

Système de scraping StrideMatch – Version pédagogique

But du document

Expliquer simplement comment fonctionne le système de scraping, quelles données il récupère, comment elles sont stockées, et ce que cela apporte à StrideMatch – sans entrer dans les détails techniques.

1. En une phrase : à quoi sert le scraping pour StrideMatch ?

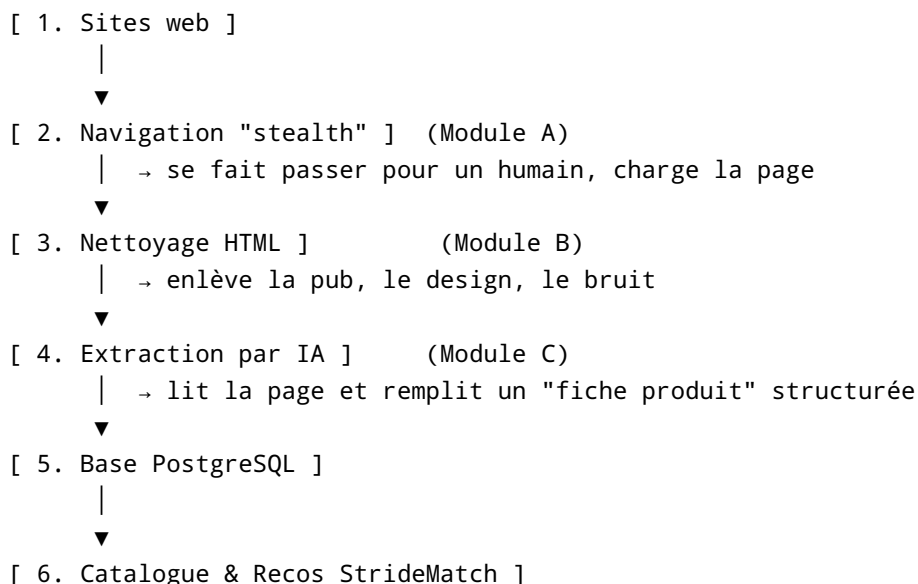
Le scraping est un **robot de lecture** qui parcourt des sites spécialisés (RunRepeat, blogs running, comparateurs de prix...), récupère les informations sur les chaussures de running, les nettoie, les structure, puis les stocke dans une base de données unique pour alimenter les **recommandations StrideMatch**.

Sans scraping : données incomplètes, à jour difficilement, beaucoup de saisie manuelle.

Avec scraping : **catalogue riche, objectif, à jour, quasi automatique**.

2. Vue d'ensemble du système

Le système fonctionne comme une petite usine en 3 grandes étapes.



Rôle de chaque module (en langage simple)

Module A – Navigation “stealth” (Playwright)

- Ouvre la page comme un vrai utilisateur (même type de navigateur, même taille d'écran, mêmes temps

de pause).

- Permet de contourner les protections anti-robots *sans forcer* : l'idée est de se comporter comme un visiteur normal.

Module B – Nettoyage HTML (BeautifulSoup)

- La page brute contient : texte + images + publicités + scripts de tracking + mise en page...

- Le module garde **uniquement le contenu utile** (titres, paragraphes, tableaux), et supprime tout le reste.

- Résultat : le texte est 20 fois plus léger → **moins cher et plus simple à analyser avec l'IA**.

Module C – Extraction par IA (GPT + schémas de données)

- L'IA reçoit le texte nettoyé + un "gabarit" de fiche produit (schéma).

- Elle extrait : marque, modèle, poids, drop, dimensions, scores de laboratoire, points forts, points faibles, etc.

- Le résultat est un **JSON structuré**, validé par un schéma qui impose des types (nombre, texte, liste...) et des limites (ex : un poids entre 100 et 500 g).

3. Ce que le système récupère concrètement

Aujourd'hui, une fiche StrideMatch pour une chaussure contient **une quarantaine de champs** que l'on peut regrouper en 6 blocs.

L'idée n'est pas de mémoriser tous les champs, mais de comprendre **les grandes familles**.

