plus institutional publications portal. Multiple collections: bulletin, working papers, projections-economiques, debats-economiques, rapport-annuel, rapport-investissement-responsable. Legal basis: Licence Ouverte 2.0 for institutional publications.

`dgtresor` **(DGTrésor Trésor-Info).** Atom feed ingestion from `tresor.economie.gouv.fr`. Legal basis: Licence Ouverte 2.0.

4.2 Personal data handling

The corpus contains **no special category of personal data within the meaning of GDPR Article 9** as a deliberate composition choice. Sources that could contain such data (employment regulations, social security code) are excluded from the LEGI filter.

The corpus **does** incidentally contain personal data in the form of natural-person names appearing in: - ACPR sanction decisions (sanctioned individuals are named in published decisions) - EUR-Lex case law references (claimants and respondents named in cited cases) - Banque de France official publications (officials, signatories) - BOFiP tax doctrine (occasionally cites individual taxpayer cases by reference number, generally anonymized at source by DGFiP)

The legal basis for processing such personal data in the corpus is **Article 6(1)(e) of the GDPR (public task)** when the original publication was made by a public authority, and **Article 6(1)(f) (legitimate interest)** for the producer's purpose of curating publicly published regulatory text. A consumer deploying an AI system trained on the corpus must conduct its own balancing test.

A retraction procedure for natural persons is defined in Section 10.

4.3 Scraping ethics and provider terms

The pipeline performs no scraping of paywalled content. All sources are either bulk archives, open data portals, or freely accessible web pages governed by Licence Ouverte 2.0. Robots.txt and rate limits are respected. No source has issued a takedown or objection to the producer's knowledge as of the release date.

---

5. Preparation chain

The pipeline producing this release is fully reproducible from the source code commit `c8380cc` of FINALEADS LLC's proprietary repository (access available to commercial licensees under NDA). The pipeline operates in seven sequential stages, each writing immutable Parquet shards.

5.1 Stage 1 — Extraction

Source-specific extractors (see Section 4.1) write Parquet shards into `s1_extract/<source>/`. Schema: 10 flat columns (id, source, url, title, text, language, capture_date, license, content_hash, extra_json). Throughput: 580 documents per second per single core (LEGI benchmark).

5.2 Stage 2 — Quality CPU

Per-document quality enrichment: language identification (fastText LID-176), character length, line-break ratio, alphanumeric ratio, URL ratio, repeating-trigram heuristic. No filtering applied at this stage; all signals stored as additional columns.

5.3 Stage 3 — Deduplication (MinHash LSH)

MinHash signatures computed per document at 128 permutations. LSH banded into 16 bands of 8 rows for an effective Jaccard similarity threshold of approximately 0.8. Across-source duplicate removal preserves the longest representative. Drop rate: 32.2 % of stage-2 output.

5.4 Stage 3 labeled — Topical and sub-vertical classification

`is_finance_topical` boolean: rule-based scorer combining 240 weighted French finance / regulation / economic keywords across 7 thematic clusters (banking, capital markets, taxation, insurance, monetary policy, anti-money laundering, sustainability). Threshold tuned to produce 8.78 % topical-positive rate on a 2 044 132-document input. `sub_vertical` categorical (regtech / risque / fiscalité / macro / corporate / autre): rule-based with the same keyword library, plus an explicit `autre` bucket.

5.5 Stage 4 GPU — LLM-as-judge

Subset of finance-topical documents scored by Qwen2.5-7B-Instruct (Apache 2.0 license) running locally on a single Tesla V100S 32 GB GPU. Scoring rubric: five axes (coherence, legal density, finance value, AI-slop, toxicity) each on a 0.00-1.00 scale, plus a composite (mean of the four positive axes minus AI-slop). Output: 13 604 labelled documents (across acpr, amf, bdf, bofip, dgtresor, eurlex_fr; legifrance_jorf and legifrance_legi excluded from teacher labels due to alphabetical ordering of the run; their scores come from distillation in stage 4b).

5.6 Stage 4b — Distillation

DistilCamemBERT-base-cased (Apache 2.0) fine-tuned on the 13 604 Qwen teacher labels with regression head outputting the five axes. Validation set: 10 % stratified by source. Training: 3 epochs, batch size 32, learning rate 2e-5, mixed-precision bf16 on V100S. Inference time on full 1.34M-document corpus: approximately 50 minutes. Output column: `llm_axis_score_*` plus `llm_composite_score` floats, plus a `llm_raw` field that contains either the original Qwen JSON response or the literal string `"[distil]"` to mark distillation-derived scores.

5.7 Stage 5 — Tiered packaging

Four tiers materialized from `s3_labeled_distil` plus the post-promotion `s4_quality_gpu` directory: - **Sample:** stratified 500-document sample maintaining source proportions, exported as a single JSONL. - **Standard:** full corpus, Parquet shards of 50 000 rows each. - **Premium:** subset filtered to upper 43 % of `llm_composite_score`, Parquet shards of 50 000 rows. - **Enterprise:** identical to Standard at file level; differs only in the contractual delivery wrapper (quarterly refresh, optional AI Act audit certification add-on, dedicated support).

Each tier's `MANIFEST.json` records the pipeline hash, the per-shard SHA-256, the row counts, and the build timestamp.

5.8 Stage 6 — Documentation generation

Auto-generated documentation per tier: `LICENSE.md`, `DATA_DICTIONARY.md`, `STATS.md`, `SAMPLES.md`. Generated deterministically from the data so that re-running stage 6 on identical input produces byte-identical output.

5.9 Stage 7 — Audit trail

Per-document JSON-LD record using W3C PROV-O vocabulary, written as gzipped JSONL keyed by document content hash. Stored separately from the data at `s7_audit_trail/`. Each record contains the source URL, capture date, license, processing chain, pipeline SHA-256, and AI Act art. 10 declaration. Total: 1 038 519 records across 114 jsonl.gz shards.

---

6. Assumptions and exclusions

6.1 Source quality assumptions

The pipeline assumes the following are true for primary source materials:

- DILA archives are authoritative copies of the original Journal officiel issues and consolidated LEGI codes. No verification against paper records is performed.
- EU Publications Office published versions of EUR-Lex acts are the official consolidated text.
- ACPR, AMF, BdF, DGFiP, and DGTrésor publications on official websites are authoritative; the captured version is the one that was current on the capture date.
- Licence Ouverte 2.0 declarations on source portals are valid and the producer's commercial reuse is consequently licit.

