Université du Québec en Outaouais
Département des sciences administratives

SPID

SEC Project Intelligence Database

Provenance des données, fonctionnement du pipeline d'extraction
et portrait statistique de 121 099 projets stratégiques
extraits des filings de 2 492 sociétés cotées américaines (2010–2026)

DuckDB · 10-K · 10-Q · 8-K · Extraction par LLM · Graphe de connaissances · Embeddings

Manel Kammoun
manel.kammoun@uqo.ca

Charli Tandja Mbianda
djerrycharli.tandjambianda@uqo.ca

Simon-Pierre Boucher
simon-pierre.boucher@uqo.ca

Gatineau — Pavillon Alexandre-Taché, 283, boulevard Alexandre-Taché

Rapport technique — 28 octobre 2026

Résumé

Ce rapport documente **SPID** (*SEC Project Intelligence Database*), une base de données qui reconstruit le portefeuille de *projets stratégiques* des grandes sociétés cotées américaines à partir du texte brut de leurs documents réglementaires déposés auprès de la Securities and Exchange Commission (SEC). Plutôt que de lire les formulaires 10-K, 10-Q et 8-K comme des documents financiers, SPID les traite comme une source d'*intelligence sur les projets* : usines, centres de données, acquisitions, programmes de R&D, initiatives d'IA, transition énergétique, etc.

Le document répond à trois questions. **D'où viennent les données ?** Elles proviennent du corpus EDGAR de l'UQO : les filings de 2 492 sociétés de l'univers **S&P 500** ∪ **Russell 3000** (2 564 sociétés, extension du 27–28 octobre 2026) téléchargés depuis l'API publique de la SEC, soit 527 548 documents et quelque 350 Go de HTML brut répartis sur dix nœuds du cluster MacLustr, découpés en sections. **Comment fonctionne le pipeline ?** Cinq étapes reprenables (ingestion des 10-K, extraction par grand modèle de langage, résolution des mentions en projets, graphe de connaissances, embeddings) persistent leurs résultats dans un unique fichier DuckDB de 3,2 Go. **Que contient la base ?** 1 123 098 sections analysées, 308 065 mentions de projets, 121 099 projets résolus répartis en 21 types et 11 secteurs, un graphe de 183 472 nœuds et 121 099 vecteurs sémantiques de 384 dimensions. Une trentaine de figures et de tableaux dressent le portrait statistique du jeu de données, et une section finale en expose les limites.

Chiffres clés de la base `spid.duckdb` (exécution complète, 28 octobre 2026)

- Entreprises couvertes : **2 492** (S&P 500 ∪ Russell 3000)
- Période des dépôts : **janv. 2010 – oct. 2026**
- Sections de filings analysées : **1 123 098**
- Filings contenant au moins un projet : **135 304**
- Mentions de projets extraites : **308 065**
- Projets résolus : **121 099**
- Types de projets (taxonomie) : **21**
- Secteurs GICS : **11**
- Projets avec montant divulgué : **56 241** (46,4 %)
- Montant médian divulgué : **130 M$ US**
- Localisations distinctes : **14 508**
- Technologies distinctes : **8 311**
- Partenaires distincts : **36 822**
- Nœuds / arêtes du graphe : **183 472** / **204 456**
- Embeddings (all-MiniLM-L6-v2, 384-d) : **121 099**
- Taille du fichier DuckDB : **3,2 Go**

Table des matières

1 Vue d'ensemble 4

2 Provenance des données 4
- 2.1 La source primaire : EDGAR . . . 4
- 2.2 Le corpus EDGAR de l'UQO . . . 5

3 Fonctionnement du pipeline 6
- 3.1 Étape 1 — Ingestion des 10-K . . . 7
- 3.2 Étape 2 — Extraction par grand modèle de langage . . . 8
  - 3.2.1 Flux unifié . . . 8
  - 3.2.2 Maîtrise du coût : le « gate » et le « focus » . . . 8
  - 3.2.3 Consigne et schéma de sortie . . . 9
  - 3.2.4 Post-traitement et identifiant stable . . . 9
  - 3.2.5 Le backend déterministe `rules` . . . 9
  - 3.2.6 Exécution sur le corpus complet . . . 9
- 3.3 Étape 3 — Résolution des mentions en projets . . . 10
- 3.4 Étape 4 — Graphe de connaissances . . . 11
- 3.5 Étape 5 — Embeddings et recherche sémantique . . . 11
- 3.6 Interfaces : ligne de commande et API . . . 11
- 3.7 Coût, débit et reproductibilité . . . 12

4 Schéma de la base de données 12

5 Portrait statistique des projets 13
- 5.1 Couverture temporelle et formulaires . . . 13
- 5.2 Types de projets . . . 14
- 5.3 Secteurs . . . 15
- 5.4 Statut et cycle de vie . . . 17
- 5.5 Montants divulgués . . . 18
- 5.6 Entreprises . . . 20
- 5.7 Localisations, technologies et partenaires . . . 21

6 Graphe de connaissances et espace sémantique 21

7 Qualité des données et limites 22

8 Reproduire ce rapport 23

A Taxonomie des 21 types de projets 23

B Fichiers du dépôt `project-db/` 24

1 Vue d'ensemble

La stratégie d'une entreprise se matérialise par des *projets* : la construction d'une usine, l'ouverture d'un centre de données, l'intégration d'une société acquise, le lancement d'un programme de R&D, la migration d'un système ERP vers l'infonuagique. Ces initiatives sont pourtant dispersées, non structurées et enfouies dans les sections narratives des documents réglementaires. Aucune base commerciale (Capital IQ, PitchBook, Gartner) ne les recense de manière systématique et *centrée sur le projet*.

SPID comble cette lacune. L'unité d'analyse n'est ni le document ni l'entreprise, mais le **projet**, décrit par un type (taxonomie contrôlée de 21 catégories), un statut, un montant, des localisations, des technologies, des partenaires, des bénéfices attendus, des risques et une chronologie de mentions dans les filings successifs.

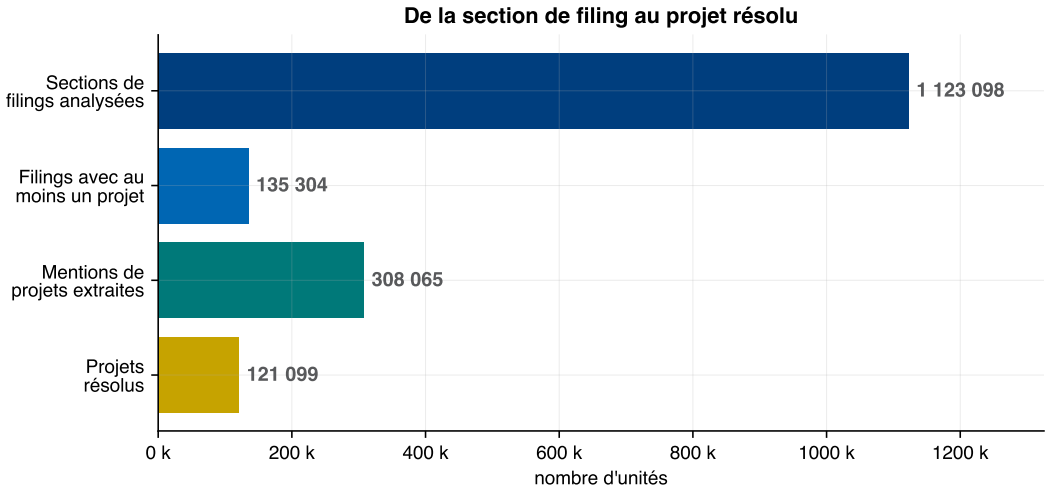


Figure 1. L'entonnoir SPID : des 1 123 098 sections de filings analysées aux 121 099 projets résolus. Chaque projet agrège en moyenne 2,1 mentions.

Le dépôt `project-db/` est volontairement léger (environ 1 900 lignes de Python dans le paquet `spid/`) : il s'appuie sur le corpus EDGAR déjà constitué par l'UQO (section 2) et y ajoute la « couche projet » (section 3). Le résultat tient dans un unique fichier `data/db/spid.duckdb` (section 4), dont la section 5 dresse le portrait statistique.

2 Provenance des données

2.1 La source primaire : EDGAR

Toutes les données textuelles proviennent d'**EDGAR** (*Electronic Data Gathering, Analysis, and Retrieval*), le système de dépôt électronique de la SEC. Toute société cotée aux États-Unis y publie obligatoirement :

- le **10-K**, rapport annuel exhaustif (activités, facteurs de risque, propriétés, rapport de gestion) ;
- le **10-Q**, rapport trimestriel abrégé ;
- le **8-K**, rapport d'événement courant déposé dans les quatre jours ouvrables suivant un fait important (acquisition, contrat, nomination, communiqué de résultats).

EDGAR est public, gratuit et accessible par API (`https://www.sec.gov/edgar`), sous réserve d'un *User-Agent* déclaratif et d'une limite de 10 requêtes par seconde. Le corpus a été téléchargé avec l'identifiant `UQO-Research simon-pierre.boucher@uqo.ca`.

2.2 Le corpus EDGAR de l'UQO

SPID est construit *par-dessus* le corpus `UQO_RESEARCH_DB`, qui réside dans le répertoire parent `../data` et que SPID consomme en lecture seule. Ce corpus a été constitué par la chaîne `src/acquisition` → `src/parsing` → `src/storage` du dépôt amont.

Table 1. Actifs du corpus source utilisés par SPID.

Actif	Volume	Contenu et usage par SPID
`data/raw/sec/`	≈ 350 Go, 2 564 entreprises	HTML brut de chaque dépôt, rangé par `<cik>/<form>/<accession>/`, conservé sur le nœud qui l'a traité (corpus S&P 500 sur M3U96b).
`parquet/filings/`	527 548 filings	Une ligne par dépôt : `accession_number`, `cik`, `ticker`, `company_name`, `form_type`, `filing_date`, `fiscal_year`.
`parquet/filing_sections/`	1 342 283 sections	Texte pré-segmenté des 10-K, 10-Q et 8-K (7,3 Go) : `section_id`, `accession_number`, `section_name`, `item_number`, `text`, `word_count`.
`parquet/companies/`	2 564 lignes	Métadonnées : secteur GICS, code SIC, État d'incorporation, bourse, indicateurs `in_sp500` / `in_russell3000`.