3.1. Bloc 1 – Identification (qui suis-je ?)

Rôle	Exemples
Identifier la chaussure	Marque (Nike, Asics...), Nom du modèle (Pegasus 41 ...), Catégorie (route, trail...), Genre (H/F/unisexe)
Retrouver la source	URL d'origine (RunRepeat, blog FR...), date du scraping

Utilité côté produit

Filtrer facilement ("toutes les Nike trail"), tracer d'où vient l'information, vérifier ou mettre à jour une fiche.

3.2. Bloc 2 – Spécifications techniques de base

Ce sont les **critères que la plupart des coureurs connaissent** :

- **Poids (g)** : léger pour la vitesse, plus lourd pour la protection.
- **Drop (mm)** : différence de hauteur talon / avant-pied (influence la foulée).
- **Hauteur d'amorti au talon / à l'avant-pied (stack)** : protection, "épaisseur de mousse".
- **Prix** : en devise standard (USD ou converti).
- **Usage** : compétition, entraînement quotidien, récupération...
- **Type de support** : neutre, stabilité, pronateur.
- **Type de foulée recommandé** : talon, médio-pied, avant-pied.

Utilité côté produit

Créer des filtres simples : "moins de 250 g", "drop 4-6 mm", "chaussure neutre pour marathon", etc.

3.3. Bloc 3 – Tests de laboratoire (l'avantage unique)

C'est le cœur de la valeur ajoutée : des **mesures objectives**, réalisées dans un laboratoire indépendant (RunRepeat).

Quelques exemples parlants :

- **Absorption des chocs au talon / avant-pied** : capacité de la chaussure à amortir l'impact (intéressant pour coureurs lourds, longues distances).
- **Retour d'énergie (%)** : à quel point la mousse "renvoie" l'énergie (zoomX vs mousse classique...).
- **Dureté de la mousse** : plus ou moins ferme (impact sur confort, stabilité).
- **Flexibilité & rigidité en torsion** : chaussure plus souple ou plus rigide (guidage du pied).
- **Adhérence (coefficient de traction)** : accroche sur sol mouillé.
- **Respirabilité** : test de ventilation du mesh.
- **Comportement au froid** : combien la mousse durcit à -10°C.

Pourquoi c'est clé pour StrideMatch ?

- Ces données sont **rares** (très peu de sites les ont).
 - Elles permettent des recommandations **objectives** ("si tu fais 85 kg et tu cours sur bitume, cherche une absorption de choc au talon élevée").
-

3.4. Bloc 4 – Dimensions & fit (ajustement)

Objectif : adresser **le vrai problème de l'e-commerce** : les retours à cause d'un mauvais fit.

Principales mesures :

- **Largeur de l'avant-pied (toebox)** : pieds fins, normaux, larges.
- **Hauteur de l'avant-pied** : espace pour les orteils, oignons, ongles fragiles.
- **Largeur générale du chaussant** : équivalent des largeurs B, D, 2E, 4E.
- **Largeur de la semelle au talon et à l'avant-pied** : stabilité.
- **Épaisseur de la semelle externe** : durabilité.
- **Épaisseur de la langue** : confort avec les lacets.

Impact business

Mieux dimensionner les recommandations = **moins de retours**, plus de satisfaction.

3.5. Bloc 5 – Durabilité

Mesures pour estimer **la durée de vie** de la chaussure :

- Résistance de l'avant-pied (test d'abrasion).
- Résistance de la mousse au talon dans le temps.
- Usure de la semelle externe après test.

Ces scores permettent d'estimer, par exemple : "cette chaussure tiendra environ 900 km, soit plus que la moyenne du marché".

3.6. Bloc 6 – Métadonnées & ressenti utilisateur

Ce que l'on récupère ou dérive :

- **Largeurs disponibles** (normal, large, extra-large...).
- **Saison d'usage** (hiver, été, toutes saisons).
- **Semelle amovible** (pour orthèses).
- **Éléments réfléchissants** (sécurité de nuit).
- **Note globale (score)**.
- **Listes de points forts / points faibles** extraits automatiquement des avis.

Utilité

Habiller la fiche produit et enrichir les filtres ("compatible semelles orthopédiques", "bon pour la nuit", etc.).