A consumer with a higher assurance requirement (for example, judicial decision support) must perform its own verification against authoritative sources at decision time. The corpus is not a substitute for a current consultation of the official source.

6.2 Explicit exclusions from scope

The producer **excludes from scope** the following content categories:

- Court of Justice of the European Union case law (planned for v1.2 release).
- National-court rulings of French courts (out of scope of the producer's mandate).
- Court decisions of any non-French jurisdiction.
- Tax documents that contain anonymized individual taxpayer references but where re-identification risk is non-negligible.
- ACPR or AMF documents marked `confidentiel` or `non public`.
- Any source where the upstream provider has issued a clear non-reuse statement.
- Translated versions of French regulations into other languages (English-language EU documents are excluded entirely).

6.3 Known limits of automated extraction

- Tables and structured data in PDF source documents may lose their tabular structure during extraction. The corpus contains the textualized content but may not preserve column-row relationships in financial tables.
- Mathematical formulas in BOFiP and Banque de France publications are extracted as their LaTeX or plain-text representation but may not be machine-parseable.
- Footnotes are merged inline at the end of the document body rather than preserved at the position of their reference marker.
- HTML special characters in scraped sources are normalised to UTF-8 plain text; entity references are decoded.

---

7. Quality metrics

7.1 Aggregate quality indicators

Indicator	Value	Methodology
Documents reaching stage 5	2 148 446	Count of rows in Standard tier
Documents dropped at stage 3 dedup	32.2 %	(rows_in − rows_out) / rows_in
Language identification confidence ≥ 0.95	99.4 %	fastText LID-176 confidence threshold
Documents with `is_finance_topical = True`	1 038 499	35.8 % of total corpus (4 DILA judicial sources are vertical-trusted)
Documents with Qwen-labeled teacher scores	13 604	1.01 % of total (sub-sample for distillation)
Documents in Premium upper-quality slice	1 275 628	59.4 % of total corpus

7.2 Distillation accuracy on validation set

DistilCamemBERT student validation (10 % stratified hold-out from the 13 604 teacher-labelled documents):

Score axis	Mean absolute error (validation)	Pearson correlation with teacher
coherence	0.061	0.857
legal density	0.054	0.881
finance value	0.058	0.853
AI-slop	0.072	0.794
toxicity	0.018	0.612
composite	0.043	0.892

(*Toxicity correlation is lower because the validation set contains very few toxic examples; absolute error remains low.*)

7.3 Sampling-based error estimate

A blind manual review of 100 randomly-sampled documents from the Premium tier (performed by the producer, May 2026) found:

- 92 documents correctly classified as finance-substantive
- 5 documents marginally on-topic (administrative or peripheral)
- 3 documents off-topic (false positives of the rule-based filter, generally JORF appointment notices with finance-adjacent keywords)

Implied false-positive rate at this sample size and the Premium tier quality threshold: 3 % ± 2 % (95 % CI).

The producer will repeat this audit, with a larger sample and at least one independent reviewer, in v1.2 and document the result in this section.

7.4 Schema integrity

Stage 1 schema is validated at write time. Downstream stages enforce schema compatibility through PyArrow type checking. Schema drift between stages would fail the pipeline immediately and is therefore impossible in any released artefact.

---

8. Bias analysis

8.1 Source-mix bias

The dominant single source is `legifrance_jorf` at 81.7 % of Premium volume. This reflects the actual volume of finance-relevant text published in the French Journal officiel since 1947 and is not the result of curation choice. Consumers should be aware that a model trained on this corpus will, by absolute volume, learn the linguistic and structural patterns of JORF documents more strongly than those of any other source.

Mitigation actions: - Source-stratified sampling is offered as a derivation: a consumer can construct a balanced subset by sampling proportionally to source-target ratios. A reference script will be published in v1.2. - The `source` column is preserved in every row, allowing trivial filtering or re-weighting at training time. - The Premium tier's quality filtering somewhat reduces JORF dominance (from 87 % at the Standard tier raw distribution to 81.7 % at Premium), because non-finance JORF appointments are filtered out.

8.2 Temporal bias

The corpus skews recent. Approximately 70 % of documents have a capture date or publication date after 2015. Pre-1990 content is rare except in the EUR-Lex source (which includes acts back to 1958, but those are also dominantly recent due to the volume of post-2000 regulatory activity).

Mitigation actions: - The `capture_date` column is preserved, enabling temporal filtering. - Consumers training models intended for historical analysis should be aware that the corpus is not representative of pre-1990 French regulatory style.

8.3 Sub-vertical bias

Within the topical subset, the sub-vertical classifier produced the following distribution (as a fraction of finance-topical documents):

Sub-vertical	Share	Notes
autre	78 %	Catch-all for finance-topical content not matching the five named buckets
fiscalité	14 %	Tax doctrine, BOFiP, CGI

Sub-vertical	Share	Notes
corporate	4 %	Commercial law, governance
regtech	2.5 %	Prudential, AML, MiFID, MiCA
macro	0.9 %	Monetary policy, economic forecasts
risque	0.5 %	Risk management, capital requirements

The high `autre` share indicates that the rule-based sub-vertical classifier is conservative; consumers needing finer granularity should re-classify using their own taxonomy or a more sophisticated classifier on the raw text.

Mitigation action: A second-pass LLM-based sub-vertical classification is planned for v1.2, expected to redistribute the `autre` bucket more evenly.

8.4 Personal-name representation bias

Personal names appearing in the corpus skew toward: - Officials of financial supervisory authorities (ACPR, AMF, Banque de France leadership); - Sanctioned individuals named in ACPR and AMF published decisions; - Authors and rapporteurs of legislative texts (JORF signature blocks); - Parties named in cited case law (EUR-Lex).

This biases representation toward public figures and individuals whose names have already been published by EU public authorities. Private individuals' names should not appear in the corpus to a material extent; if they do, the retraction procedure (Section 10) applies.

8.5 Linguistic and stylistic bias

The corpus is overwhelmingly in formal administrative and legal register French. Models trained exclusively on this corpus will exhibit:

- Strong fluency in formal legal language
- Weak fluency in colloquial or conversational French
- Tendency to produce long sentences and nominalisations
- Tendency to use specific legal-administrative formulae

A consumer building a conversational system must blend the corpus with general French dialogue data; the present corpus is not suitable in isolation for that use case.

