Document stratégique (E1) : Du scraping à l'intelligence artificielle – Étude des options technologiques

Table des matières

1. Analyse approfondie du contexte et des problématiques

Le secteur de l'horlogerie de luxe connaît une transformation numérique progressive, stimulée par la montée des plateformes de revente, l'émergence d'outils de pricing automatisé, et l'évolution du comportement des consommateurs vers des décisions d'achat plus informées. Dans ce contexte, la donnée devient une ressource stratégique, permettant de comparer les modèles, de suivre les tendances de prix, d'authentifier les produits ou encore de personnaliser les recommandations.

Cependant, cette donnée est encore **majoritairement non structurée**, issue de fiches produits très hétérogènes publiées sur des sites comme EveryWatch.com. La description des montres est souvent rédigée librement, avec des vocabulaires variables, des abréviations, des fautes ou encore des doublons partiels. Ces conditions nuisent fortement à l'exploitation automatique des informations.

Dans le cadre de ce projet, l'enjeu est donc **d'industrialiser la collecte, la structuration et l'analyse de données horlogères**, en automatisant au maximum les étapes de récupération, d'interprétation sémantique, de catégorisation et d'enrichissement des données issues des plateformes spécialisées.

Les **problématiques centrales** peuvent être formulées ainsi :

- **Comment extraire de manière fiable des informations techniques (matériau, taille, référence, calibre, etc.) à partir de textes libres et non normalisés ?**
- **Comment rendre ces données exploitables en aval pour des usages métiers (veille concurrentielle, pricing prédictif, sourcing automatisé) ?**
- **Comment articuler différentes briques d'intelligence artificielle pour optimiser la valeur ajoutée générée tout en maîtrisant les coûts et la complexité du projet ?**

À travers cette étude, nous proposons une cartographie des options d'intelligence artificielle mobilisables dans une optique d'**automatisation intelligente** du traitement des données horlogères. Cependant, il est impératif de souligner que **l'intelligence artificielle ne peut être envisagée de manière isolée**. Son efficacité repose sur la **qualité et l'accessibilité des données en amont**, or dans notre cas, ces données sont **non accessibles via des API officielles**, et **non structurées dans les pages web**.

Ainsi, **le web scraping constitue une brique technologique stratégique**, préalable à toute exploitation par l'IA. Le scraping permet de **collecter massivement les informations** visibles sur le site EveryWatch, en contournant l'absence d'API, tout en respectant les conditions légales et techniques d'accès aux données publiques.

L'intelligence artificielle intervient **en complémentarité**, en apportant de la valeur dans les étapes de **compréhension, structuration, catégorisation, prédiction, ou interaction** autour de ces données.

2. Architecture technologique : du scraping aux briques d'intelligence artificielle

Scraping des données horlogères : prérequis technique incontournable

Le scraping désigne l'action d'extraire automatiquement des informations depuis une page web. Dans notre projet, cette technique est **essentielle pour accéder aux données produits de EveryWatch.com**, en l'absence d'API publique.

Nous avons mobilisé l'outil **Playwright en Python**, qui permet de simuler le comportement d'un navigateur humain et de **récupérer de manière fiable** les fiches montres (titre, image, lien, description, estimation, etc.). Cette première étape garantit une **collecte massive, reproductible et industrialisable** des données brutes du site.

Le scraping est donc la **colonne vertébrale technique** du projet, sur laquelle les briques d'IA viennent ensuite se greffer. Il permet de :

- Constituer un **corpus de travail solide** pour l'IA ;
- Déployer des processus de mise à jour automatisée ;
- Réduire drastiquement le **temps humain de collecte** (estimé à 2–3 min par fiche en manuel).

Traitement des blocs de texte par intelligence artificielle (modèles OpenAI)

Une fois les données collectées par scraping, notamment les descriptions textuelles riches mais hétérogènes des montres, se pose la question de leur structuration. Ces descriptions contiennent de nombreuses informations techniques essentielles (taille, matériau, mouvement, numéro de série, présence d'accessoires, etc.), mais elles sont exprimées dans un langage naturel libre, sans format standard. Leur traitement nécessite donc une capacité à comprendre le sens et la structure implicite du texte.

Afin d'automatiser cette étape d'extraction, nous avons retenu l'utilisation de **modèles de traitement du langage naturel avancés fournis par OpenAI**, en particulier via l'API GPT-4. Ce choix repose sur la capacité démontrée de ces modèles à interpréter des blocs de texte complexes, même en présence de formulations variées, d'abréviations, ou d'informations techniques imbriquées.

