

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[2][3]. 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 du Groupe a fait l'objet d'un déploiement réussi, mais cela 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

I	Contexte métier	6
1	Description du Groupe	7
1.1	Le métier du Groupe Pomona	7
1.2	La décentralisation	9
1.2.1	Les Directions fonctionnelles	9
1.2.2	Les clients du Groupe	9
1.2.3	Premier niveau de décentralisation : les branches	10
1.2.4	Le second niveau de décentralisation : les succursales	14
2	La gestion de l'information produit	16
2.1	L'information produit	16
2.1.1	Utilisations de l'information produit	16
2.1.2	Des produits bruts aux produits transformés	18
2.1.3	Les grands types d'information	18
2.2	Le processus	23
2.2.1	Le fournisseur est propriétaire des informations produit	23
2.2.2	La notion de produit et d'article	23
2.2.3	Les contrôles	25
2.3	Les outils informatiques associés	28
2.3.1	Les branches faiblement outillées	28
2.3.2	Le GIP	28
2.3.3	Le PIM	29

II	Les données	32
3	Le périmètre produit	33
3.1	Accessibilité de la donnée en fonction des branches	33
3.2	Les branches déployées	33
3.3	Les types de produit	33
4	Les données utilisables	34
4.1	Données structurées	34
4.2	Données non structurées	34
4.3	Pièces jointes	34
4.3.1	Fiches techniques fournisseur	35
4.3.2	Étiquettes produit	35
4.3.3	Fiches logistiques fournisseur	35
4.3.4	Fiches techniques et argumentaires Pomona	35
4.4	Récapitulatif de la complétude des données	35
4.5	Analyse qualitative des données	35
4.6	Les données « manuellement étiquetées »	35
III	Les objectifs de ce projet	36
5	Les cas d’usage	37
5.1	Objectifs : Qualité et productivité	37
5.2	La préalimentation d’information	37
5.3	Le contrôle à la saisie fournisseur	37
5.4	L’aide aux vérifications Pomona	37
5.5	Les contrôles en masse asynchrones	37
6	Les types de données à récupérer	38
6.1	La composition produit	38
6.2	Les données nutritionnelles	38
6.3	Les données logistiques	38

7	Le choix du cas d'usage	39
7.1	Les multiples formats	39
7.2	Les informations « spatialisées »	39
7.3	La complexité dans la représentation des données logistiques . . .	39
7.4	La moindre représentation des étiquettes	39
IV	Construction du modèle	40
8	Les principes généraux	41
8.1	Contenu du texte d'une liste d'ingrédients	41
8.2	Limitation à l'identification des listes d'ingrédients	41
8.3	Conversion de documents en texte	42
9	Construction d'un modèle simple « ouvert »	43
9.1	Extraction des données	43
9.2	Conversion en blocs de texte	44
9.3	Train/Test split	44
9.4	Entraînement du modèle	44
9.5	Calcul de la similarité	44
9.6	Illustration des résultats obtenus	44
10	Utilisation des données manuellement étiquetées	45
10.1	Chargement des données manuellement étiquetées	45
10.2	Train/Test split	45
10.3	Entraînement du modèle	45
10.4	Illustration des prédictions obtenues	45
11	Mesure de la performance	46
11.1	Précision	46
11.1.1	Approche naïve	46
11.1.2	Avec du « text-postprocessing »	46
11.2	Similarité cosinus	46

11.3	Fonction de <i>loss</i> spécifique	46
11.3.1	Distance de Levenshtein	47
11.3.2	Distance de Damerau-Levenshtein	47
11.3.3	Distance de Jaro	47
11.3.4	Distance de Jaro-Winkler	47
11.4	Cross-validation des modèles précédents	47
11.4.1	Modèle « ouvert »	47
11.4.2	Modèle entraîné sur les données étiquetées manuellement	47
12	Transfer learning	48
12.1	Principe du pré-entraînement	48
12.2	Illustration de l'impact sur la performance	48
13	Hyperparameter tuning	49
13.1	Les paramètres ajustables	49
13.2	Application d'une grid search	49
V	Travaux subséquents	50
14	Opérationnalisation de cette maquette	51
14.1	Client et sponsor métier	51
14.2	Définition des règles de gestion	51
14.3	Mise en place d'une organisation projet	51
14.4	Industrialisation du code	51
14.5	Monitoring de la performance du modèle	51
15	Extension des fonctionnalités offertes	52
15.1	Prise en compte de nouveaux types de pièces jointes	53
15.2	Utilisation d'outil d'OCR pour les pdf non structurés	53
15.3	Mise en place d'outil de spatialisation des textes	53
15.4	Construction d'outils d'extraction de données connexes à la com- position	53