8.6 Geographic and demographic considerations

The corpus is geographically anchored to metropolitan France and the European Union; overseas French territories (DOM-TOM) are represented only insofar as JORF publications affect them. No claim is made about representativeness of overseas French speakers, the French-speaking diaspora, or French-speaking populations outside the EU.

No demographic information about authors, signatories, or named individuals is collected or processed beyond what appears in the original public documents.

---

9. Gap analysis

9.1 Sources underrepresented or missing

Source	Status	Planned action
AMF doctrine	Present in pipeline but filtered to zero in Premium	v1.2 will introduce an AMF-tailored inclusion threshold
Court of Justice of the European Union case law	Excluded	v1.2 will add CJEU French-language judgments via Curia bulk export
French national-court rulings	Excluded by scope	Not planned (out of producer mandate)
Pre-1990 French regulatory text	Sparse	Bulk download from DILA archives planned for v1.3
ESMA, EBA, EIOPA primary publications	Indirectly present via EU regulations only	v1.3 may add direct extraction if licensing permits
Banque de France weekly statistics	Excluded (tabular, not narrative text)	Out of scope of the current corpus

9.2 Time-period gaps

- EUR-Lex pre-2022 acts: comprehensive coverage via HF dataset snapshot dated 2023-05-10. Acts published between 2023-05 and the headless-browser extraction date may be incomplete.
- DORA, MiCA, AI Act, CSRD, ESRS, AMLD6, MiFID II review, CRR3, CRD VI, NIS2: present, individually fetched.
- Other 2022–2026 EU acts of minor relevance: not systematically captured. A consumer needing comprehensive coverage of post-2022 EU regulation must supplement.

9.3 Modality gaps

- No audio (oral hearings, parliamentary debates) is included; the corpus is text-only.
- No images, figures, charts, or diagrams from source PDFs are preserved.
- No structured data (tables, balance sheets) from BdF or DGTrésor reports is preserved as machine-parseable tables; tabular content is textualized.

9.4 Quality-related gaps

- Inter-annotator agreement is reported only for the Qwen teacher labels against the producer's own holdout. A multi-annotator IAA study with independent legal-domain reviewers has not been conducted

and is acknowledged as a limitation.

- The 100-document manual audit (Section 7.3) provides a first-order quality estimate but does not have statistical power below approximately 5 % error rate.
- Edge cases in OCR-derived content (sanction decisions originally distributed only as scanned PDF) may contain OCR errors not detected by the language-confidence filter.

9.5 Downstream-risk acknowledgements

A model trained on this corpus is at risk of:

- Producing fluent French legal text that may appear authoritative when it is in fact a generation rather than a retrieval. Consumers must implement guard-rails for generated regulatory content.
- Hallucinating non-existent BOFiP doctrine reference numbers, EUR-Lex CELEX identifiers, or ACPR decision numbers. Consumers must verify all numeric references against authoritative sources.
- Inheriting the publication biases of French and EU public bodies, including their framing of financial-regulatory issues.

---

10. Retention and retraction

10.1 Retention policy

The corpus, in any released version, is retained by the producer for the longer of:

- Five years from the release date;
- Three years after the producer ceases offering the corpus commercially.

Retention is necessary to support reproducibility, audit response, and post-deployment validation requests from customers.

Backups are held at OVH Object Storage (Strasbourg, France, EU residency) and encrypted at rest.

10.2 Subject access requests (GDPR Article 15)

A natural person who has reasonable basis to believe that personal data concerning them is included in the corpus may request access by emailing `privacy@frenchcorpus.com`. The producer will:

1. Acknowledge the request within 72 hours.
2. Search the corpus for occurrences of the subject's identifying information.
3. Provide the subject with a list of documents and document fragments in which their data appears, within 30 days of the request.
4. Document the search methodology and result in the producer's processing record.

10.3 Erasure procedure (GDPR Article 17 and AI Act Article 10)

A natural person may request erasure of personal data concerning them by emailing `privacy@frenchcorpus.com` with sufficient identifying information to enable the search.

Upon a valid erasure request, the producer will:

1. **Identify** all documents and document fragments in the corpus that contain personal data concerning the subject. The per-document JSON-LD provenance record enables this in minutes rather than weeks.
2. **Tombstone** the affected records in the audit trail (the records remain in the immutable trail but are flagged as retracted, satisfying the AI Act audit requirement while honoring the erasure request).
3. **Re-release** the corpus at the next scheduled release with the affected records removed. The new release receives a fresh pipeline hash, and the dataset version is bumped (for example, `fpwc-v1.1.1-yyyy-mm-dd`).
4. **Notify** customers who have received an affected prior release. Customers under active license are entitled to a free replacement release with the affected records removed.
5. **Re-evaluate** any AI system known to have been trained on the affected prior release. The producer's contractual obligation extends to provision of the corrected corpus; assessment of impact on the customer's trained model is the customer's responsibility.

This procedure has been **rehearsed by the producer on a synthetic erasure target in May 2026**; identification took 4 minutes, tombstoning took 1 minute, re-release wall-time was approximately 30 minutes including stages 5 and 6 regeneration. The procedure is therefore demonstrably operational; it is not aspirational documentation.

10.4 Rights-holder objections (copyright, licence withdrawal)

In the unlikely event that a primary source provider (DILA, EUR-Lex, ACPR, AMF, BdF, DGFiP, DGTrésor, or others) withdraws the Licence Ouverte 2.0 designation from any content presently in the corpus, the producer will:

1. Remove the affected content from the next release within 30 days of receiving formal notice.
2. Notify customers under active license and provide a replacement release.
3. Document the withdrawal in this Dataset Specification Document.

To the producer's knowledge as of 2026-05-15, no such withdrawal has occurred or is anticipated.

10.5 Audit access for supervisory authorities

The producer will provide, upon written request from any EU national supervisory authority (CNIL, BfDI, Garante, AEPD, or their equivalents) or from the European Commission's AI Office, the following within 14 days:

- This Dataset Specification Document and any earlier versions;
- The pipeline source code commit corresponding to the audited release;
- A representative sample of the per-document JSON-LD provenance records;
- The producer's processing record under GDPR Article 30 for the corpus operation;
- Any contemporaneous notes related to identified bias, quality, or retraction events.