Le modèle est sollicité à l'aide d'un **prompt structuré** demandant explicitement de retourner un dictionnaire JSON contenant les principaux attributs horlogers recherchés. En cas d'absence d'une information dans le texte, le modèle retourne une valeur `null`, garantissant une structure homogène en sortie.

Exemple d'attributs extraits automatiquement :

- `case_size`
- `material`
- `movement`
- `glass`
- `reference_number`

- `serial_number`
- `movement_number`
- `case_number`
- `limited_numbered`
- `accessories`

Cette méthode présente plusieurs **avantages décisifs** :

- Elle ne nécessite **aucun entraînement supervisé** préalable : le modèle est immédiatement opérationnel.
- Elle est **hautement adaptable** : le prompt peut être modifié pour s'ajuster à d'autres marques, langues ou styles rédactionnels.
- Elle permet un **traitement en série** rapide des blocs de texte extraits, avec un taux de structuration élevé dès les premiers tests.

En comparaison avec des méthodes classiques d'analyse syntaxique (regex, spaCy, etc.), le recours à OpenAI permet de **réduire fortement la complexité de développement** tout en obtenant une interprétation fine, y compris sur des cas ambigus ou des expressions inhabituelles.

Ce traitement constitue ainsi la **première brique d'intelligence artificielle effectivement intégrée au projet**, en complément direct du module de scraping.

Comparaison des approches de structuration textuelle

Critère	**OpenAI (GPT-4)**	**Regex (règles manuelles)**	**spaCy (NER classique)**
Efficacité sur des textes complexes	Excellente compréhension même sans format fixe	Faible, dépend strictement de la syntaxe	Moyenne, nécessite des patterns clairs
Facilité de mise en œuvre	Très simple via API et prompt	Assez simple, mais fastidieux	Complexe : nécessite entraînement ou pipelines
Temps de développement initial	Très rapide (prompt + appel API)	Long si nombreux cas à couvrir	Moyen à long, selon qualité des entités
Coût d'exécution	Moyen à élevé (API payante, facturation au token)	Très faible	Gratuit (open source)
Adaptabilité à d'autres marques / langues	Excellente (prompt modifiable à volonté)	Très faible (tout est à réécrire)	Moyenne, dépend des modèles pré-entraînés
Robustesse aux variations de style	Très bonne tolérance	Nulle	Faible à moyenne
Maintenance / scalabilité	Faible besoin de maintenance (prompt évolutif)	Maintenance élevée si structure évolue	Maintenance moyenne
Besoins en données d'apprentissage	Aucun	Aucun	Oui si modèle personnalisé

3. Recommandations stratégiques argumentées

Dans le cadre de ce projet, plusieurs approches d'intelligence artificielle ont été explorées pour structurer automatiquement les blocs de texte extraits par scraping depuis le site EveryWatch. Ces blocs contiennent des informations techniques déterminantes pour la valorisation des données horlogères (taille, mouvement, verre, numéro de série, etc.), mais leur hétérogénéité rend leur traitement complexe.

Nous avons identifié trois scénarios technologiques viables, correspondant à trois niveaux de sophistication, et analysé leurs apports respectifs selon des critères de coût, d'efficacité, de complexité de mise en œuvre et de maintenabilité.

Scénario 1 — Extraction par GPT-4 via l'API OpenAI *(choix retenu)*

Ce scénario repose sur l'utilisation de l'API OpenAI (modèle GPT-4) pour interpréter les blocs de texte, via un prompt conçu pour extraire les attributs techniques au format JSON. L'approche est **zero-shot**, ne nécessitant ni entraînement préalable ni jeux d'exemples annotés.

L'outil se comporte comme un lecteur expert, capable d'extraire correctement des entités même en présence de descriptions informelles, d'abréviations, ou de variations de formulation. En cas d'absence d'information dans le texte, la clé correspondante est systématiquement renseignée avec une valeur `null`, assurant une structure uniforme en sortie.

Avantages stratégiques :

- Mise en œuvre rapide (prompt + appel API),
- Excellente tolérance aux variations de style et aux cas ambigus,
- Maintenance minimale : le prompt peut être adapté à mesure que les besoins évoluent,
- Adaptabilité élevée à d'autres marques, langues ou plateformes.