15.5	Élargissement aux données nutritionnelles	53
15.6	Extraction « opportuniste » d'informations complémentaires	53
15.7	Évaluation de la performances sur d'autres familles de produits .	53
VI	Figures, tableaux et bibliographie	54
VII	Exemple de documents fournisseur	58
A	Fiches techniques	59
B	Étiquettes produit	60
VIII	Le code utilisé	61
C	Extraction de données du PIM	62
D	Conversion des pièces jointes en textes	63
E	Identification des listes d'ingrédients	64

Première partie

Contexte métier

Chapitre 1

Description du Groupe

L’objet de l’ensemble de cette première partie est de donner sur le Groupe Pomona des éclairages nécessaires à la compréhension du cas d’usage développé. Bien d’autres aspects sur la société, pourraient être mentionnés (ex : des indicateurs sur l’activité, l’histoire du Groupe. . .) mais ils seront omis car non indispensables à la compréhension du sujet. Plus de détails sur le Groupe sont accessibles sur le site web de la société[8].

1.1 Le métier du Groupe Pomona

Le Groupe Pomona est une société de distribution livrée de produits alimentaires à destination des professionnels des métiers de bouche. L’activité du Groupe consiste uniquement à acheter et revendre de la marchandise, à l’exclusion de toute activité de fabrication ou de transformation¹. Le Groupe Pomona est une société de *distribution*. Elle ne possède d’ailleurs pas d’actif industriels (autre que des entrepôts logistiques) ni d’agrément pour transformer les marchandises.

1. de très rares cas de transformation existent (ex : mûrissage de fruits, filetage de poisson) mais sont extrêmement exceptionnels

Cette activité d'achat/vente se fait dans la majorité des cas sous le régime du *négoce*, à savoir que le Groupe acquiert la propriété des marchandises qu'il commercialise avant de la céder à ses clients. L'autre régime est celui dit de la *prestation (logistique)*. Dans ce cas, par le jeu d'écritures comptables, la valorisation du stock disparaît des comptes du Groupe. Néanmoins, indépendamment de cet aspect purement comptable, l'ensemble :

des flux de documents : commandes d'achat, factures fournisseur, commandes de vente, factures clients

des flux financiers : paiements fournisseur, paiements client

des flux physiques : réception et stockage, préparation et expédition

restent largement inchangés.

Pour résumer, l'activité de l'ensemble des entités du Groupe pourraient se résumer via le schéma présenté en figure 1.1

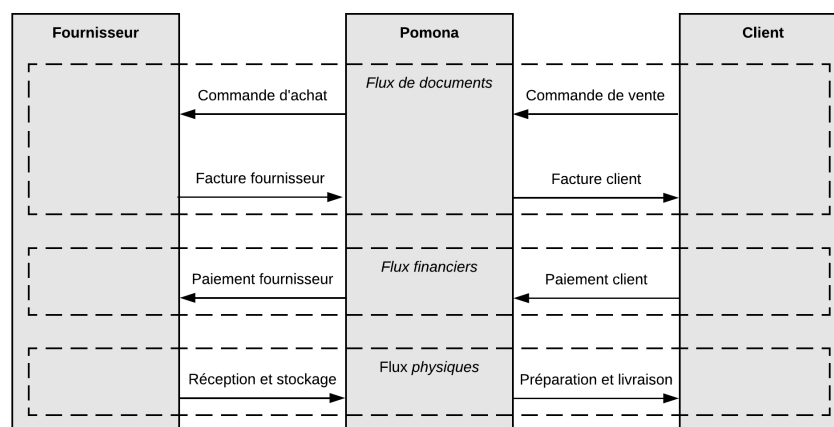


FIGURE 1.1 – Les flux métier avec les partenaires commerciaux

Le métier du Groupe est d'être un grossiste, qui achète et revend des produits alimentaires² sans produire ou transformer quoi que ce soit.