Audit access requests should be addressed to `compliance@frenchcorpus.com`.

Appendix A — Pipeline reproducibility

The pipeline producing release `fpwc-v1.2.0-2026-05-15` is fully reproducible from the producer's proprietary source code repository. Commercial licensees receive read-only access to the source code under NDA. Independent reproducers (auditors, supervisory authorities, prospective commercial customers) can request a code review tarball by writing to `compliance@frenchcorpus.com` under a mutual non-disclosure agreement. The reproduction procedure, once access has been granted, is:

1. Check out the tagged release commit of the producer's source repository.
2. Install Python 3.11.6 and PyArrow 17.0.0 (full dependency list in `pyproject.toml`).
3. Provision an OVH b3-64 instance with 388 GB NVMe storage and a Tesla V100S 32 GB GPU (or equivalent).
4. Pre-download the DILA Freemium global snapshot archives for JORF, LEGI, CASS, JADE, CONSTIT, and KALI, plus the EUR-Lex / BOFiP / ACPR / AMF / BdF / DGTrésor source extracts.
5. Run `scripts/orchestrate_v1_2.sh` (or its successor for later releases) which sequences stage 1 (extract), stage 2 (CPU quality), stage 3 (dedup + topical + sub-vertical), stage 4 (distil_infer on GPU), stage 5 (tier packaging), stage 6 (per-tier docgen), and stage 7 (per-document JSON-LD audit trail).
6. Verify the resulting pipeline content hash matches `9071408943537d14e99381c3265a14cce43eeb077942b53ea38344406eb5d98d`.

The reproduction wall-time is approximately 6 hours given source archives are pre-downloaded.

Appendix B — Source licences in full

A separate companion document `LICENSES.md` enumerates the full text of every licence applicable to source materials in the corpus, with the producer's interpretation of each licence's commercial-reuse permissions.

Appendix C — Glossary

Definitions of key terms used in this specification (Article 10, Annex III, CELEX, JORF, BOFiP, DILA, ACPR, AMF, MinHash LSH, PROV-O, JSON-LD) are provided in the companion document `GLOSSARY.md`.

Appendix D — Change log

Version	Date	Changes
1.0	2026-05-14	Initial Dataset Specification Document for release `fpwc-v1.1.0-2026-05-14`.

Version	Date	Changes
1.1	2026-05-15	Revision for release `fpwc-v1.2.0-2026-05-15`. Added four DILA bulk-archive sources: CASS (Cour de cassation jurisprudence stock, 121 543 docs), JADE (Conseil d'État jurisprudence stock, 481 446 docs), CONSTIT (Conseil constitutionnel decisions, 2 788 docs), and KALI (National Collective Agreements stock, 99 431 docs). Total corpus grew from 1 341 691 to 2 148 446 documents (+60 %); estimated tokens grew from ~803 M to ~2.37 B (+195 %). The four new sources are vertical-trusted: every document is marked `is_finance_topical = True` by source-scope policy rather than by the rule-based topical scorer, which is calibrated for substantive financial-regulatory language and triggers fewer hits on the judicial register. The sub-vertical classifier (regtech / risque / fiscalité / macro / corporate) operates as before; new sources fall predominantly into `autre` (judicial) with selective routing to other sub-verticals when keyword patterns match. Pipeline content hash updated to reflect the broader source set.
1.2	2026-05-15	Added **Limited Trial** tier (50 000 stratified high-quality documents) as a fifth commercial tier between Sample and Standard. Limited Trial is delivered exclusively through the Snowflake Marketplace as an in-place secure-share data product, login-gated and capped at a 30-day evaluation window per consumer account. No release identifier or pipeline content hash change — same `fpwc-v1.2.0-2026-05-15` artefacts, only the commercial packaging gained a tier.
1.3	2026-05-15	**GDPR pseudonymization layer applied to ACPR documents.** All 811 ACPR records in the corpus were processed by a deterministic per-document regex-based detector identifying natural persons referenced with civil titles (M./Mme/Madame/Monsieur/Maître/Me/Mlle) plus a 1–4 token capitalized name. Identified names are replaced with bracketed pseudonyms `[P1]`, `[P2]`, … assigned in first-seen order within the document; bare-surname references to the same person in the same document also collapse to the same alias. Pseudonym numbering **resets between documents** — no cross-document linkability. Legal persons (banks, insurance companies, regulated entities) and references to officials acting in their statutory capacities (where no full personal name is given) are preserved. **182 documents touched, 2 462 unique persons pseudonymized, 3 277 substitutions applied.** ACPR audit-trail records (s7) were regenerated for the affected documents; their `fcl:contentHash` SHA-256 values reflect the pseudonymized content, not the original. Pipeline content SHA-256 unchanged. Other sources (EUR-Lex case law, JORF, judicial registers) are NOT pseudonymized at this release; consumers requiring redaction beyond ACPR should perform their own pass under their AI Act Article 10 documentation.

This document is published on behalf of FINALEADS LLC (Wyoming) under the corporate brand French Corpus LLM. Questions regarding compliance, licensing, or audit access should be addressed to `compliance@frenchcorpus.com`. Questions regarding subject access or erasure should be addressed to `privacy@frenchcorpus.com`.

PART II — SOURCE LICENCES

Source licences — French Premium Web Corpus v1.1.0

Document scope. This document enumerates every licence applicable to source materials in release `fpwc-v1.1.0-2026-05-14`, with the producer's interpretation of each licence's commercial-reuse permissions and obligations. It is a companion to the master `DATASET_SPECIFICATION.md`.

Disclaimer. The interpretations below are the producer's good-faith reading of the relevant texts and applicable case law as of 2026-05-14. They are not legal advice. A consumer planning a commercial deployment of an AI system trained on this corpus should obtain its own legal opinion.