4. Comment sont stockées les données ? (Vue base de données)

Aujourd'hui, les données sont encore dans **un fichier Excel**. Ce n'est pas adapté à un produit SaaS à long terme.

La cible est une base **PostgreSQL** bien structurée.

4.1. De Excel à PostgreSQL

Limites d'Excel (situation actuelle) : - Colonne en doublon (FR / EN).

- Beaucoup de champs vides (prix, liens, dimensions).
- Pas de validation : poids absurdes possibles, unités mélangées.
- Pas de traçabilité (on ne sait pas d'où vient la donnée ni quand elle a été mise à jour).

Avantages d'une base PostgreSQL : - Performante même avec des **milliers de chaussures**.

- Contrôles automatiques : par exemple "le poids doit être entre 100 et 500 g".
- Relation avec d'autres tables : historique de prix, logs de scraping, candidats doublons...
- Intégration directe dans une application (API, back-office, front...).

4.2. Schéma logique (simplifié)

Table: running_shoes

- ├ Identification (marque, modèle, catégorie, genre, URL source...)
- ├ Specs de base (poids, drop, stack, prix, usage, type de foulée...)
- ├ Tests lab (amorti, retour d'énergie, flexibilité, adhérence...)
- ├ Dimensions & fit (largeurs, hauteurs, épaisseurs...)
- ├ Durabilité (scores de résistance, usure estimée...)
- ├ Métadonnées (largeurs dispo, saison, score, pros/cons...)
- └ Champs techniques (dates de création, mise à jour, etc.)

Tables satellites

└ scraping_logs	→ historique des runs de scraping
└ price_history	→ évolution des prix dans le temps
└ duplicate_candidates	→ candidats doublons à valider
└ scraping_sources	→ liste des sites, fréquence, priorité

L'idée est de **garder Excel comme point de départ**, mais que la vérité de référence devienne la base PostgreSQL.

5. Où en est-on par rapport au catalogue actuel ?

5.1. Diagnostic du fichier Excel existant

- ~140 modèles de chaussures.
- 54 colonnes, dont la moitié en doublon FR/EN.
- Plusieurs colonnes importantes **vides** (prix, liens, certaines dimensions).
- Pas d'URL source → impossible de revalider automatiquement.

5.2. Complémentarité avec le scraping

Aujourd'hui : - Excel couvre une partie des **champs basiques**, mais **de manière incomplète**.

- Le scraping apporte : - Les mêmes champs de base, mais **remplis automatiquement**.

- Surtout, **tout le bloc "tests de labo" + dimensions détaillées + durabilité**.

Objectif : - Importer les 140 modèles Excel dans PostgreSQL.

- Enrichir ces modèles avec les données scrapées (RunRepeat et autres).

- Atteindre un **taux de complétude proche de 100 %** sur les champs importants.

6. Extension à 15 sites (et pas seulement RunRepeat)

Le système a été pensé pour **ne pas dépendre d'un seul site**.

6.1. Types de sites visés

- **RunRepeat** : laboratoire + avis → données factuelles très riches.
- **Sites FR de tests running** (Chaussure Running, Running Addict, Journal du Trail...) : retours terrain, avis détaillés, vocabulaire français.
- **Comparateurs de prix** : meilleures offres, promotions, stock.

6.2. Pourquoi l'approche IA simplifie tout

Avec une approche classique, il faudrait **coder un parser spécifique par site**.

Avec l'IA, on peut garder **le même "cerveau" (prompt + schéma)** pour tous les sites :

Texte nettoyé du site X



[Même prompt IA]



Fiche structurée StrideMatch

Résultat : - Ajout d'un nouveau site = quelques minutes de configuration, pas une journée de développement.

- Le système reste **souple** aux changements de mise en page.

7. Automatisation, dédoublonnage et supervision humaine

7.1. L'automatisation (scheduler)

Un planificateur (type Celery Beat) lance automatiquement :

- **Tous les jours :**
 - Nouvelles chaussures et mises à jour de RunRepeat.
 - Prix sur le comparateur choisi.
- **Toutes les semaines :**
 - Les principaux sites FR et EN de tests (nouvelles reviews, nouveaux modèles).
- **Tous les mois :**
 - Un "grand ménage" : re-scraping complet, nettoyage de la base, génération d'un rapport.

