

Enthic

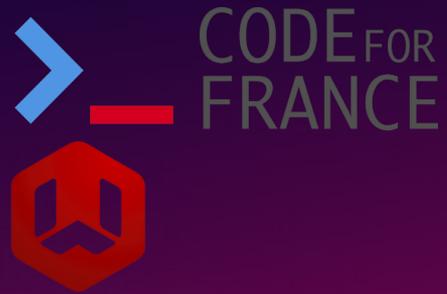
Portail des données financières des entreprises
Françaises

Christophe Brun, PapIT
Grenoble le 15/10/2020



Génèse

- L'association Code For France
- L'association Wexample



Demande d'aide liée à la taille du jeu de donnée. 4.5Go zippé pour les 3 premières années. L'insertion des données en base est lente et le site également. Seul quelques centaines de milliers d'entreprises sont en base.

Investigation du jeu de donnée

- Maintenu par l'INPI
- Contient des bilans comptables, (salaires, participation, taxes et impôts, nombre de salariés, chiffre d'affaires, etc)
- Un zip par jour contenant tous les bilans déposés ce jour et son md5 eux-même dans un zip
- Le bilan est un XML et l'INPI fournit un XSD pour vérifier le format du fichier 😊
- Documentation technique expliquant le contenu du XML 😊
- Impossible de trouver un bilan sans un algorithme 😭

Base de donnée Open Source

- EU Open Data Directive launched in 2017 :

Towards a list of High-Value datasets.

Stimulate the publishing of dynamic data and the uptake of Application Programme Interfaces (APIs). 🥲

In France consolidated version January 2017

- Participation de l'INPI :

Cette ouverture des données s'inscrit dans le cadre d'une politique gouvernementale volontariste, dont l'objectif est de favoriser l'émergence de services innovants à valeur ajoutée pour l'économie. Date de 2017.

Base de donnée Open Source, les sources

- L'INPI avec les données RNCS, propriété intellectuelle
- www.data.gouv.fr Plateforme ouverte des données publiques françaises.

Base de donnée Open Source, exemple du FINESS

- Ce jeu de données provient d'un service public certifié, liste des établissements du domaine sanitaire et social.
- Pas d'API, il faut la développer.
 - CSV (donc pas d'outil de validation de XSD)
 - Pas de header, on ne sait pas à quoi correspond les colonnes, XML a toujours un tag.
 - Avenue du 11 novembre, transformé (par une ouverture dans Excel probablement) en 11-nov
 - Numéro supposé de téléphone vide ou 0, que veut dire 0.
 - Date d'autorisation, peu probable 0201-12-26, 0980-11-12, 0192-08-01 (Peu faire planter un programme)
- Réaction de la communauté :
 - *FAQ : Pourquoi les colonnes n'apparaissent pas sur le fichier proposé ?*
 - *Pour permettre l'utilisation d'automates, la structure du fichier s'affiche en format XSD. Pour retrouver les colonnes d'un format type tableur, il faut vous référer à la description du fichier disponible en format PDF.*
 - *Tam Kien Duong : Code python pour obtenir un CSV propre, rajoute des entêtes enlève une colonne inutile sauvegarde au format utf-8*

Objectifs

- Distribuer de manière simple et gratuite des données financières sur les entreprises Françaises
 - API aux standards Swagger pour la description et Hydra (partiel) pour le contenu.
 - Accès complet à la base pour investiguer et extraire des données absentes de l'API
- Communiquer sur le potentiel de ce jeu de donnée pour créer des usages (Application type Yuka pour les particuliers, BI, etc)

Moyens

- Motivation
- Des compétences en développement
- Le jeu de donnée ouvert par l'INPI sur les bilans comptables depuis 2017

A éviter

- Utiliser le système de fichier pour écrire ou lire sur le disque dure les XML (désavantage par rapport au CSV).
- Charger toutes les données en base.
 - Recherche d'*Enum* pour transformer les chaînes de caractère en types numériques
 - Filtrer les données, incorrectes ou que l'on ne sait pas exploiter
- Modification en base de données



Batch pour le traitement des données et web pour la distribution à décorréler

Développement du schéma de la base

- Conditionne le développement du batch
- Plusieurs bilans par an mais un ligne souhaitée



identity

siren	INT
denomination	VARCHAR(100)
ape	SMALLINT
postal_code	CHAR(5)
town	VARCHAR(25)

bundle

siren	INT
declaration	YEAR
accountability	TINYINT
bundle	TINYINT
amount	FLOAT

request

uri	VARCHAR(100)
agent	VARCHAR(200)
parameter	VARCHAR(100)
created	TIMESTAMP

Contraintes liées au développement

- Peu de moyens humain et financier et peu de temps
 - Technologies open source, Linux, Python, MySQL
 - Low Code, utilisation de programmes existants et langages de haut niveau.