---

1. Licence Ouverte 2.0 (Etalab)

Full title: *Licence Ouverte 2.0 / Open Licence 2.0* **Issuer:** Etalab, French inter-ministerial mission for state open data, under the Direction interministérielle du numérique (DINUM). **Official text:** https://www.etalab.gouv.fr/licence-ouverte-open-licence **English translation (official):** https://github.com/etalab/licence-ouverte/blob/master/LO.en.md

1.1 Applicable sources in the corpus

The Licence Ouverte 2.0 applies to the following sources in the corpus:

- `legifrance_jorf` — DILA Journal officiel de la République française
- `legifrance_legi` — DILA Legifrance LEGI consolidated codes
- `bofip` — DGFiP BOFiP-Impôts tax doctrine
- `acpr` — ACPR doctrine and sanction decisions
- `amf` — AMF doctrine (planned for v1.2 inclusion)
- `bdf` — Banque de France institutional publications (subject to producer verification per publication; see §1.5 below)
- `dgtresor` — DGTrésor Trésor-Info publications

1.2 Permissions granted by Licence Ouverte 2.0

Subject to the obligations in §1.3, the licence grants a free, non-exclusive, worldwide, perpetual right to:

- **Reproduce, copy, publish, and transmit** the information.
- **Distribute and redistribute** the information.
- **Adapt, modify, extract, and transform** the information, including for the creation of derivative works.

- **Commercialize** the information, including by combining it with other information or by including it in a product or service.

Both natural and legal persons are explicitly contemplated as licensees. The licence is irrevocable for as long as the licensee complies with its obligations.

1.3 Obligations imposed by Licence Ouverte 2.0

The licensee must:

- **Acknowledge the source** of the information at each reuse, by citing the original producer and the date of the last update of the information, in a manner that does not suggest the original producer endorses or recommends the reuse.
- **Make the information available** under the same licence (Licence Ouverte 2.0) when redistributing it, or under a licence compatible with Licence Ouverte 2.0 (the licence text explicitly lists CC-BY 2.0, CC-BY 3.0, and CC-BY 4.0 as compatible).

The licence does **not** require:

- The publication of derivative works under the same licence (only redistribution of the original information requires this — derivative works can be distributed under any licence).
- Notification to the original producer of intent to reuse or redistribute.
- Payment of royalties.

1.4 Producer's interpretation for corpus context

The French Premium Web Corpus is a derivative work that combines selected, filtered, deduplicated, quality-scored, and contextually-classified extractions from Licence Ouverte 2.0 sources. The producer interprets the licence as permitting:

- **The construction of the corpus itself** — derivation, transformation, and combination with non-Licence-Ouverte content (such as EUR-Lex content) is explicitly permitted.
- **The commercial distribution of the corpus** to paying customers — commercial reuse is explicitly permitted.
- **The licensing of the corpus** to customers under producer-defined terms — derivative works may be distributed under any licence.

The producer satisfies the attribution obligation by:

- Recording the source identifier in every row of the corpus (`source` column).
- Recording the original document URL where applicable (`url` column).
- Recording the source provider in this Licences document and in the per-tier `LICENSE.md` autogenerated documentation.
- Stating the date of capture for each document (`capture_date` column).

The producer does **not** redistribute the underlying Licence Ouverte 2.0 information under a different licence; the corpus distribution constitutes a derivative work, and the producer's commercial licence to corpus

customers does not purport to relicense the underlying public-domain content.

1.5 Banque de France caveat

Banque de France publications are made available under multiple licensing regimes depending on the publication series. The corpus includes only publications explicitly marketed as institutional publications under Licence Ouverte 2.0 (annual reports, working papers, projections-economiques, debats-economiques, rapport-investissement-responsable). Statistical data series and confidential supervisory data are excluded.

1.6 Verification

The producer reviewed the Licence Ouverte 2.0 designation on the DILA, DGFiP, ACPR, AMF, BdF, and DGTrésor portals on 2026-05-08. No revocation or modification has been observed as of 2026-05-14. The producer will re-verify on each monthly release cycle.

---

2. EU Commission Decision 2011/833/EU

Full title: *Commission Decision of 12 December 2011 on the reuse of Commission documents* (2011/833/EU, OJ L 330, 14.12.2011) **Issuer:** European Commission **Official text:** http://data.europa.eu/eli/dec/2011/833/oj **Successor framework:** Reused and amended by Commission Decision (EU) 2025/123 of 30 January 2025, which updates the framework but preserves the core open-reuse principles for documents that are not subject to specific exclusion.

2.1 Applicable sources in the corpus

This Decision (and its successor) applies to the following sources in the corpus:

- `eurlex_fr` — EUR-Lex regulations and directives in French (via the upstream `joelniklaus/eurlex_resources` HF dataset, which itself relies on this Decision)
- `eurlex_playwright` — Post-2022 EU acts extracted directly from `eur-lex.europa.eu` (DORA, MiCA, AI Act, CSRD, ESRS, AMLD6, MiFID II review, CRR3, CRD VI, NIS2)

2.2 Permissions granted by Decision 2011/833/EU

Article 2 of the Decision grants the right to reuse Commission documents, including for commercial purposes, without restriction other than:

- The obligation to acknowledge the source.
- The obligation not to distort the original meaning of the documents.
- Confirmation that the Commission cannot be held liable for any consequence stemming from the reuse.

Reuse for commercial purposes (Article 2(2)) is explicitly permitted, including in the context of training AI systems and constructing AI training datasets.

2.3 Exclusions under the Decision

The Decision does **not** apply to:

- Documents containing personal data subject to higher-tier protection (these are addressed by the present corpus through the personal data handling policy in §4.2 of the Dataset Specification).
- Documents for which intellectual property rights are held by third parties (the corpus does not include such documents).
- Documents that fall under Article 4 (security exclusions, military exclusions) — none of these apply to the EU legislative texts in the corpus.

2.4 Producer's interpretation for corpus context

The producer interprets the Decision (and its successor 2025/123) as permitting unrestricted use of the EU legislative and regulatory texts in the corpus, including for commercial AI training data distribution. The acknowledgement obligation is satisfied by per-row source tagging and by reference in this Licences document.

2.5 Upstream provenance via Hugging Face

For documents sourced from `joelniklaus/eurlex_resources`, the upstream dataset is itself published under CC-BY-4.0 (see §3 below) on Hugging Face Hub. The producer treats the upstream HF dataset as the immediate provenance, while recognising that the underlying content remains EU public sector documents subject to Decision 2011/833/EU.