Sélection des entreprises (`src/acquisition/sp500.py`, `universe.py`). L'échantillon initial (juin 2026) était l'indice **S&P 500** ; il a été étendu le 27 octobre 2026 au **Russell 3000**. Faute de liste officielle gratuite, le Russell 3000 est approché par les positions de l'ETF iShares IWV (fichier `latest-holdings.csv` du 23 octobre 2026) : 2 587 lignes « actions » dont 2 555 appariées à un CIK (les CVR, titres en escrow et banques déposant auprès de la FDIC sont écartés). L'univers fusionné (`data/metadata/universe.json`) compte **2 564 sociétés** : 491 présentes dans les deux indices, 9 S&P 500 seulement, 2 064 Russell seulement ; 65 d'entre elles (émetteurs étrangers déposant des 20-F/40-F) n'ont aucun 10-K/10-Q/8-K. Pour le S&P 500, La liste des constituants (ticker, nom, CIK, secteur et sous-industrie GICS, date d'ajout) est lue dans la table de la page Wikipédia *List of S&P 500 companies* et mise en cache dans `data/metadata/sp500_constituents.json`. Chaque ticker est ensuite associé à son identifiant SEC (*Central Index Key*, CIK, complété à 10 chiffres) via le fichier public `company_tickers.json` de la SEC (`cik_mapper.py`). Il s'agit d'un **instantané courant** de l'indice, non d'une composition historique : les sociétés sorties de l'indice avant 2026 sont absentes (biais de survie, section 7).

Téléchargement. Modules `sec_client.py` et `downloader.py`. Pour chaque CIK, l'index des dépôts est lu à `data.sec.gov/submissions/CIK{cik}.json` (bloc `recent` puis fichiers d'archives paginés). Sont retenus les formulaires **10-K**, **10-Q** et **8-K** au sens strict (les amendements `/A` sont exclus) déposés à partir du **1er janvier 2010**. Le document primaire de chaque dépôt est téléchargé dans `data/raw/sec/{cik}/{form}/{accession}/`. Le client respecte la limite de la SEC (intervalle minimal de 0,1 s, soit 10 requêtes/s), effectue jusqu'à 5 tentatives avec attente exponentielle (1, 2, 4, 8, 16 s) et tient un journal JSONL des fichiers reçus qui rend le téléchargement reprenable.

Nettoyage et segmentation amont (`src/parsing/`). Le HTML est nettoyé avec `lxml` (balises `script`, `style`, `meta`, `link` supprimées ; tableaux convertis en texte tabulé ; fichiers de plus de 50 Mo ignorés). Le découpage repère les en-têtes « ITEM *n* » et « PART I/II » par expressions régulières : les **10-Q** sont découpés en 7 sections (états financiers, MD&A, risques de marché, contrôles, litiges, facteurs de risque, ventes non enregistrées) ; les **8-K**, trop courts et hétérogènes, sont conservés en une seule section `full_body`. Toute

section de moins de 20 mots est rejetée. Les **10-K**, en revanche, n'avaient été laissés qu'en HTML brut : c'est la première tâche de SPID (section 3.1).

Table 2. Composition du fichier `filings.parquet` (527 548 dépôts, 2 499 CIK).

Formulaire	Dépôts	Premier dépôt	Dernier dépôt
8-K	400 606	2010-01-04	2026-10-27
10-Q	95 167	2010-01-05	2026-10-27
10-K	31 775	2010-01-08	2026-10-27

Table 3. Sections éligibles à l'extraction (au moins 30 mots), par formulaire, après ingestion des 10-K par SPID.

Formulaire	Sections	Entreprises (CIK)
8-K	400 601	2 499
10-K	143 704	2 464
10-Q	578 793	2 494
Total	**1 123 098**	**2 499**

Pourquoi seulement 84 entreprises pour les 10-Q ?

Le corpus amont n'a segmenté les 10-Q que pour un sous-ensemble de sociétés ; pour les autres, les 10-Q existent en HTML brut mais n'ont pas été découpés. Les 8-K et les 10-K couvrent en revanche la quasi-totalité de l'échantillon (497 CIK chacun) et fournissent 85 % des mentions de projets (figure 4).

3 Fonctionnement du pipeline

Le traitement s'organise en cinq étapes séquentielles. Chacune lit la sortie de la précédente dans DuckDB, y écrit la sienne de façon **idempotente** (clés stables, `INSERT OR REPLACE`) et peut être relancée sans double comptage.

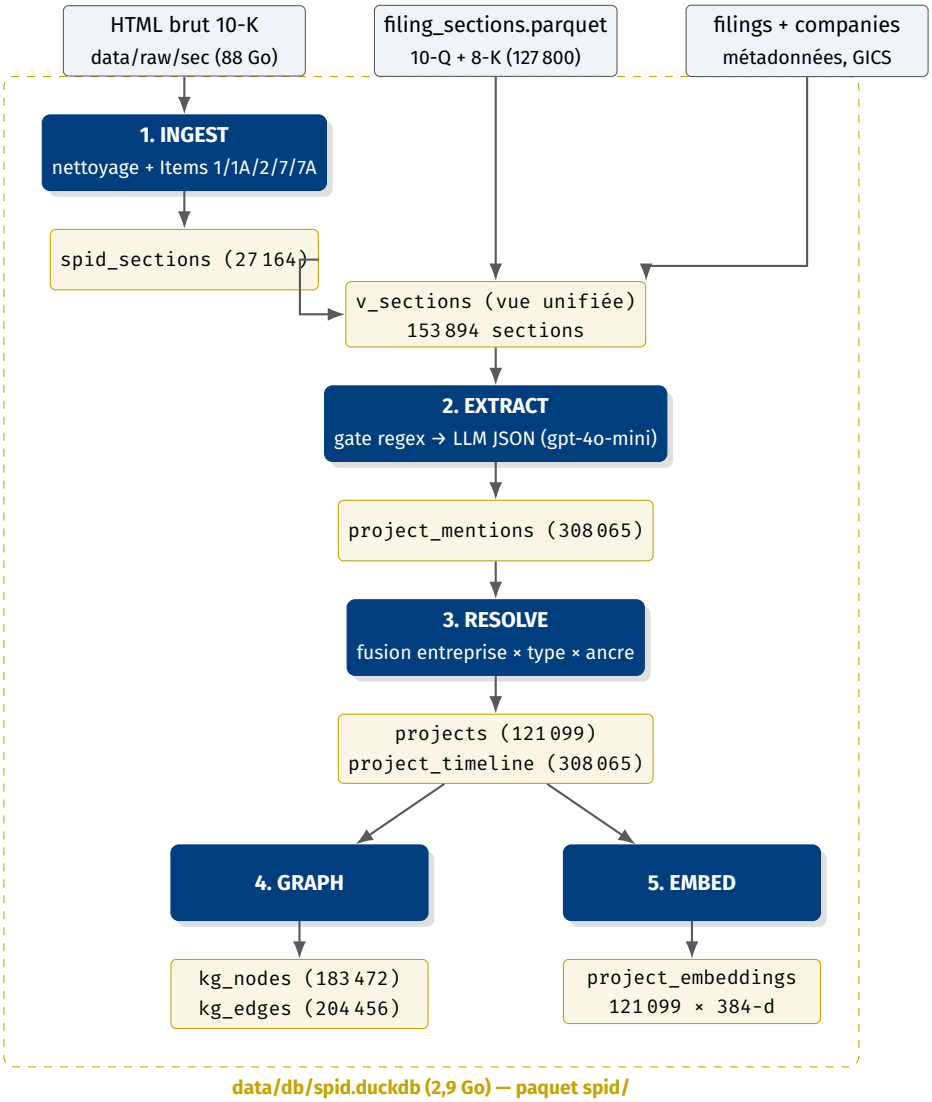


Figure 2. Architecture du pipeline SPID. Les rectangles gris sont les entrées (corpus EDGAR de l'UQO), les rectangles bleus les cinq étapes de traitement et les rectangles ocre les tables DuckDB produites.

3.1 Étape 1 — Ingestion des 10-K

Module : `spid/ingest.py`. Pour chaque dépôt 10-K du répertoire `data/raw/sec/<cik>/10-K/<accession>/` :

1. **Sélection du document primaire** : le plus gros fichier `.htm`, `.html` ou `.txt` du répertoire (les pièces jointes et les exhibits sont écartés).

2. **Nettoyage HTML** : suppression des balises `script`, `style` et `table` avec BeautifulSoup (parseur `lxml`), conversion en texte, dé-échappement des entités, normalisation des espaces. Contrairement au nettoyeur amont, les tableaux sont *supprimés* : les chiffres présentés en tableau dans les 10-K n'atteignent donc pas l'extracteur.

3. **Découpage en Items** : l'expression régulière `ITEM\s+(\d+[AB]?)\s*[.\-:—]` repère les en-têtes et ne conserve que les Items les plus denses en projets : **1** (*Business*), **1A** (*Risk Factors*), **2** (*Properties*), **7** (*MD&A*) et **7A** (*Quantitative Disclosures*). Les occurrences de moins de 400 caractères (renvois de la table des matières) sont écartées ; pour chaque Item, la plus longue occurrence est retenue et tronquée à 600 000 caractères. Si aucun Item n'est reconnu, le corps complet est conservé (`full_body`, 1 230 cas).

4. **Persistance** dans `spid_sections` (identifiant UUID, compte de mots), par lots de 400 lignes. Une accession déjà ingérée est sautée.