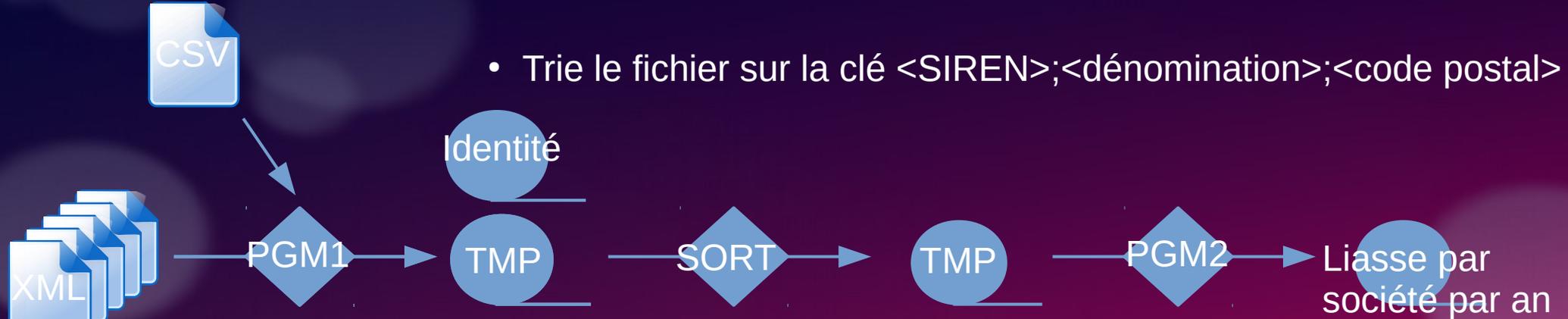
Format du XML

1 Bilan comptable = 1 XML



Batch

- Paramétrage des liasses à extraire



- Trie le fichier sur la clé <SIREN>;<dénomination>;<code postal>

- Créer deux CSV
- Filtrer les données
- Nettoie une fraction des erreurs de saisie

- Trie le fichier sur la clé <SIREN>;<année>;<liasse>

- Somme les montants d'une même clé <SIREN>;<année>;<liasse>

Programme 1

- Lit le fichier de paramétrage avec les liasses
- Extrait les identités des entreprises
- Extrait les liasses
- Modifie ou rejette certaines erreurs de saisie
- Transforme tous les Enum en type MySQL numérique économe en taille, car les Enum n'existent pas en MySQL. Comparaison de types numériques est souvent plus rapide que les chaînes de caractère. Les durées des requêtes augmentent avec taille de la base.
- Le programme est plus lent avec Pypy que CPython. 

SORT et programme 2

- SORT, trie les fichiers avec séparateur ou de longueur fixe.
 - Tri est rapide car lie séquentiellement, évite une recherche dans un fichier ou un stockage en mémoire

000000001	2017	7	3000
000000001	2017	7	500
000000001	2018	9	4000
000000002	2017	8	500

- Programme 2, fait la somme pour une même
<SIREN>;<année>;<liasse>

000000001	2017	7	3500
000000001	2018	9	4000
000000002	2017	8	500

2 fichiers identiques au tables

- 2 CSV, un pour l'identité, l'autre pour les liasses fiscales.
- Ils peuvent être insérés directement par MySQL en ligne de commande (low code).
- Pas de traitement ultérieur, la base est 100 % disponible pour la lecture.

Table de stockage des requêtes entrantes

- Les paramètres des requêtes (SIREN, dénomination, etc.) sont stockés
- Aucune IP pour des questions de vie privée
- Permettra éventuellement de savoir ce qui intéresse le plus, pour faire évoluer l'application

Serveur Web, les standards

- API JSON définit par un fichier Swagger avec son front end swagger <https://api.enthic.fr/> . La spécification OpenAPI (<https://swagger.io/>) .
- JSON (partiellement) au format JSON-LD (<https://json-ld.org/>) (JavaScript Object Notation for Linked Data) quand c'est possible.