Inconvénients :

- Coût variable selon le volume de requêtes et la longueur des textes,
- Dépendance à une infrastructure tierce (OpenAI),
- Nécessite une gestion rigoureuse des quotas et des clés API.

Ce scénario a été retenu en raison de sa **pertinence immédiate, sa performance qualitative, et sa compatibilité avec un déploiement rapide et scalable.**

Scénario 2 — Extraction par expressions régulières (regex)

Ce scénario repose sur la construction manuelle de règles syntaxiques (regex) pour identifier des entités précises dans les descriptions. Il s'agit d'une méthode déterministe, peu coûteuse, et historiquement utilisée dans les pipelines de text mining.

L'approche consiste à définir des patrons (patterns) pour chaque attribut attendu, en tenant compte des variations connues : ex. `\d{2,3}mm` pour la taille, `calibre\s+[A-Z0-9\.\-]+` pour le mouvement, etc.

Avantages stratégiques :

- Coût de déploiement nul,
- Interprétabilité maximale (chaque règle est explicite),
- Rapidité d'exécution.

Inconvénients :

- Faible robustesse : dès que le format varie ou que le champ est mal formulé, la règle échoue,
- Maintenance difficile : toute évolution du corpus implique des ajustements manuels,
- Inefficace sur des cas complexes ou ambigus (inversion syntaxique, éléments multiples, absence d'unités).

Ce scénario a été écarté en raison de sa **fragilité face à la diversité linguistique des fiches produits**, et de son manque d'évolutivité. Il peut toutefois servir de **solution de secours** ou de filtre préliminaire dans un pipeline mixte.

Scénario 3 — Extraction via spaCy (NER classique)

Ce scénario s'appuie sur spaCy, une librairie open source de NLP capable de réaliser de l'extraction d'entités nommées (NER). Il existe deux possibilités dans ce cadre :

1. Utiliser un modèle pré-entraîné généraliste (peu pertinent ici),
2. Entraîner un modèle personnalisé sur un corpus d'exemples annotés (nécessite une phase de labellisation manuelle).

Le modèle peut apprendre à reconnaître les attributs horlogers à partir de phrases d'exemple telles que :
"Boîtier en acier 36 mm, mouvement automatique calibre 240, verre saphir."

Avantages stratégiques :

- Approche open source, sans coût d'API,
- Possibilité de personnalisation fine,
- Traitement local (pas de dépendance à un service externe).

Inconvénients :

- Temps de développement important (labellisation + entraînement),
- Nécessite un volume significatif de données annotées,
- Résultats incertains sur des textes peu normalisés,
- Difficulté à gérer des cas marginaux ou implicites.

Ce scénario a été exploré mais écarté, car il ne répondait pas aux **contraintes de temps** et de **ressources humaines disponibles pour constituer un corpus d'apprentissage suffisant**. Il pourrait être reconsidéré dans un cadre de projet plus long terme.

Conclusion générale

Ce document a permis d'explorer les différentes options technologiques mobilisables pour répondre à la problématique centrale du projet : structurer et valoriser automatiquement les données horlogères issues de plateformes non normalisées, dans un contexte d'absence d'API et de forte hétérogénéité des formats.

La phase de **scraping automatisé via Playwright** constitue la fondation technique indispensable, garantissant une collecte massive et reproductible des fiches produits. Sur cette base, plusieurs scénarios d'intégration de l'intelligence artificielle ont été envisagés, allant d'approches classiques par expressions régulières à des modèles entraînés de type spaCy, jusqu'aux solutions d'IA générative proposées par OpenAI.

L'analyse croisée de ces scénarios, selon des critères de faisabilité, de robustesse, de coût, et d'évolutivité, a conduit à recommander l'intégration des **modèles GPT-4 via l'API OpenAI** pour le traitement des blocs de texte. Cette solution offre une performance immédiate, une excellente adaptabilité à la diversité des formulations, et une évolutivité maîtrisée via la modification simple du prompt.

Cette approche permet ainsi de construire un pipeline intelligent, du scraping à la structuration automatisée, ouvrant la voie à des usages avancés de la donnée : prédiction de prix, classification automatique, ou encore interactions conversationnelles avec une base horlogère enrichie.

La stratégie retenue repose donc sur une articulation claire entre automatisation de la collecte et intelligence sémantique du traitement, avec pour finalité une base de données horlogères exploitable à haute valeur ajoutée, adaptée à une industrialisation progressive.