Table 4. Sections 10-K produites par l'ingestion (31 559 dépôts 10-K, 143 704 sections, 991 millions de mots).

section_name	Item du 10-K	Sections	Mots (moy.)	Mots (méd.)
`business`	Item 1 — Activités	29 285	7 154	5 515
`risk_factors`	Item 1A — Facteurs de risque	29 158	11 828	9 476
`md_and_a`	Item 7 — Rapport de gestion (MD&A)	28 919	11 077	9 905
`properties`	Item 2 — Propriétés	28 236	614	188
`quantitative_disclosures`	Item 7A — Risques de marché	26 876	1 482	522
`full_body`	Corps complet (repli)	1 230	47 719	44 514

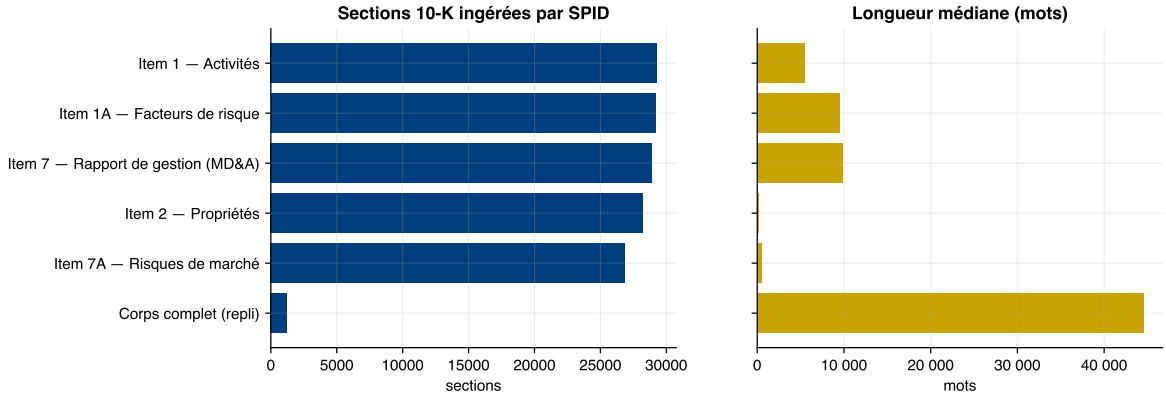


Figure 3. Sections 10-K ingérées par SPID : nombre par Item et longueur médiane en mots. Le rapport de gestion (Item 7) et les facteurs de risque (Item 1A) sont les sections les plus longues ; les propriétés (Item 2) sont courtes mais riches en localisations.

3.2 Étape 2 — Extraction par grand modèle de langage

Modules : `spid/pipeline.py` (moteur), `spid/extract/base.py` (interface), `spid/extract/llm.py` (extracteur LLM), `spid/extract/rules.py` (extracteur déterministe). Le backend est *enfichable* : le même moteur alimente indifféremment le LLM ou les règles.

3.2.1 Flux unifié

La vue `v_sections` réunit les sections 10-Q/8-K du corpus amont et les sections 10-K ingérées, jointes aux métadonnées du dépôt (CIK, ticker, formulaire, date) et de l'entreprise (secteur, sous-industrie, SIC, état). Seules les sections d'au moins 30 mots sont éligibles.

3.2.2 Maîtrise du coût : le « gate » et le « focus »

- **Project gate.** Avant tout appel, une expression régulière de vocabulaire-projet (`_PROJECT_GATE`, racines *project, initiative, program, invest, expan, construct, build, launch, develop, deploy, implement, acqui, partner, facilit, capacity, capital, migrat, transform, pipeline, plant, data center, automat, sustainab, renewable, restructur*) filtre les sections. Une section sans signal de projet ne déclenche *aucun* appel API.

- **Focus.** Si la section dépasse 9 000 caractères, seules les *fenêtres* de texte entourant chaque signal (220 caractères avant, 380 après) sont conservées, triées, fusionnées lorsqu'elles se chevauchent et concaténées jusqu'à un budget de 9 000 caractères. Le message final (« SECTION : *nom* » + texte

focalisé) est plafonné à 14 000 caractères. Ce mécanisme est déterminant pour les 8-K, transmis en une seule section `full_body`.

3.2.3 Consigne et schéma de sortie

Le modèle (`gpt-4o-mini`, température 0, mode JSON `response_format=json_object`, 2 000 jetons de sortie au maximum, délai de 120 s) reçoit une consigne système d'analyste financier et renvoie un objet `{"projects": […]}` dont chaque élément respecte le schéma suivant :

schéma d'une mention de projet (sortie JSON du LLM)

```
{
  "project_name": "",   // nom court (officiel si mentionné)
  "project_type": "",   // EXACTEMENT un des 21 types de la taxonomie
  "description":  "",   // 1 phrase, ancrée dans le texte (pas de spéculation)
  "objective":    "",   // le but / le pourquoi
  "status": "planned | in_progress | completed | mentioned",
  "amount_usd": 0,      // montant en USD si indiqué, sinon null
  "locations": [], "technologies": [], "partners": [], "suppliers": [],
  "benefits":  [], "risks": [], "years": [],
  "confidence": 0.0     // 0..1 : à quel point c'est un vrai projet concret
}
```

Règles imposées au modèle (extrait textuel de la consigne) : « *Extract ONLY real initiatives, not generic risk-factor boilerplate or hypotheticals. If the text describes no concrete project, return {"projects" : []}. Do not invent facts. Numbers/locations must appear in the text. Prefer fewer high-quality projects over many vague ones.* » Les appels sont parallélisés dans un `ThreadPoolExecutor` (16 *workers* lors de l'exécution complète) ; trois tentatives sont faites sur les codes 429 et 5xx (attentes de 1,5, 3 et 4,5 s), après quoi la section est considérée sans projet.

3.2.4 Post-traitement et identifiant stable

Chaque projet renvoyé est normalisé : le `project_type` est ramené à la taxonomie (correspondance exacte, puis floue ; défaut `technology_deployment`) ; le statut est validé (défaut `mentioned`) ; une mention sans description est rejetée ; les montants sont convertis en nombre ; les listes sont tronquées à 8 éléments de 120 caractères ; les années ne sont retenues qu'entre 1980 et 2040 ; la confiance vaut 0,6 par défaut. Si le modèle ne renvoie ni technologies ni années, les expressions régulières de la taxonomie les cherchent dans la description. Les mentions de confiance inférieure à **0,45** sont écartées par le moteur. Un identifiant **stable** est calculé :

$$\texttt{mention_id} = \text{SHA-1}\big(\texttt{accession} \mid \texttt{section_id} \mid \texttt{project_type} \mid \texttt{ancre}\big)[0{:}20], \quad \texttt{ancre} = \texttt{type:nom[0:40]}.$$

Ré-exécuter l'extraction réinsère les mêmes lignes (`INSERT OR REPLACE`) sans doublon : l'extraction est **incrémentale et reprenable**.

3.2.5 Le backend déterministe `rules`

Sans coût d'API, `rules.py` découpe la section en phrases, ne garde que celles de 20 à 700 caractères qui passent le *gate*, les classe par les signaux de la taxonomie (au plus deux types), repère un nom de projet explicite (« Project X », « X Program »), puis calcule une confiance : base 0,40, +0,20 pour un nom explicite, +0,18 pour un montant, +0,10 pour une localisation, +0,06 pour une année, +0,08 pour un statut déterminé, +0,06 pour un verbe d'engagement (*announced, committed*), −0,08 pour un modal d'incertitude (*may, could, potential*). Les phrases sous 0,30 sont abandonnées. Ce backend sert de base de comparaison et d'amorçage ; la base actuelle est intégralement issue du backend `llm`.

3.2.6 Exécution sur le corpus complet

Le pilote `run_full_extract.py` traite les CIK par lots (15 puis 20 CIK) pour borner la mémoire (la matérialisation de toutes les sections d'un seul coup dépassait la limite de 2 Go des tampons de chaînes Arrow).

Le run S&P 500 de juin 2026 a couvert 153 894 sections (500 CIK, 34 lots, 16 appels parallèles) ; l'extension Russell 3000 des 27–28 octobre 2026 a ajouté 968 134 sections (81 lots de 20 CIK, 48 appels parallèles, une douzaine d'heures d'extraction effective) en s'appuyant sur le registre `extracted_sections`, qui remplace l'ancien journal `logs/extract_done_ciks.txt` et garantit qu'aucune section n'est facturée deux fois. Résultat : **308 065** mentions, toutes issues du backend `llm`.

Table 5. Six mentions tirées au hasard parmi celles de confiance $\geq 0{,}90$ et de montant > 100 M$ (types énergie, construction). Montants en M$ US.

Ticker	Dépôt	Type	Nom et description extraits	M$	Conf.
AEE	8-K 2022-06-23	Transition énergétique	**Renewable Generation Expansion** Adding a total of 2,800 megawatts of renewable generation by 2030 and 4,700 megawatts by 2040.	4 300	0.90
MGEE	10-Q 2018-05-08	Transition énergétique	**Saratoga Wind Farm** In December 2017, the PSCW authorized construction of a 66MW wind farm, consisting of 33 turbines, located near Saratoga, Iowa.	108	1.00
CLMT	8-K 2026-09-01	Construction d'usine	**Renewable Fuels Facility Expansion** Montana Renewables, LLC is developing and expanding its renewable fuels facility in Great Falls, Montana, with a focus on repurposing existing equipment.	816	0.90
AOSL	10-K 2019-08-23	Construction d'usine	**Joint Venture for Semiconductor Facility** We formed a joint venture with the Chongqing Funds to construct a power semiconductor packaging, testing and wafer fabrication facility in Chongqing, China.	330	1.00
NEE	8-K 2012-09-28	Construction d'usine	**Canyon Wind Wind Generating Facilities** Canyon Wind, LLC is developing and constructing wind generating facilities with a generating capability totaling approximately 219 megawatts located in Arizona and Michigan.	232	0.90
BHVN	8-K 2026-08-26	Pipeline de médicaments	**Opakalim Development** Opakalim is an investigational, selective Kv7.2/7.3 potassium channel activator currently in Phase 2/3 development for focal epilepsy.	400	0.90

3.3 Étape 3 — Résolution des mentions en projets

Module : `spid/resolve.py`, entièrement en SQL DuckDB. Un même projet est en général mentionné dans plusieurs filings et sur plusieurs années (le 8-K d'annonce, puis les 10-K et 10-Q de suivi). La résolution regroupe les mentions par une clé exacte à trois composantes :

$$\texttt{gkey} = \texttt{cik} \mid \texttt{project_type} \mid \text{ancre}, \qquad \text{ancre} = \begin{cases} \text{première localisation (minuscules)} & \text{si présente,} \\ \text{premier mot de 4 lettres ou plus du nom} & \text{sinon,} \\ \texttt{general} & \text{à défaut,} \end{cases}$$

et `project_id` $=$ MD5(`gkey`). Il n'y a pas de similarité floue : deux mentions fusionnent si et seulement si elles partagent entreprise, type et ancre. Pour chaque groupe, une ligne de `projects` est produite :

- `project_name` et `description` : ceux de la mention la plus confiante (`arg_max` sur `confidence`) ;
- `canonical_location` : l'ancre géographique ; `technologies` : union dédupliquée ;
- `status` : celui du dépôt le plus récent (`arg_max` sur `filing_date`) ;

- `total_amount_usd` : le **maximum** des montants divulgués (et non la somme, pour ne pas compter plusieurs fois un même montant répété de filing en filing) ;
- `n_mentions`, `n_filings` (accessions distinctes), `avg_confidence` ;
- `first_seen` / `last_seen` : dates de dépôt extrêmes ; `first_year` / `last_year` : années extrêmes citées dans le texte.