Serveur Web, le framework Flask

- Parfois catégorisé comme microframework
- *Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications.*
<https://palletsprojects.com/p/flask/>

Serveur Web, recherche de patterns pour les réponses

- Chaque pattern donne une classe.
 - Deux types d'identité, avec ou sans SIREN/dénomination
 - Deux types de liasses, multiples (liste d'année) ou simple (moyenne ou année)
- Indirectement hérite de la classe *Response* de Flask

```
# http://api.enthic.fr/company/denomination/thales/2017
@application.route("/company/denomination/<string:denomination>/<string:year>", methods=['GET'], strict_slashes=False)
@insert_request
def company_denomination_year(denomination, year):
    """
    Retrieve company information for a given year by company denomination. Path
    is /company/denomination/<string:denomination> GET method only and no strict slash.
    :param denomination: String, denomination of the company.
    :param year: Year of results to return, default is None.
    :return: HTTP Response as application/json. Contain all known information.
    """
    return YearDenominationCompany(denomination, year)
```

Serveur Web, enregistrement des requêtes

```
# http://api.enthic.fr/company/denomination/thales/2017
@application.route("/company/denomination/<string:denomination>/<string:year>", methods=['GET'], strict_slashes=False)
@insert_request
def company_denomination_year(denomination, year):
    """
    Retrieve company information for a given year by company denomination. Path
    is /company/denomination/<string:denomination> GET method only and no strict slash.
    :param denomination: String, denomination of the company.
    :param year: Year of results to return, default is None.
    :return: HTTP Response as application/json. Contain all known information.
    """
    return YearDenominationCompany(denomination, year)
```

- Un décorateur permet d'exécuter une routine avant et/ou après une fonction. Le temps d'insertion en base ne s'ajoute donc pas au temps de réponse.

Serveur Web, les points d'API

- Résumé sur <https://api.enthic.fr/>

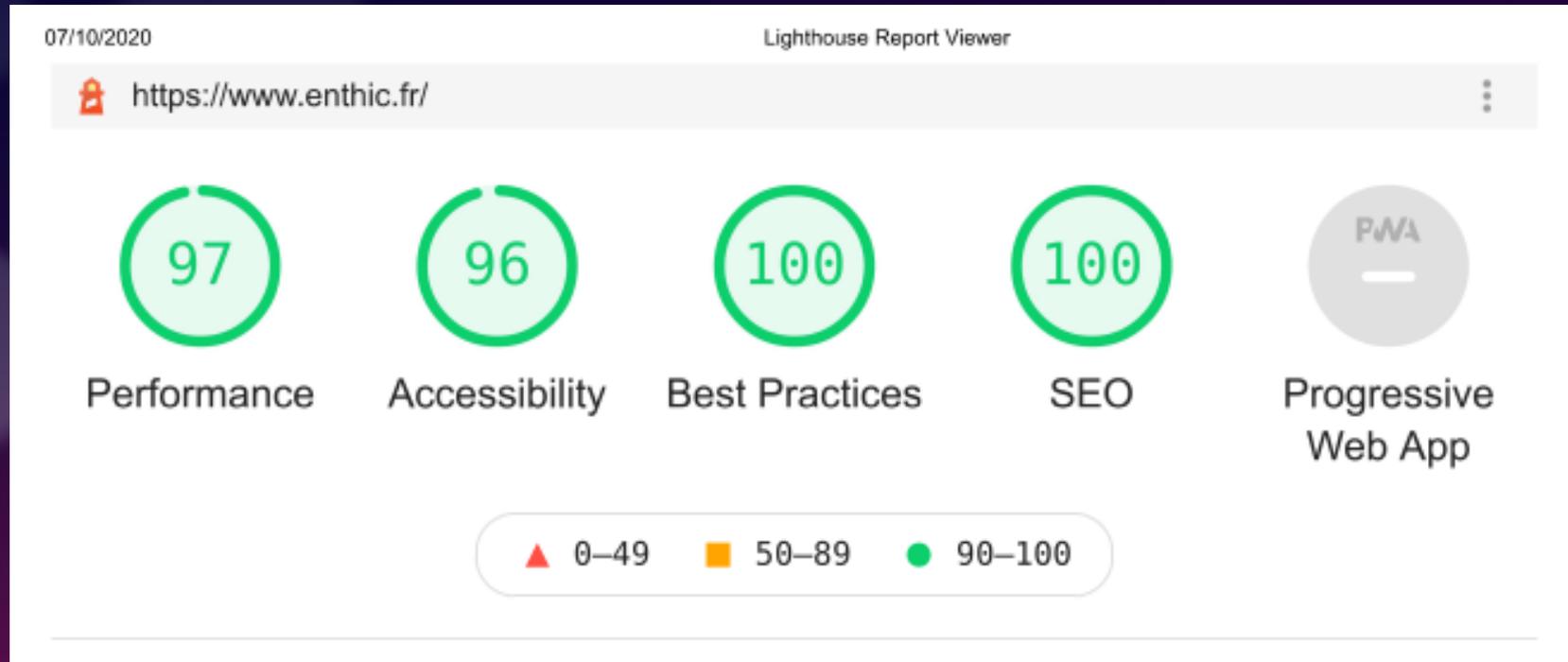
The screenshot shows the Swagger UI for the API 'Swagger French Company Account and Ethical Score' (version 2020.4.2). The page includes the Swagger logo, the API URL 'https://api.enthic.fr/swagger.json', and an 'Explore' button. Below the title, there is a description of the API and links to contact the developer or find out more about Swagger. A 'Schemes' dropdown menu is set to 'HTTP'. The page lists several related projects: 'Enthic' (with website https://enthic.fr), 'PapIT' (with website https://papit.fr), 'Enthic API code documentation' (with Sphinx documentation https://api.enthic.fr/documentation/index.html), and 'Wexample Labs' (with website https://wexample.com). At the bottom, there are three API endpoints listed:

- GET** `/siren/{siren}` Retrieve all company information by SIREN.
- GET** `/siren/{siren}/{year}` Retrieve company information by SIREN for a given year.
- GET** `/siren/{siren}/average` Retrieve company yearly average information by SIREN.

Intégration continue

- Ensemble d'outils et de pratiques qui facilitent l'évolution d'une application
- Packaging du code dans un package Python classique
- Auto-documentation avec **sphinx**. Génère la documentation à partir du README et des docstrings
<https://api.enthic.fr/documentation/index.html>
- Testing avec le framework de test **Pytest**, ± 1 ligne de code pour une de test. 708 tests pour le serveur Flask.
- SonarQube évalue la qualité de tout le code
- Jenkins build un client JS, Java et Python avec leurs tests tous les jours <https://jenkins.wexample.com/job/enthic-client-builder/>

Site web pour communiquer, enthic.fr



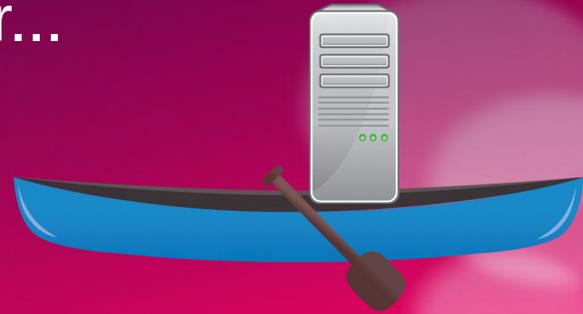
Usage, Front End

- Un moteur de recherche d'entreprise
- Vue détaillé des comptes

<https://opencompaniesdata.netlify.app/recherche/detail?id=300963915>

Impact Score sélectionné par Data For Good

- Impact Score : *Transmettre des critères simples et concrets essentiels à l'ensemble des entreprises françaises, loin du green ou social washing.*
- Actuellement basé sur un questionnaire d'Oxfam uniquement. Ils étudient la possibilité de benchmarker les entreprises par rapport aux données financières
- 6 Data Scientists bénévoles étudient la base Enthic
- Impact Score, impacte déjà le serveur...



Prise de contact

- Ministère de la transition écologique
- Société de conseil en financement de projet innovants (CIR)

Difficultés

- Difficulté à associer une maison mère à ses filiales et donc d'avoir une vue d'ensemble d'un grand groupe. Exemple total plus de 200 filiales probables
<https://opencompaniesdata.netlify.app/recherche?text=total>
- Erreur de saie. Orano group à saisie une année en M€ au lieu d'€

Évolutions

- Création de nouveaux indexes pour faciliter les recherches.
- Faciliter le paramétrage de l'application
- Intégrer à la base et aux API les résultats des recherches faites par data for good.

Sponsors et remerciements



- JetBrains leader des environnements de développement intégrés fournit gracieusement une licence *All Products*



- Le projet est supporté techniquement et moralement par l'association Wexample, réseau de Freelance et Laboratoire d'innovations des métiers du web.



- Léonard Michelet (Code For France)
<https://github.com/leonarf/OpenCorporateFacts>