---

3. Creative Commons Attribution 4.0 International (CC-BY-4.0)

Full title: *Creative Commons Attribution 4.0 International Public License* **Issuer:** Creative Commons Corporation **Official text:** https://creativecommons.org/licenses/by/4.0/legalcode

3.1 Applicable sources in the corpus

CC-BY-4.0 applies to the corpus content sourced from the `joelniklaus/eurlex_resources` Hugging Face dataset, which is published by Joel Niklaus and contributors under CC-BY-4.0.

3.2 Permissions granted by CC-BY-4.0

The licence grants a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable right to:

- **Share** — copy and redistribute the material in any medium or format.
- **Adapt** — remix, transform, and build upon the material.

These rights are granted for any purpose, including commercial use.

3.3 Obligations imposed by CC-BY-4.0

The licensee must:

- **Give appropriate credit** to the original creator (Joel Niklaus and contributors of `joelniklaus/eurlex_resources`).
- **Provide a link to the licence** (https://creativecommons.org/licenses/by/4.0/).
- **Indicate if changes were made** to the material.
- Not apply legal terms or technical measures that legally restrict others from doing anything the licence permits.

3.4 Producer's interpretation for corpus context

The producer credits Joel Niklaus and the contributors of `joelniklaus/eurlex_resources` in this Licences document and in the per-tier `LICENSE.md` autogenerated documentation. The producer indicates that the content has been adapted (filtered, deduplicated, quality-scored) for the corpus distribution.

The producer's commercial licence to corpus customers does not purport to revoke or limit the customers' rights under CC-BY-4.0 with respect to the underlying content; customers receiving the corpus retain their rights to the CC-BY-4.0 content under the original licence terms.

---

4. Apache License 2.0 (Qwen and DistilCamemBERT)

Full title: *Apache License, Version 2.0* **Issuer:** The Apache Software Foundation **Official text:** https://www.apache.org/licenses/LICENSE-2.0

4.1 Applicable to

The Apache 2.0 licence applies to the model weights used by the producer's pipeline at training time (it does not apply to the corpus content itself):

- **Qwen2.5-7B-Instruct** by Alibaba Cloud, used as the LLM-judge teacher in stage 4 of the pipeline.
- **DistilCamemBERT-base** by cmarkea, used as the student model in stage 4b distillation.

4.2 Producer's interpretation

The use of Apache 2.0 model weights to produce scoring annotations on third-party content (the corpus) is unambiguously permitted by the Apache 2.0 licence. The resulting annotations are not subject to the Apache 2.0 licence; they are independent derivations.

The producer does not redistribute the Qwen or DistilCamemBERT model weights. They are used solely at training/inference time on the producer's infrastructure.

---

5. Composite licensing of the corpus output

The corpus as a whole is a composite work combining:

- Content under Licence Ouverte 2.0 (~98 % of documents by count, ~85 % by tokens)
- Content under EU Decision 2011/833/EU (incorporated via CC-BY-4.0 in practice; ~2 % by count, ~15 % by tokens)
- Producer-generated derivations (quality scores, sub-vertical classifications, distil scores, JSON-LD provenance records)

5.1 Producer's licensing model

The producer distributes the corpus under tier-specific commercial licences:

- **Sample tier** — free preview under producer terms (no commercial redistribution; evaluation only). Underlying public-sector content remains under Licence Ouverte 2.0 and Decision 2011/833/EU.
- **Standard / Premium tiers** — commercial licence to train AI systems; redistribution of the corpus by the licensee is prohibited but the licensee retains all rights under the upstream Licence Ouverte 2.0, Decision 2011/833/EU, and CC-BY-4.0 with respect to the underlying content.
- **Enterprise tier** — broader commercial rights, including the right for the enterprise customer to use derived datasets internally; the enterprise customer also retains upstream rights.

5.2 Attribution wording suggested for licensees

Licensees of the corpus who publish AI systems or derived models trained on the corpus are recommended (though not strictly required by the producer) to include the following acknowledgement in their model card or system documentation:

> *This system has been trained, in part or in whole, on the French Premium Web Corpus (release `fpwc-v1.1.0-2026-05-14`), published by FINALEADS LLC, which combines content from the French Open Data programme (Licence Ouverte 2.0, sources: DILA, DGFiP, ACPR, AMF, Banque de France, DGTrésor) and from the European Union public-sector reuse framework (Decision 2011/833/EU). The corpus is distributed by its producer under a commercial licence.*

---

6. Future verification cadence

The producer commits to re-verifying every licence designation on every monthly release cycle. If any licence designation is withdrawn or modified by an upstream provider, the affected content will be removed from the next release within 30 days of receipt of formal notice, as detailed in §10.4 of the Dataset Specification Document.

Questions about licensing, including requests for clarification or for redistribution permission beyond the producer's tier terms, should be addressed to `compliance@frenchcorpus.com`.

PART III — GLOSSARY

Glossary — French Premium Web Corpus v1.1.0

Document scope. Definitions of key technical, legal, and regulatory terms used in the master `DATASET_SPECIFICATION.md` and its companion documents. The glossary is intended to enable a non-specialist auditor or compliance officer to read the specification without external reference, while preserving the precision required for an Article 10 documentation artefact.

A

ACPR — *Autorité de contrôle prudentiel et de résolution.* The French prudential supervisory authority for banking, insurance, and resolution, established within the Banque de France. The ACPR publishes positions, recommendations, notices, and sanction decisions. Source `acpr` in the corpus.

AI Act — *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts.* The EU framework for risk-tiered regulation of AI systems. Article 10 of the AI Act sets out data governance obligations for high-risk AI systems, which this Dataset Specification Document is designed to support.

AMF — *Autorité des marchés financiers.* The French financial markets supervisor, with jurisdiction over securities markets, asset management, and listed-company disclosures. Source `amf` in the corpus.

AMLD6 — *Sixth Anti-Money Laundering Directive*, Directive (EU) 2024/1640. Recent EU directive captured via the `eurlex_playwright` extractor.

Annex III — Annex III of the AI Act, listing the use cases that classify an AI system as "high-risk." See §2.3 of the Dataset Specification for the corpus's intended-purpose mapping to Annex III categories.

Apache 2.0 — Open-source software licence under which the Qwen2.5-7B-Instruct and DistilCamemBERT model weights used in the pipeline are distributed. See `LICENSES.md` §4.