Chaque mention datée est conservée dans `project_timeline` (dépôt, formulaire, statut, montant, confiance, extrait de 300 caractères), à raison d'une ligne par mention, ce qui permet de suivre l'évolution d'un projet de son annonce à son achèvement. Contrairement à l'extraction, la résolution (comme le graphe et les embeddings) reconstruit ses tables intégralement à chaque exécution.

Résultat : 308 065 mentions → 121 099 projets ; 60,2 % des projets ne reposent que sur une mention, tandis que les projets les plus suivis cumulent jusqu'à 185 mentions (tableau 10).

3.4 Étape 4 — Graphe de connaissances

Module : `spid/graph.py`. Les projets résolus sont projetés en un graphe orienté de cinq types de nœuds et trois relations :

Nœud	Préfixe		Relation
Company (2 492)	`C:cik`	`owns` →	Project
Project (121 099)	`P:project_id`	`located_in` →	Location
Location (14 508)	`L:lieu`	`uses` →	Technology
Technology (8 551)	`T:techno`		
Partner (36 822)	`PR:nom`		

Les propriétés sont stockées en JSON (entreprise : cik, ticker, secteur ; projet : type, statut, montant, nombre de filings, première et dernière observation ; arête `owns` : montant). Le graphe est exporté en parquet et en Cypher (`data/exports/kg_import.cypher`, 6,6 Mo, instructions `MERGE`) pour Neo4j. Les nœuds de casse variable (partenaires, noms d'entreprise) sont agrégés par `GROUP BY` avec `any_value()` afin de respecter la clé primaire. Les 7 640 nœuds `Partner` sont créés à partir des mentions mais la relation `partners_with` n'est pas encore matérialisée : ils n'ont pas d'arête entrante dans la version actuelle.

3.5 Étape 5 — Embeddings et recherche sémantique

Module : `spid/embed.py`. Le texte de chaque projet (« *nom. type. description.* technologies : … ») est encodé en un vecteur de **384 dimensions** normalisé (norme L2 unitaire) par le modèle local `all-MiniLM-L6-v2` (`sentence-transformers`, lots de 256), sans appel API. Les vecteurs sont stockés dans `project_embeddings`. La recherche sémantique (`spid semantic "AI projects launched by banks"`) encode la requête, calcule la similarité cosinus par un simple produit matriciel sur les 19 227 vecteurs (recherche exhaustive, sans index approché), sur-échantillonne cinq fois le nombre de résultats demandés puis applique les filtres SQL éventuels.

3.6 Interfaces : ligne de commande et API

commandes du paquet spid (exécutées depuis project-db/)

```
PY=../.venv/bin/python
PYTHONPATH=. $PY -m spid ingest --tickers NVDA,AAPL,MSFT     # 1. sectionner les 10-K bruts
PYTHONPATH=. $PY -m spid extract --backend llm --workers 16  # 2. extraire (ou --backend rules)
PYTHONPATH=. $PY -m spid resolve                             # 3. mentions -> projets + timeline
PYTHONPATH=. $PY -m spid graph                               # 4. graphe + export Cypher/parquet
PYTHONPATH=. $PY -m spid embed                               # 5. embeddings MiniLM
PYTHONPATH=. $PY -m spid analytics | stats                   # tableaux de bord agrégés
PYTHONPATH=. $PY -m spid search "data center" --type data_center --min-amount 5e8
PYTHONPATH=. $PY -m spid semantic "AI projects launched by banks"
```

```
PYTHONPATH=. $PY -m spid all --backend llm                    # tout le pipeline
uvicorn spid.api:app --port 8088                              # API HTTP : /projects /semantic
/analytics
```

La configuration passe par `project-db/.env` : `OPENAI_API_KEY`, `SPID_LLM_BASE_URL`, `SPID_LLM_MODEL` (défaut `gpt-4o-mini`), `SPID_EMBED_MODEL`, `SPID_BACKEND`. L'interface étant compatible OpenAI, deux variables suffisent pour viser la passerelle du cluster de calcul de l'UQO (`llm.maclustr.io`) ou tout modèle local.

3.7 Coût, débit et reproductibilité

- **Coût** : de l'ordre du dixième de cent US par section grâce au *gate* et au *focus* ; l'exécution complète est estimée entre 100 et 130 $ US de `gpt-4o-mini`.
- **Débit** : appels LLM parallèles (16 *workers*) ; résolution, graphe et analytique sont du SQL DuckDB pur et passent à l'échelle de millions de mentions.
- **Reprenabilité** : identifiants stables, `INSERT OR REPLACE`, journal des CIK terminés ; une interruption ne re-facture jamais le travail déjà fait.

4 Schéma de la base de données

La base est un unique fichier DuckDB (`data/db/spid.duckdb`, 3,2 Go, dernière écriture le 28 octobre 2026). Elle contient huit tables physiques (dont le registre `extracted_sections` ajouté en octobre 2026), trois vues sur le corpus amont et une vue unifiée.

Table 6. Tables persistées dans `spid.duckdb`.

Table	Lignes	Colonnes
`spid_sections`	27 164	section_id, accession_number, section_name, item_number, text, word_count
`project_mentions`	308 065	mention_id, accession_number, cik, section_id, section_name, project_name, project_type, description, objective, amount_usd, amount_raw, locations[], technologies[], partners[], suppliers[], benefits[], risks[], years[], status, confidence, backend, ticker, company_name, form_type, filing_date, fiscal_year, sector
`projects`	121 099	project_id, cik, ticker, company_name, sector, project_type, project_name, canonical_location, technologies[], status, total_amount_usd, n_mentions, n_filings, first_seen, last_seen, first_year, last_year, avg_confidence, description
`project_timeline`	308 065	project_id, accession_number, filing_date, form_type, status, amount_usd, confidence, snippet
`kg_nodes`	183 472	node_id, node_type, label, props (JSON)
`kg_edges`	204 456	src, rel, dst, props (JSON)
`project_embeddings`	121 099	project_id, model, dim, vector (FLOAT[384])

Table 7. Vues définies dans `spid.duckdb`.

Vue	Définition
`src_filings`	`read_parquet(…/parquet/filings/filings.parquet)`
`src_sections`	`read_parquet(…/parquet/filing_sections/filing_sections.parquet)`
`src_companies`	`read_parquet(…/parquet/companies/companies.parquet)`
`v_sections`	`src_sections` ∪ `spid_sections`, jointes à `src_filings` (cik, ticker, formulaire, date, exercice) et `src_companies` (secteur, sous-industrie, SIC, état)

Point d'attention : chemins des vues

Les trois vues `src_*` pointent en dur vers le corpus parquet du nœud maître du cluster (`~/UQO_RESEARCH_DB/data/parquet/`… sur M3U96b, où le pipeline s'exécute désormais). Sur toute autre machine (copie du laptop, serveur `pdb-api`), les huit tables physiques restent parfaitement lisibles, mais toute requête sur `v_sections` ou `src_filings` échoue tant que les vues ne sont pas recréées avec le chemin local du corpus (`Store()` les recrée automatiquement à l'ouverture en écriture). Par ailleurs, la colonne `fiscal_year` de `project_mentions` est vide pour les mentions du run de juin 2026 (elle est renseignée depuis octobre 2026) : les analyses temporelles s'appuient sur `filing_date`.

5 Portrait statistique des projets

5.1 Couverture temporelle et formulaires

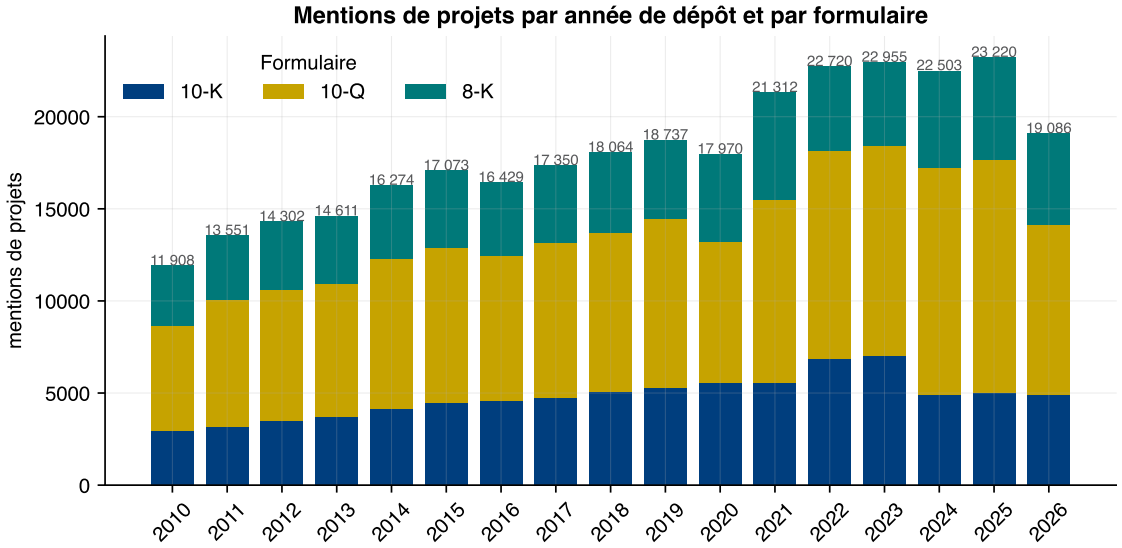


Figure 4. Mentions de projets par année de dépôt et par formulaire (74 566 en 8-K, 81 495 en 10-K, 152 004 en 10-Q). Le flux est stable, entre 2 000 et 2 800 mentions par an ; 2026 est une année partielle (dépôts jusqu'au 26 mai).

