Extraction de données relatives aux produits alimentaires à partir de documents non structurés

Pierre MASSÉ

Juin 2020

Résumé

La gestion de l'information produit est devenu un enjeu de société majeur ces dernières années. Les scandales sanitaires récents ont déclenché une prise de conscience collective des consommateurs, en parallèle de la mise en place de réglementations de plus en plus contraignantes pour l'ensemble des acteurs de la filière[1][2]. À ce titre, le Groupe Pomona a lancé ces dernières années un projet majeur de refonte des processus et des outils de gestion de l'information produit.

La première filiale a fait l'objet d'un déploiement réussi, mais qui a toutefois mis en évidence le fait que des gains à la fois en qualité et en productivité restent accessibles.

La mise en place d'outils mettant en oeuvre les principes du Machine Learning appliqués au traitement du langage permettrait d'aider les opérationnels de la gestion de l'information à interpréter plus vite et mieux les documents mis à disposition par les fournisseurs du Groupe.

Le présent rapport détaille la mise en place d'un outil permettant d'extraire les listes d'ingrédients des fiches techniques transmises par les fabricants des produits.

Table des matières

Première partie

Contexte métier

Chapitre 1

Description du Groupe

1.1 Le métier du Groupe Pomona

1.2 Les deux niveaux de décentralisation

1.3 Les branches

1.3.1 Les branches RHD

Préciser ici le non recouvrement des produits entre les branches

1.3.2 Les branches spécialistes

Dire que là, entre elles pas trop, mais avec les branches RHD, si.

1.3.3 L'étranger

1.3.4 Recouvrements des gammes de produits

Mettre ici un schéma représentant les gammes de produits

Chapitre 2

La gestion de l'information produit

2.1 L'information produit

2.2 Le processus associé

2.3 Le PIM (Product Information Management)

Deuxième partie

Les données

Chapitre 3

Le périmètre produit

3.1 Accessibilité de la donnée en fonction des branches

3.2 Les branches déployées

3.3 Les types de produit

3.4

Chapitre 4

Les données utilisables

4.1 Données structurées

4.2 Données non structurées

4.3 Pièces jointes

4.3.1 Fiches techniques fournisseur

4.3.2 Étiquettes produit

4.3.3 Fiches logistiques fournisseur

4.3.4 Fiches techniques et argumentaires Pomona

4.4 Analyse qualitative des données

Montrer qu'un sondage basique fait que la qualité actuelle est perfectible

Mettre également la distribution numérique des produits par fournisseur et insister sur la difficulté posée par de multiples formats

4.5 Les données « manuellement étiquetées »

Montrer comment elles ont été produites

Expliciter les règles de gestion qui ont été listées pendant l'étiquetage manuel

Evaluer la cohérence entre étiquettes manuelles et contenu du PIM

Troisième partie

Construction d'un modèle

Quatrième partie

Les objectifs de ce projet

Chapitre 5

Les cas d'usage

5.1 Objectifs : Qualité et productivité

5.2 La préalimentation d'information

5.3 Le contrôle à la saisie fournisseur

5.4 L'aide aux vérifications Pomona

5.5 Les contrôles en masse asynchrones

Chapitre 6

Le type de données à récupérer

6.1 La composition produit

6.2 Les données nutritionnelles

6.3 Les données logistiques

Cinquième partie

Travaux subséquents

Chapitre 7

Opérationnalisation de cette maquette

7.1 Client et sponsor métier

7.2 Définition des règles de gestion

7.3 Mise en place d'une organisation projet

7.4 Industrialisation du code

Prochaines étapes : opérationnalisation via API

Documentation

Chapitre 8

Extension des fonctionnalités offertes

Sixième partie

Figures et tableaux

Liste des tableaux

Table des figures

Septième partie

Bibliographie

Bibliographie

[1] Conseil de l'Union Européenne. Règlement n°1169/2011 dit inco, nov 2011. https://www.senat.fr/europe/textes_europeens/ue0120.pdf.

[2] Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes. Étiquetage des denrées alimentaires : nouvelles règles européennes, jan 2015. https://www.economie.gouv.fr/dgccrf/etiquetage-des-denrees-alimentaires-nouvelles-regles-europeennes.

Huitième partie

Exemple de documents fournisseur

Annexe A

Fiches techniques

Annexe B

Étiquettes produit

Neuvième partie

Le code utilisé

Annexe C

Extraction de données du PIM

Annexe D

Conversion des pièces jointes en textes

Annexe E

Identification des listes d'ingrédients