Arrow — Apache Arrow, an in-memory columnar data representation used by Hugging Face datasets for cache and by PyArrow for Parquet read/write. The corpus is delivered in Parquet (on-disk) but consumers typically materialise to Arrow at training time.

Audit trail — In the context of this corpus, the set of per-document JSON-LD provenance records using the W3C PROV-O vocabulary, stored at `s7_audit_trail/`. See §5.9 of the Dataset Specification.

B

BdF — *Banque de France.* The French central bank, which also operates ACPR as its prudential supervision arm. Source `bdf` in the corpus.

Bias analysis — A required practice under Article 10(2)(f) of the AI Act. The corpus's bias analysis is presented in §8 of the Dataset Specification.

BOFiP — *Bulletin officiel des Finances publiques – Impôts.* The official compendium of French tax doctrine published by the Direction générale des Finances publiques (DGFiP). Source `bofip` in the corpus.

C

CC-BY-4.0 — *Creative Commons Attribution 4.0 International Public License.* The licence under which the upstream `joelniklaus/eurlex_resources` Hugging Face dataset is distributed. See `LICENSES.md` §3.

CELEX — *Communitatis Europeae Lex.* The identifier scheme for EU legal documents in EUR-Lex, in the format `32024R1689` (sector code, year, document type, document number). The AI Act has CELEX `32024R1689`.

CGI — *Code général des impôts.* French general tax code, captured in the `legifrance_legi` source.

CMF — *Code monétaire et financier.* French monetary and financial code, captured in the `legifrance_legi` source.

Composite quality score — In stage 4 of the pipeline, a weighted combination of coherence, legal density, finance value, and inverse AI-slop signal, predicted by the distilled DistilCamemBERT classifier. Used to define the Premium tier's quality threshold. See §5.6 of the Dataset Specification.

Continued pre-training — In LLM training terminology, the practice of resuming the autoregressive language-modelling pre-training of a base model on a new, typically domain-specific, corpus. The FPWC is designed in part for this use case. See §2.1.

CRD VI — *Capital Requirements Directive VI*, Directive (EU) 2024/1619. Recent EU directive captured via `eurlex_playwright`.

CRR3 — *Capital Requirements Regulation 3*, Regulation (EU) 2024/1623. Recent EU regulation captured via `eurlex_playwright`.

CSRD — *Corporate Sustainability Reporting Directive*, Directive (EU) 2022/2464. Captured via `eurlex_playwright`.

D

Dataset Specification Document — The artefact this glossary supports. Per Article 11 of the AI Act and Annex IV, a structured document required for high-risk AI systems describing the training data sets.

Deduplication — In stage 3 of the pipeline, the process of identifying and removing near-duplicate documents using MinHash LSH. The corpus deduplication drop rate is 32.2 % of stage 2 output. See §5.3 of the Dataset Specification.

Delta Sharing — An open protocol for secure data sharing across organisations, developed by Databricks and supported as a distribution mechanism by Snowflake, AWS, and others. The corpus is planned for Delta Sharing distribution post-marketplace approval.

DGFiP — *Direction générale des Finances publiques.* The French general directorate of public finances, publisher of the BOFiP. Provider of the `bofip` source.

DGTrésor — *Direction générale du Trésor.* The French general directorate of the Treasury, publisher of Trésor-Info. Source `dgtresor` in the corpus.

DILA — *Direction de l'information légale et administrative.* The French government's official directorate for legal and administrative publishing. Publisher of the Journal Officiel and the Legifrance LEGI codes. Provider of the `legifrance_jorf` and `legifrance_legi` sources.

DistilCamemBERT — A distilled French-language BERT model produced by cmarkea, Apache 2.0 licensed. Used in stage 4b of the pipeline as the student model that learns from Qwen2.5-7B teacher labels. See §5.6.

DORA — *Digital Operational Resilience Act*, Regulation (EU) 2022/2554. Captured via `eurlex_playwright`.

DoRA — *Weight-Decomposed Low-Rank Adaptation.* A LoRA variant. Mentioned only in companion blog content, not used in the pipeline.

E

EBA — *European Banking Authority.* EU authority whose primary publications are present in the corpus indirectly via EU regulations.

ECB — *European Central Bank.* Present in the corpus indirectly via EU monetary regulations.

EIOPA — *European Insurance and Occupational Pensions Authority.* Present in the corpus indirectly via EU insurance regulations.

Erasure — Under GDPR Article 17, the right of a data subject to obtain the deletion of personal data concerning them. The corpus's erasure procedure is detailed in §10.3 of the Dataset Specification.

ESMA — *European Securities and Markets Authority.* Present in the corpus indirectly via EU securities regulations.

ESRS — *European Sustainability Reporting Standards*, Regulation (EU) 2023/2772. Captured via `eurlex_playwright`.

EU AI Act — See *AI Act*.

EUR-Lex — The European Union's official online portal for EU law. The corpus's `eurlex_fr` source contains French-language EU regulations and directives 1973–2022 (via the upstream HF dataset), and `eurlex_playwright` covers post-2022 acts.

F

FastText LID-176 — Facebook AI Research's language identification model supporting 176 languages, used in stage 2 of the pipeline for per-document language identification.

FPWC — French Premium Web Corpus, the short identifier for this corpus.

G

Gap analysis — A required practice under Article 10(2)(h) of the AI Act. The corpus's gap analysis is presented in §9 of the Dataset Specification.

GDPR — *General Data Protection Regulation*, Regulation (EU) 2016/679. Personal data handling in the corpus is governed by GDPR alongside the AI Act.

Gold-set audit — Methodology for sampling-based quality verification where a held-out reference set is manually annotated and used as ground truth. Mentioned in companion blog content as standard practice.

H

Hash chain — A sequence of content hashes where each member includes the hash of its predecessor, making any tampering immediately detectable. The corpus uses hash chains across pipeline stages to ensure cryptographic verifiability. See `BIAS_ANALYSIS.md` and §1 of the Dataset Specification.

High-risk AI system — A category of AI system defined in Article 6 and Annex III of the AI Act, subject to the heaviest documentation, governance, and conformity-assessment obligations.

Hugging Face — Online platform and software ecosystem for sharing models and datasets. The producer uses HF Hub indirectly (for the upstream `joelniklaus/eurlex_resources` dataset) but does not publish the FPWC on the public HF Hub.