Visuellement :

[Celery Beat (agenda)]

├ Tâches quotidiennes (RunRepeat, prix)

├ Tâches hebdo (sites FR/EN)

└ Tâches mensuelles (rescraping global, rapport)



[Workers] → [Scraping + IA] → [PostgreSQL]

7.2. Détection et gestion des doublons

Même modèle peut apparaître sur plusieurs sites (nom légèrement différent, poids mesuré différemment, etc.).

Le système : - Compare **nom, marque, poids, drop, prix...** pour calculer un **score de similarité**.

- Si la similarité est très forte → fusion automatique.

- Si la similarité est moyenne → la chaussure passe dans une **file de validation** pour un humain.

Scraping multi-sites



[Calcul de similarité]

- |— Score $\geq 0,90$ → fusion automatique
- |— 0,70–0,90 → à valider dans un dashboard
- |— $< 0,70$ → nouvelle chaussure

Un petit dashboard (web) permet de : - Voir les paires suspects (“Vomero 17” vs “Vomero Plus”).
- Décider : **fusionner** / **garder séparé** / **ignorer**.

Temps humain estimé : quelques minutes par semaine.

7.3. Règles de mise à jour

Certaines mises à jour peuvent être faites **sans intervention humaine** :

- Prix légèrement modifié (ex : -10 %).
- Ajout d’une nouvelle largeur disponible.
- Passage de “en stock” à “rupture”.

En revanche, les changements qui touchent au **profil technique** de la chaussure (poids très différent, drop qui change, etc.) déclenchent une **validation humaine**.

8. Limites, risques et cadre légal

8.1. Limites techniques

- Certains sites utilisent des protections anti-bot (Cloudflare, CAPTCHA...).
- La structure HTML peut changer (refonte de site), ce qui peut dégrader la qualité de l’extraction.
- La qualité des données dépend aussi de la **fiabilité des sources** (une erreur sur un blog reste possible).

Mesures de mitigation : - Comportement “humain” (délais, nombre de requêtes limité).

- Surveillance des taux d’erreurs par site + alertes.

- Croiser les sources pour détecter les incohérences (ex : poids très différent).

8.2. Cadre légal (en résumé)

- Les données récupérées sont **publiques** (pas de login, pas de paywall).
- On collecte surtout des **données factuelles** (poids, dimensions, prix...) qui ne sont pas protégées par le droit d’auteur.
- On ne copie pas les contenus créatifs (photos, texte long de reviews) ; on en tire des **résumés** et des **champs structurés**.
- Le scraping est réalisé de manière **raisonnable** (pas de surcharge des serveurs).

En l’état, l’usage est **compatible avec le cadre légal UE** pour de l’agrégation de données publiques, sous réserve de rester “bon citoyen” (respect des sites, pas de comportement agressif).

9. Synthèse exécutive

- StrideMatch dispose d'un **pipeline complet** pour transformer des pages web en **fiches produits ultra détaillées** : navigation furtive → nettoyage → extraction IA.
- Les données issues des **tests de laboratoire** sont un **avantage concurrentiel fort**, surtout combinées aux mesures de fit et de durabilité.
- Le passage d'Excel à **PostgreSQL** est clé pour faire évoluer le projet vers un **vrai produit SaaS** (scalabilité, qualité, traçabilité).
- L'extension à une quinzaine de sites permet de couvrir **la quasi-totalité du marché** en nouveautés, en restant raisonnable en coût grâce à l'IA.
- Un système d'**automatisation + supervision humaine légère** (dédoublonnage, validations critiques) permet de garder un haut niveau de confiance dans les données, sans charge opérationnelle lourde.

Ce document peut servir de **base de discussion** autant avec : - un CTO (vision architecture & techno),
- un product / business (valeur métier, ROI),
- qu'un investisseur (scalabilité, différenciation, défendabilité des données).