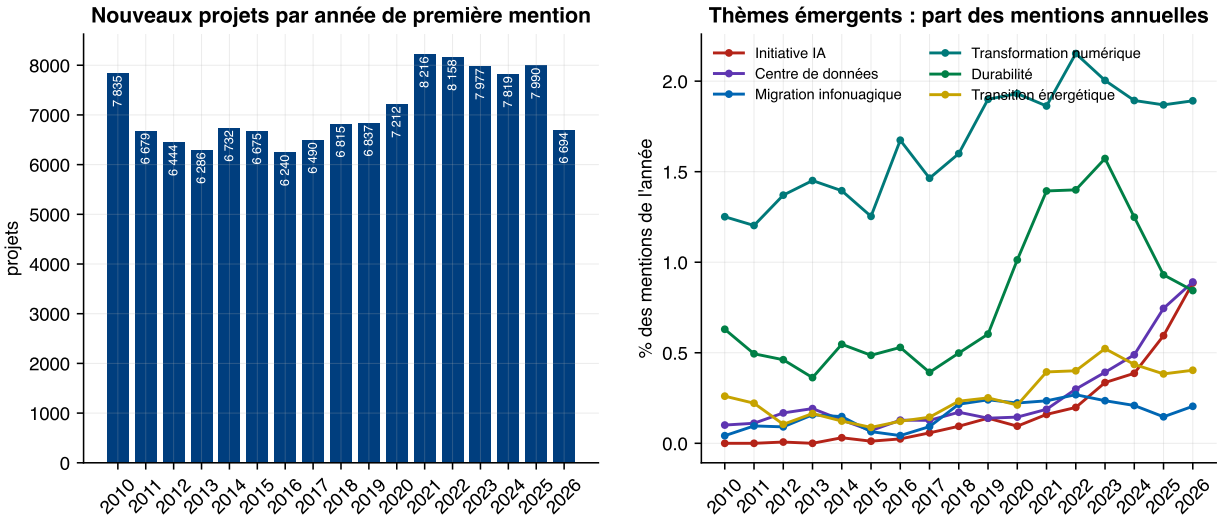


Figure 5. *Gauche* : nouveaux projets par année de première mention. *Droite* : part des mentions annuelles consacrée à six thèmes émergents. La transformation numérique décolle à partir de 2018, la durabilité culmine en 2023 et les initiatives d'IA passent de moins de 0,3 % des mentions avant 2022 à 1,7 % en 2025–2026.

5.2 Types de projets

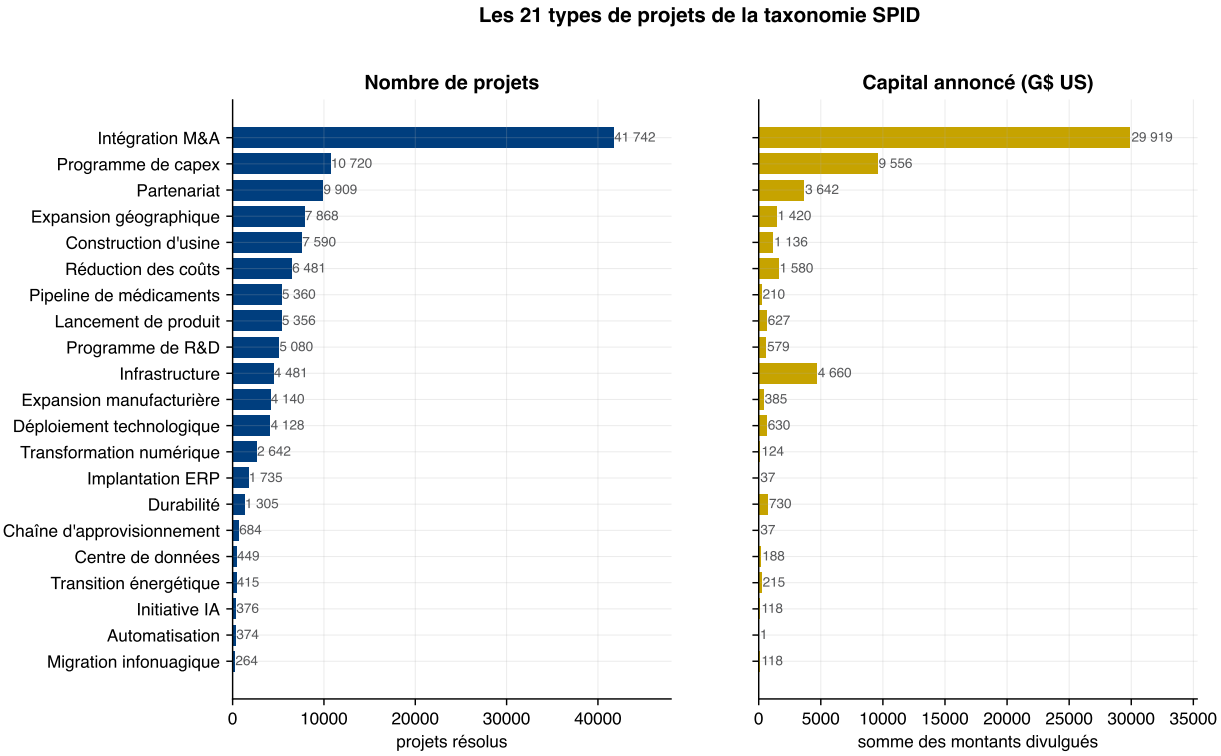


Figure 6. Répartition des 121 099 projets selon les 21 types de la taxonomie et capital annoncé par type. L'intégration post-M&A domine en nombre (38 %) comme en capital, suivie des programmes de capex.

Table 8. Les 21 types de projets : nombre de projets, capital annoncé cumulé (G$ US) et montant médian par projet (M$ US, projets avec montant).

project_type	Libellé	Projets	Capital (G$)	Médiane (M$)
ma_integration	Intégration M&A	41 742	29 918,9	155
capex_program	Programme de capex	10 720	9 555,8	300
partnership	Partenariat	9 909	3 641,6	75
geographic_expansion	Expansion géographique	7 868	1 419,7	37
plant_construction	Construction d'usine	7 590	1 135,8	56
cost_reduction	Réduction des coûts	6 481	1 580,2	50
drug_pipeline	Pipeline de médicaments	5 360	209,6	46
product_launch	Lancement de produit	5 356	627,5	27
rd_program	Programme de R&D	5 080	578,8	20
infrastructure	Infrastructure	4 481	4 660,2	48
manufacturing_expansion	Expansion manufacturière	4 140	385,3	45
technology_deployment	Déploiement technologique	4 128	630,1	135
digital_transformation	Transformation numérique	2 642	123,9	34
erp_implementation	Implantation ERP	1 735	36,9	15
sustainability	Durabilité	1 305	729,8	34
supply_chain	Chaîne d'approvisionnement	684	37,4	38
data_center	Centre de données	449	187,8	166
energy_transition	Transition énergétique	415	214,7	109
ai_initiative	Initiative IA	376	118,1	128
automation	Automatisation	374	0,8	14
cloud_migration	Migration infonuagique	264	117,8	50

5.3 Secteurs

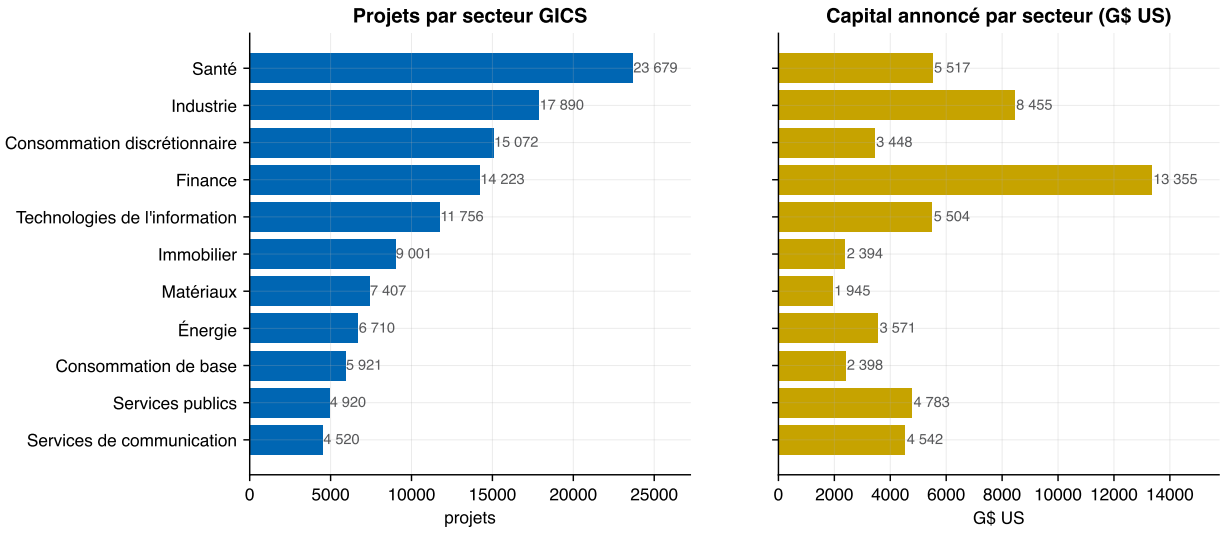


Figure 7. Projets et capital annoncé par secteur GICS. L'industrie, la santé et les technologies de l'information comptent le plus de projets ; la finance concentre le plus de capital annoncé, du fait des grandes acquisitions bancaires.

Table 9. Répartition par secteur GICS : nombre d'entreprises, de projets, projets par entreprise et capital annoncé.