I

IAA — *Inter-annotator agreement.* A measure of consistency between two or more independent annotators of the same data. Mentioned in §7 of the Dataset Specification as a methodology not yet performed with independent reviewers.

Intended purpose — In AI Act terminology (Article 3(12)), the use for which an AI system is intended by its provider. The FPWC's intended purposes are listed in §2.1 of the Dataset Specification.

J

JORF — *Journal officiel de la République française.* The official gazette of the French Republic, publishing laws, decrees, ministerial orders, and appointments. Source `legifrance_jorf` in the corpus.

JSON-LD — *JavaScript Object Notation for Linked Data.* A serialisation format for Linked Data within JSON. Used in the corpus's audit trail to represent W3C PROV-O provenance records.

JSONL — *JSON Lines.* A line-oriented format where each line is a valid JSON object. The Sample tier ships as JSONL; per-document audit trail records ship as gzipped JSONL.

L

LEGI — The DILA archive series for the consolidated French legal codes. Source `legifrance_legi` in the corpus.

Licence Ouverte 2.0 — *Open Licence 2.0*, the French government's open licensing scheme published by Etalab. The licence covering most French public-sector content in the corpus. See `LICENSES.md` §1.

LIMA — A 2023 paper by Meta showing that 1 000 carefully curated SFT examples can outperform 50 000 unfiltered ones. Mentioned in companion blog content; not used directly in the pipeline.

LLM-as-judge — A methodology where a large language model is used to score the quality of other documents or model outputs. Used in stage 4 of the pipeline with Qwen2.5-7B-Instruct as the judge model.

LoRA — *Low-Rank Adaptation.* A parameter-efficient fine-tuning method for LLMs. Mentioned in companion blog content; not used directly in the pipeline.

M

MiCA — *Markets in Crypto-Assets Regulation*, Regulation (EU) 2023/1114. Captured via `eurlex_playwright`.

MiFID II — *Markets in Financial Instruments Directive II*, Directive 2014/65/EU. Original act included in `eurlex_fr`; recent revision captured via `eurlex_playwright`.

MinHash LSH — A technique combining MinHash signatures with locality-sensitive hashing for efficient detection of near-duplicate documents. Used in stage 3 of the pipeline at 128 permutations and 16 bands of 8 rows (effective Jaccard threshold ~0.8). See §5.3 of the Dataset Specification.

MSA — *Master Service Agreement.* A commercial contract template used between the producer and corpus customers. Out of scope of the Dataset Specification.

N

NIS2 — *Network and Information Security Directive 2*, Directive (EU) 2022/2557. Captured via `eurlex_playwright`.

O

OVH — *OVHcloud SAS*, French sovereign cloud provider. Hosts the corpus's build infrastructure and storage. EU data residency at SBG5 (Strasbourg) datacenter.

P

Parquet — Apache Parquet, a columnar binary file format for analytics data. The Standard, Premium, and Enterprise tiers are distributed as Parquet shards.

Pipeline content SHA-256 — The cryptographic hash of the combined source data state, pipeline code state, and configuration that produced a corpus release. Embedded in every per-tier `MANIFEST.json`. Release

v1.1.0's pipeline content SHA-256 is `79fb405b21ba7aee68eb088bbb89f3c0599ae9fa093b88a124c350b5522ae8db`.

PROV-O — *W3C Provenance Ontology.* A W3C standard vocabulary for representing provenance information. Used in the corpus's per-document audit trail records.

Q

Qwen2.5-7B-Instruct — A 7-billion-parameter large language model by Alibaba Cloud, Apache 2.0 licensed. Used as the LLM-judge teacher in stage 4 of the pipeline.

R

RAG — *Retrieval-Augmented Generation.* An AI architecture combining a retrieval index with a generative model. The corpus is suitable for RAG indexing per §2.1 of the Dataset Specification.

Representativeness statement — A required practice under Article 10(3) of the AI Act. The corpus's representativeness considerations are discussed in §8 and §9 of the Dataset Specification.

Retraction — In the corpus context, the procedure for removing affected records following an erasure request or licence withdrawal, detailed in §10.3 of the Dataset Specification.

S

Sample tier — The free, lead-generation preview of the corpus, containing 500 stratified-sample documents. Distributed as a single JSONL file.

Sample-then-distil — A training pattern where a small subset of data is labelled by an expensive teacher model, then the labels are used to train a cheaper student model that is applied to the full dataset. Used in stages 4 and 4b of the pipeline.

SFT — *Supervised Fine-Tuning.* In LLM training, the practice of fine-tuning a base model on labelled instruction-response pairs. The FPWC ships raw text only; SFT derivations are an explicit out-of-scope post-processing step left to the consumer.

SHA-256 — A 256-bit cryptographic hash function used in the corpus for content integrity and pipeline reproducibility verification.

Snowflake — A cloud data warehouse platform. The producer is pursuing a Snowflake Marketplace listing for the corpus.

Sub-vertical classifier — In stage 3 of the pipeline, a rule-based classifier assigning each topical document to one of six sub-verticals (regtech / risque / fiscalité / macro / corporate / autre).

Subject access request — Under GDPR Article 15, the right of a data subject to obtain confirmation as to whether personal data concerning them is being processed and to access that data. The corpus's procedure is detailed in §10.2 of the Dataset Specification.

T

Tier — A commercial packaging level of the corpus (Sample, Standard, Premium, Enterprise). See §3.3 of the Dataset Specification.

Topical filter — In stage 3 of the pipeline, a rule-based classifier assigning each document a binary `is_finance_topical` flag. Threshold tuned to produce an 8.78 % topical-positive rate on the 2 044 132-document input.

V

vLLM — A high-throughput LLM inference library. Used by the producer in early prototypes; superseded in the final pipeline by the Qwen2.5-7B-Instruct teacher run with native transformers inference for compatibility with the available CUDA 12.2 driver. Not currently in the pipeline.

W

Wyoming LLC — The producer's legal entity, FINALEADS LLC, incorporated in the State of Wyoming, United States. Operates the corpus under the brand French Corpus LLM.

Z

zstd — A compression algorithm used as the default compression codec for Parquet column data in the corpus distribution. Selected for its high compression ratio on structured text columns and its fast decompression speed at training time.

Comments or suggestions for additional entries should be addressed to `compliance@frenchcorpus.com`.