Secteur	Entreprises	Projets	Projets / entr.	Capital (G$)
Santé	429	23 679	55,2	5 516,7
Industrie	386	17 890	46,3	8 454,8
Consommation discrétionnaire	267	15 072	56,4	3 448,4
Finance	446	14 223	31,9	13 355,0
Technologies de l'information	325	11 756	36,2	5 503,8
Immobilier	145	9 001	62,1	2 394,0
Matériaux	117	7 407	63,3	1 944,8
Énergie	107	6 710	62,7	3 570,6
Consommation de base	103	5 921	57,5	2 398,1
Services publics	66	4 920	74,5	4 783,0
Services de communication	101	4 520	44,8	4 541,7

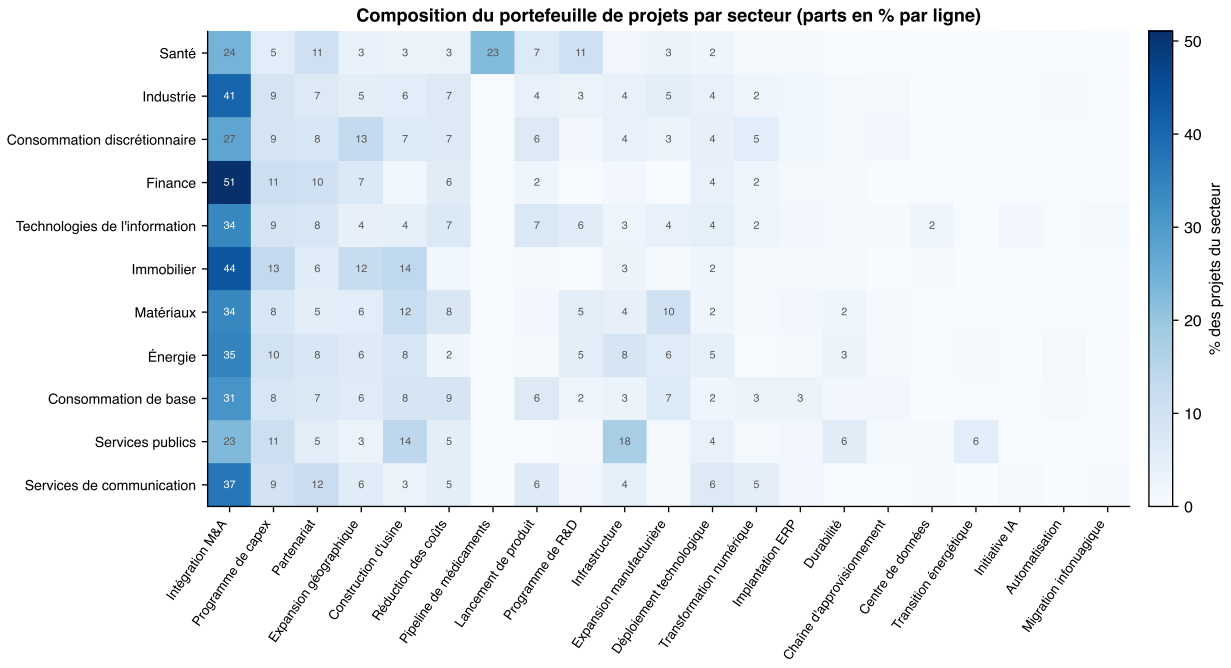


Figure 8. Composition du portefeuille de projets par secteur (parts en pourcentage de chaque ligne). Les secteurs se spécialisent : services publics dans l'infrastructure et la construction d'usines (centrales), santé dans les pipelines de médicaments et lancements de produits, consommation discrétionnaire dans l'expansion géographique, finance et immobilier dans les acquisitions.

5.4 Statut et cycle de vie

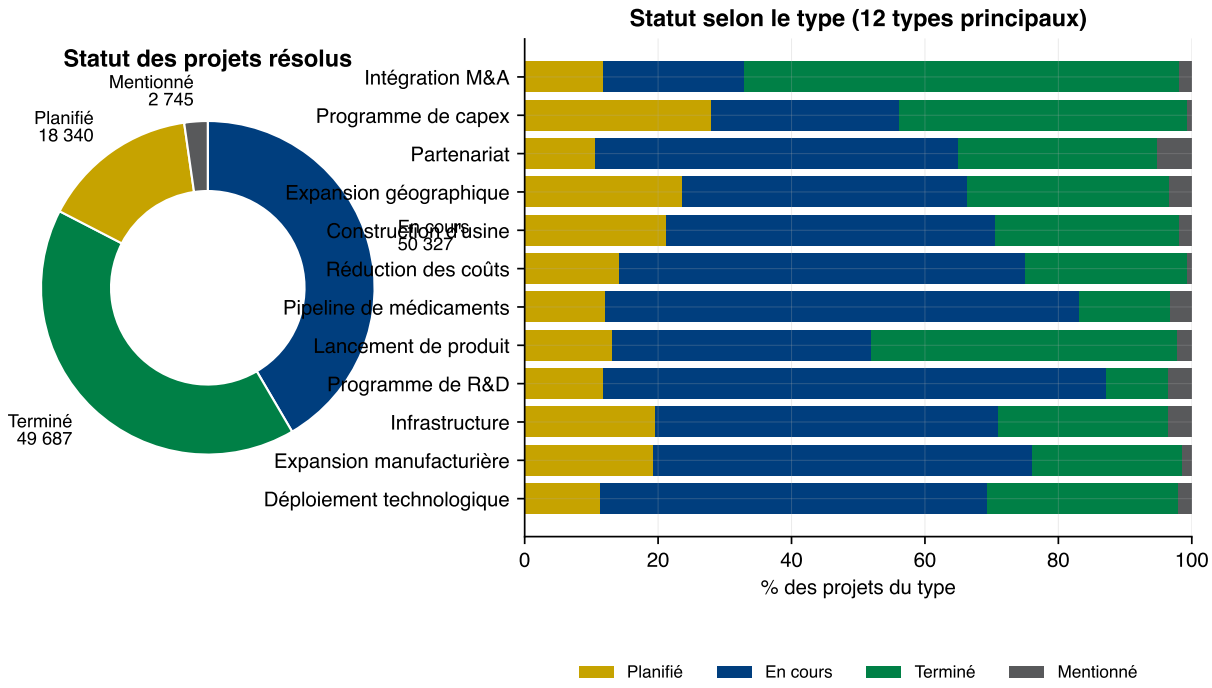


Figure 9. *Gauche* : statut courant des projets résolus. *Droite* : composition par statut des douze types principaux. Les programmes de capex et les intégrations M&A sont les plus souvent déclarés terminés ; les programmes de R&D et la transformation numérique restent majoritairement « en cours ».

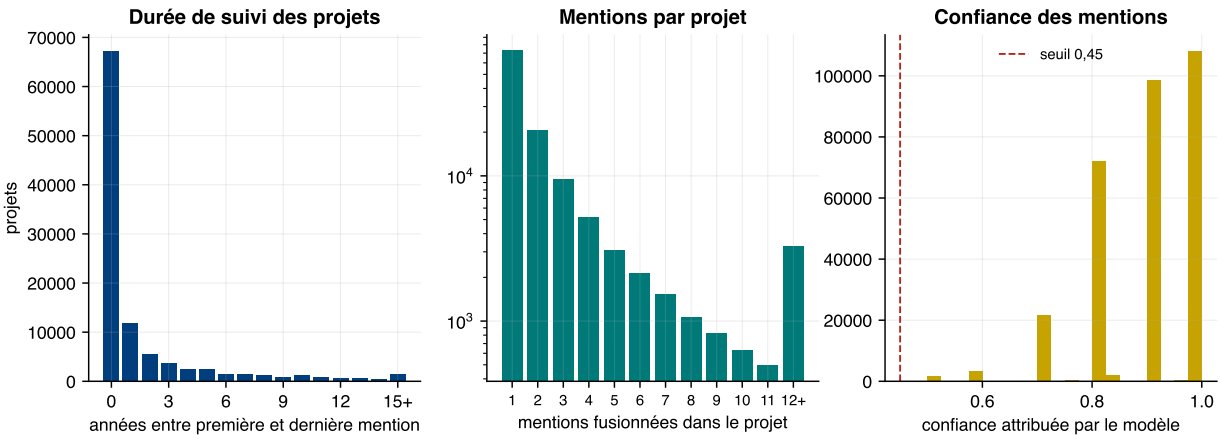


Figure 10. *Gauche* : durée de suivi (écart entre première et dernière année citée) ; 64,9 % des projets ne sont cités que sur une année, 12,4 % sur cinq ans ou plus. *Centre* : nombre de mentions fusionnées par projet (échelle log). *Droite* : confiance attribuée par le modèle à chaque mention (moyenne 0,89 ; 91,3 % des mentions à 0,80 ou plus).

Table 10. Les dix projets les plus suivis (nombre de mentions et de filings distincts). Les années sont celles citées dans le texte des filings et peuvent inclure des horizons projetés.

Ticker	Projet	Type	Statut	Années	Ment.	Filings
MIDD	Acquisition of TurboChef, CookTek, Anets, and …	Intégration M&A	Terminé	2007–2024	269	45
AME	Acquisition of Ameron Global	Intégration M&A	Terminé	2009–2026	216	98
USPH	Acquisition of Majority Interest in Physical T…	Intégration M&A	Terminé	2008–2026	212	102
NWS	Sale of News America Marketing	Intégration M&A	En cours	2020–2026	209	209
POST	Acquisition of Hearthside Food Solutions	Intégration M&A	Terminé	2012–2025	191	85
TDG	Acquisition of Jet Parts Engineering and Victo…	Intégration M&A	Terminé	2009–2033	162	87
BSX	Acquisition of Sadra Medical, Inc.	Intégration M&A	Planifié	2010–2026	147	73
RGNX	RGX-111	Pipeline de médicaments	En cours	2015–2026	145	44
AXSM	AXS-05 Development	Pipeline de médicaments	En cours	2016–2023	143	37
SWK	Acquisition of CRC-Evans International, Inc.	Intégration M&A	Planifié	2010–2029	142	51

5.5 Montants divulgués

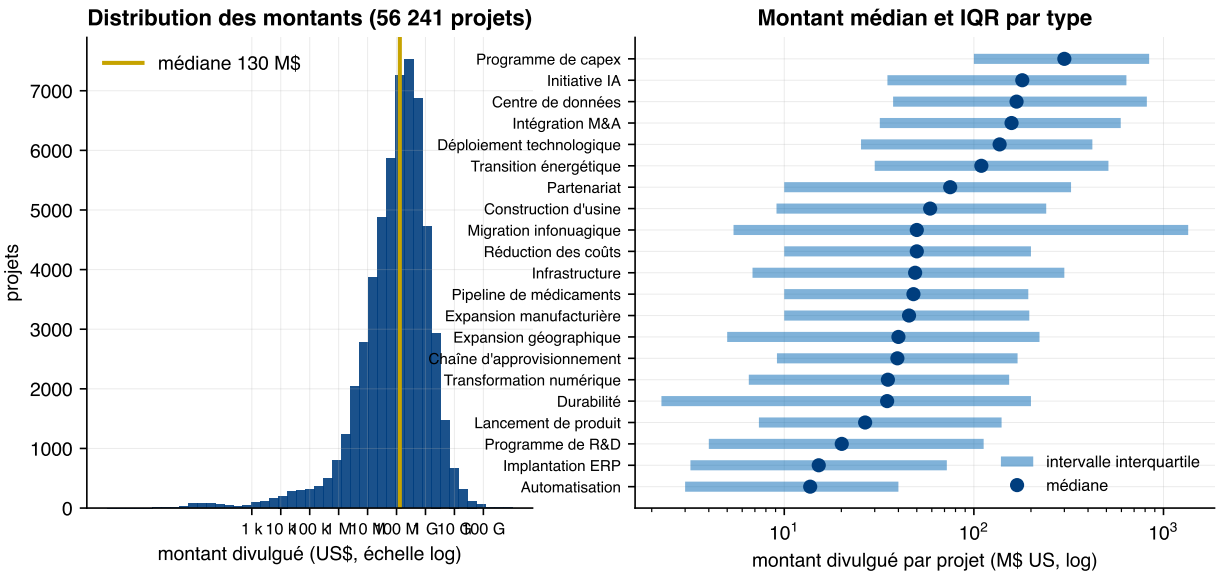


Figure 11. *Gauche* : distribution des montants divulgués (56 241 projets, échelle logarithmique) ; médiane 130 M$, moyenne 994 M$. *Droite* : médiane et intervalle interquartile par type. Les intégrations M&A et les programmes de capex sont d'un ordre de grandeur supérieur aux lancements de produits ou aux implantations ERP.

Table 11. Les quinze projets au montant divulgué le plus élevé (au moins trois mentions, hors programmes de capex). Le montant retenu est le maximum observé dans les mentions du projet ; plusieurs valeurs de cette liste sont manifestement des erreurs d'extraction (valeur d'entreprise ou chiffre d'affaires attribué à une transaction), voir la note ci-dessous.

Ticker	Projet	Type	Capital (G$)	Années	Ment.
STT	Acquisition of Goldman Sachs Administration Services	Intégration M&A	200,0	2012–2022	18
GE	Sale of NBCUniversal Interest	Intégration M&A	157,0	2013–2020	8
VZ	Wireless Transaction	Intégration M&A	130,0	2013–2020	4
VZ	Acquisition of Terremark Worldwide, Inc.	Intégration M&A	130,0	2011–2026	53
ONIT	Acquisition of Existing Servicing Platforms	Intégration M&A	119,3	2009–2026	50
IVZ	Acquisition of Guggenheim Investments' ETF Business	Intégration M&A	114,6	2010–2019	25
PMT	Correspondent Lending Business	Intégration M&A	114,5	2012–2019	12
FCNCA	Merger with Entegra Financial Corp.	Intégration M&A	110,1	2019–2023	3
LNN	Infrastructure Investment and Jobs Act	Infrastructure	110,0	2021–2026	5
FCNCA	SVBB Acquisition	Intégration M&A	107,5	2023–2023	6
RITM	Purchase of Mortgage Servicing Rights	Intégration M&A	97,0	2014–2017	3
AVGO	Acquisition of Brocade Communications Systems, Inc.	Intégration M&A	86,3	2017–2024	43
TRIP	Vacation Rentals Expansion	Expansion géographique	85,0	2011–2014	9
CHTR	TWC Transaction	Intégration M&A	85,0	2015–2016	11
AVGO	VMware Acquisition	Intégration M&A	83,8	2023–2025	6

Comment lire le « capital annoncé »

Le montant d'un projet (`total_amount_usd`) est le *maximum* des montants divulgués dans ses mentions ; le « capital annoncé » d'un type ou d'un secteur est la somme de ces maxima sur les projets concernés. Ces montants sont auto-déclarés, de portée hétérogène (coût total d'un programme, dépense annuelle, valeur d'une transaction, parfois un chiffre d'entreprise capté par erreur) et un même programme pluriannuel peut apparaître sous plusieurs projets si son nom ou sa localisation varie. Le total de 55 912 G$ sur l'ensemble de la base, comme les valeurs par type ou par secteur, se lit donc comme un **ordre de grandeur supérieur**, non comme un investissement net. Pour une analyse économétrique, les auteurs recommandent de travailler sur le montant *médian* des mentions d'un projet (table `project_timeline`) et d'écarter les valeurs extrêmes.

5.6 Entreprises

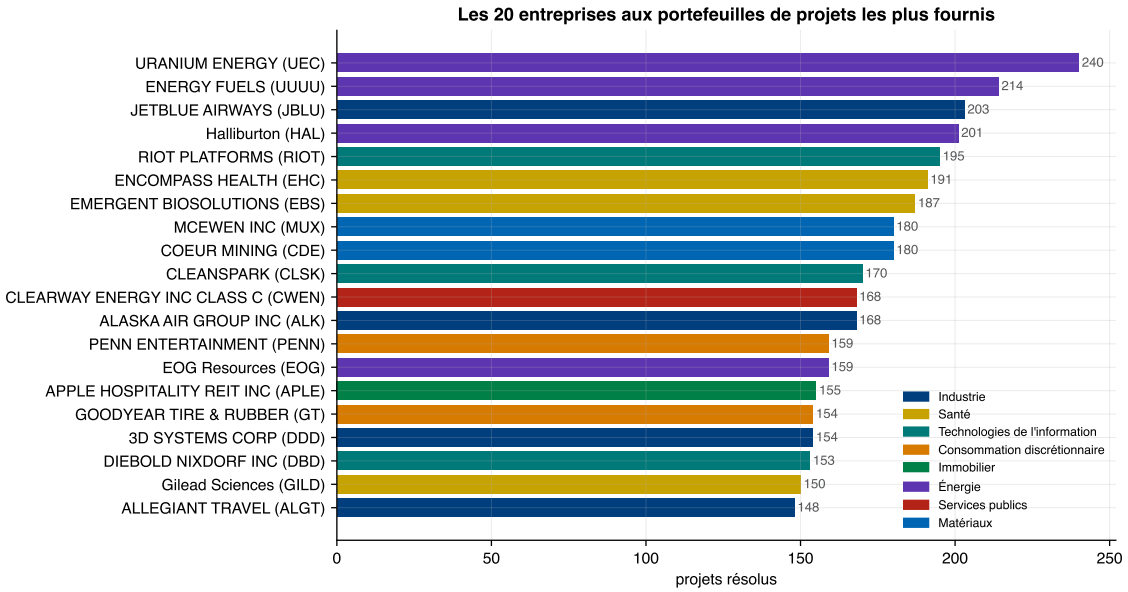


Figure 12. Les vingt entreprises aux portefeuilles de projets les plus fournis, colorées par secteur. Chaque entreprise compte en médiane 42 projets (moyenne 48,6, maximum 240). Les services publics (WEC, PPL, Duke, Xcel, PG&E) sont surreprésentés : leurs plans d'investissement réglementés sont décrits en détail dans chaque filing.

Table 12. Les quinze entreprises comptant le plus de projets résolus.

Ticker	Entreprise	Secteur	Projets	Capital (G$)
UEC	URANIUM ENERGY	Énergie	240	19,8
UUUU	ENERGY FUELS	Énergie	214	3,0
JBLU	JETBLUE AIRWAYS	Industrie	203	35,9
HAL	Halliburton	Énergie	201	33,3
RIOT	RIOT PLATFORMS	Technologies de l'information	195	4,6
EHC	ENCOMPASS HEALTH	Santé	191	7,1
EBS	EMERGENT BIOSOLUTIONS	Santé	187	24,2
MUX	MCEWEN INC	Matériaux	180	11,5
CDE	COEUR MINING	Matériaux	180	20,4
CLSK	CLEANSPARK	Technologies de l'information	170	7,2
CWEN	CLEARWAY ENERGY INC CLASS C	Services publics	168	20,2
ALK	ALASKA AIR GROUP INC	Industrie	168	31,8
PENN	PENN ENTERTAINMENT	Consommation discrétionnaire	159	21,7
EOG	EOG Resources	Énergie	159	91,3
APLE	APPLE HOSPITALITY REIT INC	Immobilier	155	17,7

5.7 Localisations, technologies et partenaires

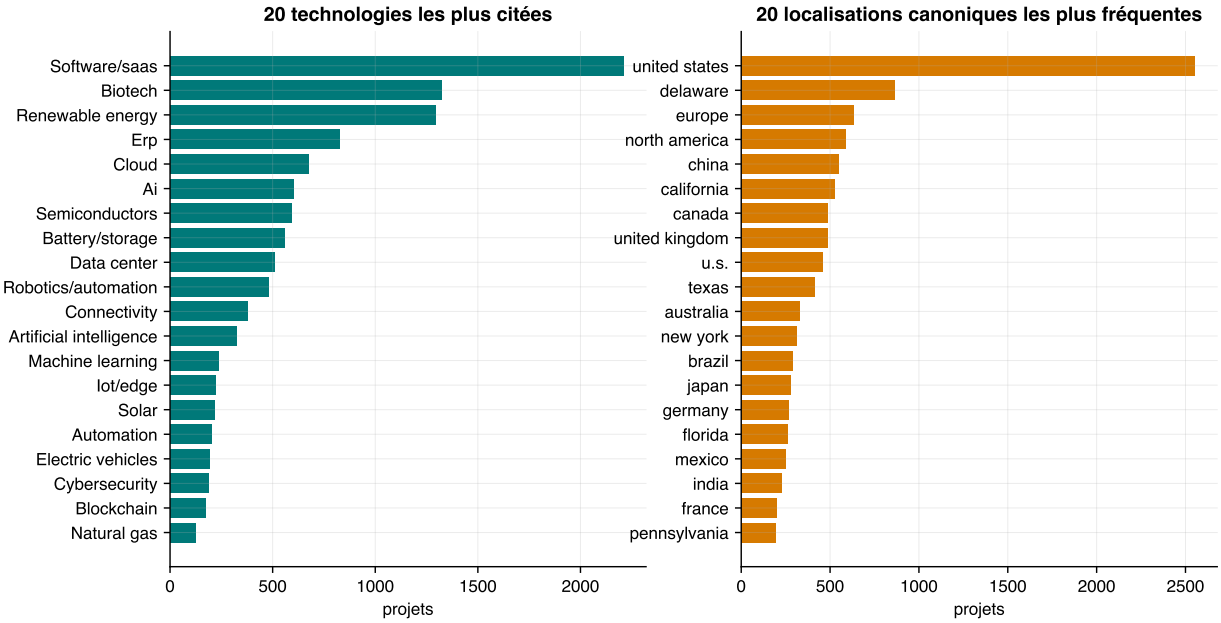


Figure 13. *Gauche* : les vingt technologies les plus citées (14,6 % des projets en portent au moins une ; 8 311 libellés distincts). *Droite* : les vingt localisations canoniques les plus fréquentes (45,1 % des projets sont localisés ; 14 508 localisations distinctes).

6 Graphe de connaissances et espace sémantique

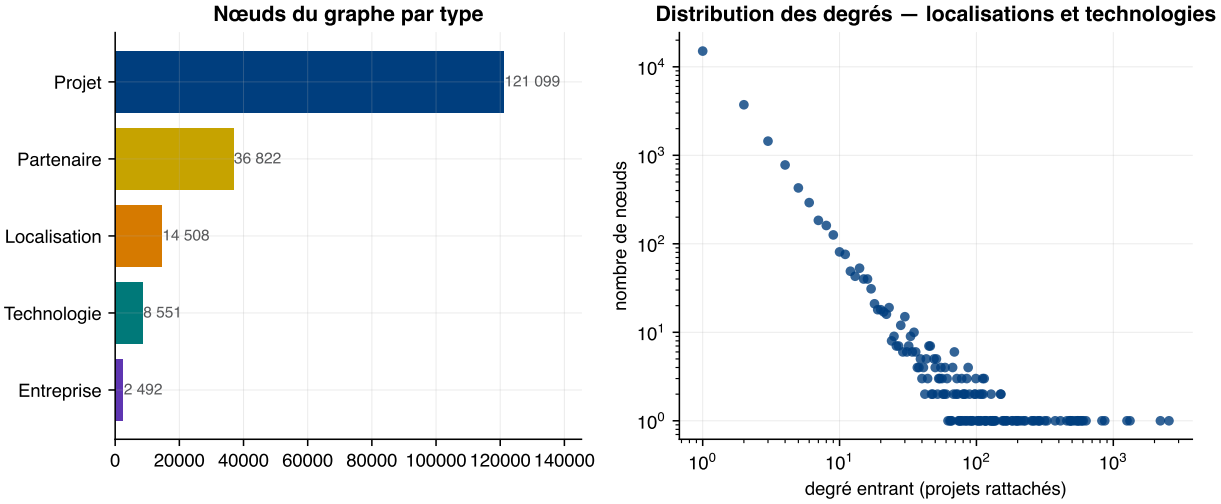


Figure 14. *Gauche* : composition du graphe (183 472 nœuds ; arêtes : 121 099 `owns`, 54 583 `located_in`, 28 774 `uses`). *Droite* : distribution des degrés entrants des localisations et technologies (log-log) ; la queue lourde traduit quelques hubs (États-Unis, Texas, Chine ; IA, énergie renouvelable, logiciel) reliés à des centaines de projets.

Le graphe permet des requêtes impossibles sur des tables plates : « quelles entreprises construisent au Texas des installations utilisant des batteries ? », « quels partenaires reviennent dans les projets d'IA du secteur financier ? ». L'export Cypher (`data/exports/kg_import.cypher`) charge le graphe dans Neo4j en

une commande. Les embeddings complètent le graphe par une recherche *par le sens* : deux projets décrits avec des mots différents mais un même objet (une « gigafactory » et une « usine de cellules de batteries ») se retrouvent voisins dans l'espace à 384 dimensions. Le projet de recherche `research/` exploite ces deux couches (détection de communautés technologiques, regroupement K-means et projection t-SNE des projets).

7 Qualité des données et limites

- **Divulgation, non réalité.** SPID mesure ce que les entreprises *disent* de leurs projets dans leurs filings. Une initiative jamais évoquée est invisible, et les incitations à divulguer varient selon les secteurs (les services publics réglementés décrivent tout, les sociétés technologiques peu).
- **Montants.** Auto-déclarés, hétérogènes en portée et absents pour la moitié des projets (voir la note de la section 5). Quelques valeurs aberrantes subsistent (montant d'un rachat d'actions ou d'une valeur d'entreprise attribué à un « projet »).
- **Échantillon non historique.** Les listes S&P 500 (mai 2026) et Russell 3000 (positions IWV d'octobre 2026) sont des instantanés : les sociétés sorties des indices entre 2010 et 2026 (faillites, acquisitions, déclassements) sont absentes, ce qui introduit un biais de survie.
- **Tableaux des 10-K supprimés.** L'ingestion SPID retire les tableaux HTML ; les montants présentés en tableau (plans de capex chiffrés, par exemple) ne sont captés que s'ils sont repris dans la prose.
- **Extraction par LLM.** Faux positifs (formules génériques prises pour des projets) et faux négatifs (projets décrits sans vocabulaire de projet) existent. Le seuil de confiance, l'ancrage textuel obligatoire et la température nulle les réduisent, mais aucun échantillon étiqueté à la main d'une taille suffisante ne permet encore d'estimer précision et rappel.
- **Résolution heuristique.** Le regroupement par entreprise, type et nom normalisé peut scinder un même projet nommé différemment d'un filing à l'autre (surestimation du nombre de projets) ou fusionner deux projets homonymes (sous-estimation). Le nom d'un projet peut aussi être une reformulation du modèle plutôt qu'un nom officiel.
- **Années citées.** `first_year` et `last_year` proviennent des années mentionnées dans le texte : elles peuvent précéder la date du dépôt (rappel historique) ou lui succéder (horizon projeté, jusqu'en 2033 ou 2040). Pour dater l'observation, préférer `first_seen` et `last_seen`.
- **Couverture des 10-Q.** Depuis l'extension d'octobre 2026, les 10-Q de 2 494 entreprises sont segmentés et extraits (578 793 sections, 52 % des sections éligibles) ; le run initial de juin 2026 n'en couvrait que 84.
- **Échantillon.** 2 564 sociétés cotées aux États-Unis, des grandes aux petites capitalisations ; les émetteurs étrangers déposant des 20-F/40-F (65 sociétés de l'ETF), les sociétés privées et les micro-capitalisations hors ETF restent hors champ.
- **Vues cassées.** Les vues `src_*` doivent être recréées avec le nouveau chemin du corpus avant toute ré-exécution du pipeline (section 4).

8 Reproduire ce rapport

docs/rapport_uqo/

```
python3 make_figures.py          # lit data/db/spid.duckdb, écrit figures/*.pdf et tables/*.tex
latexmk -lualatex rapport_spid.tex
```

Toutes les valeurs numériques du texte (macros `\nProjets`, `\medianeMontant`, etc.) et tous les tableaux sont régénérés depuis la base par le script ; aucun chiffre n'est saisi à la main.

A Taxonomie des 21 types de projets

La taxonomie (`spid/taxonomy.py`) est un vocabulaire contrôlé de 21 catégories mutuellement exclusives. Chaque catégorie est dotée d'une description transmise au modèle et de mots-clés servant au *gate*, au backend déterministe et à la correspondance floue du post-traitement.

Type	Définition retenue
`manufacturing_expansion`	Extension de la capacité de production d'un site existant (lignes, équipements, quarts).
`plant_construction`	Construction d'une nouvelle usine, centrale, raffinerie ou installation industrielle.
`data_center`	Construction, extension ou location de centres de données et d'infrastructures de calcul.
`cloud_migration`	Migration de systèmes ou de charges de travail vers l'infonuagique.
`ai_initiative`	Programme d'intelligence artificielle ou d'apprentissage automatique (produits, plateformes, modèles).
`digital_transformation`	Modernisation numérique transversale (plateformes, canaux, processus).
`erp_implementation`	Implantation ou remplacement d'un système ERP ou de gestion intégrée.
`rd_program`	Programme de recherche et développement identifié (hors médicaments).
`product_launch`	Lancement commercial d'un produit ou service nouveau.
`drug_pipeline`	Candidat-médicament, essai clinique ou dépôt réglementaire.
`ma_integration`	Acquisition, cession, fusion ou intégration post-transaction.
`geographic_expansion`	Entrée sur un nouveau marché géographique, ouverture de magasins ou de bureaux.
`capex_program`	Programme d'investissement pluriannuel chiffré (plan de capex).
`supply_chain`	Réorganisation de la chaîne d'approvisionnement, logistique, entrepôts.
`automation`	Automatisation industrielle ou robotique de processus.
`sustainability`	Programme environnemental ou de durabilité (émissions, déchets, eau).
`energy_transition`	Production renouvelable, stockage, décarbonation, électrification.
`infrastructure`	Réseaux, pipelines, lignes de transport, infrastructures lourdes.
`cost_reduction`	Programme de réduction des coûts, restructuration, fermetures.
`partnership`	Partenariat stratégique, coentreprise, alliance, contrat majeur.
`technology_deployment`	Déploiement d'une technologie identifiée ne relevant d'aucune autre catégorie (type par défaut).

B Fichiers du dépôt `project-db/`

Chemin	Rôle
`spid/config.py`	Chargement de `.env`, chemins, journalisation.
`spid/store.py`	Connexion DuckDB, création du schéma et des vues, primitives d'écriture.
`spid/taxonomy.py`	Les 21 types, descriptions, mots-clés, correspondance floue.
`spid/ingest.py`	Nettoyage et sectionnement des 10-K bruts.
`spid/extract/base.py, llm.py, rules.py`	Interface d'extraction, backend LLM (prompt, gate, focus), backend déterministe.
`spid/pipeline.py`	Moteur d'extraction : sélection des sections, parallélisme, écriture des mentions.
`spid/resolve.py`	Fusion des mentions en projets et chronologies.
`spid/graph.py`	Graphe de connaissances et exports parquet/Cypher.
`spid/embed.py`, `spid/search.py`	Embeddings MiniLM, recherche sémantique et filtrée.
`spid/analytics.py`, `spid/api.py`, `spid/__main__.py`	Agrégats, API FastAPI, interface en ligne de commande.
`run_full_extract.py`	Pilote d'extraction par lots de 15 CIK avec points de reprise.
`data/db/spid.duckdb`	La base (3,2 Go).
`data/exports/`	`kg_nodes.parquet`, `kg_edges.parquet`, `kg_import.cypher`.
`logs/`	`ingest.log`, `extract.log`, `post.log`, `extract_done_ciks.txt`.
`research/`	Projet de recherche aval : scripts 01–10, figures, tableaux et article (`paper/main.pdf`).
`docs/`	Documentation, dont ce rapport (`docs/rapport_uqo/`).