2. dans la grande majorité des cas, cf. paragraphe *Les produits non-alimentaires* page 12

1.2 La décentralisation

Le Groupe Pomona est un Groupe fortement décentralisé, avec des organisations largement indépendantes les unes des autres.

1.2.1 Les Directions fonctionnelles

Pour des raisons évidentes de recherche de synergies ou de conformité réglementaires, certaines activités restent toutefois mutualisées à la maille du Groupe. Il s'agit des organisations suivantes :

La Direction Administrative et Financière (DAF) : regroupe les équipes comptables Groupe, l'audit interne et la consolidation financière

La Direction Qualité : est en charge de définir et contrôler l'application des standards de qualité

La Direction des Systèmes d'Information (DSI) : développe et maintient en condition opérationnelles les systèmes d'information du Groupe

La Direction Technique et Logistique (DTL) : est en charge des projets immobiliers (entrepôts), des négociations avec les transporteurs et joue un rôle de conseil interne sur les sujets logistiques

La Direction des Ressources Humaines : se charge de l'ensemble des aspects en lien avec le recrutement, la paye et les sujets sociaux

La Direction Commerciale Groupe (DCG) : définit une stratégie et des bonnes pratiques commerciales et marketing

1.2.2 Les clients du Groupe

Afin de comprendre l'organisation du Groupe, il est nécessaire de connaître la typologie de ses clients. Comme mentionné précédemment, le Groupe s'adresse exclusivement aux professionnels des métiers de bouche. Aucune marchandise n'est vendue à des particuliers. Les principales typologies de clients sont les suivantes :

Les Sociétés de Restauration : elles exploitent les restaurants d'entreprise et certaines cantines d'établissement d'enseignement supérieur

Les Marchés Publics : regroupent les clients qui dépendent des collectivités (écoles, hôpitaux, prisons, ...)

La restauration commerciale : est l'ensemble des restaurants à vocation commerciale, qu'ils soient chaînés (hippopotamus, O'Tacos, ...) ou indépendants (« le restaurant du coin »)

Les spécialistes : il s'agit des détaillants spécialisés qui s'adressent aux particuliers. Boulangers, pâtisseries, bouchers, traiteurs, vente à emporter, ...

Les Grandes et Moyennes surfaces (la GMS) : sont les enseignes de la grande distribution. En général, l'accès à ces clients est compliqué par les règles mises en place par leurs centrales d'achat. Il représentent en général qu'un canal de vente d'opportunité.

Les trois premières de ces catégories représentent ce que l'on appelle la *Restauration Hors Domicile (RHD)* (ou parfois également la Restauration Hors Foyer, RHF).

1.2.3 Premier niveau de décentralisation : les branches

Le Groupe Pomona est divisé en branches, qui sont des organisations indépendantes et qui ont toute latitude pour gérer leurs stratégie et politique commerciales, la gestion de leurs achats, leur stratégie marketing, ... En particulier, les systèmes d'information ne sont pas identiques entre les différentes branches. Afin d'éviter de se concurrencer entre elles, leurs domaines d'activité respectifs ont été partitionnés par familles de produit commercialisés, segments client cibles et géographie.

Les branches RHD

Les branches RHD s'adressent comme leur nom l'indique aux clients de la Restauration Hors Domicile (cf. section 1.2.2 page 9) en France. Elles se ré-

partissent ce marché en travaillant des gammes de produits distinctes. Il s'agit des branches historiques du Groupe, qui représentent l'essentiel de son chiffre d'affaire. La répartition par produit est la suivante :

PassionFroid : spécialiste des *produits surgelés, de la viande fraîche et des produits laitiers*

ÉpiSaveurs : spécialiste des produits qui se conservent à température ambiante : *produits d'épicerie, conserves, boissons et consommables de cuisine non-alimentaires*

TerreAzur : spécialites des *Fruits et Légumes frais, et Produits De la Mer frais*

La non-concurrence entre les branches est assurée par le fait qu'elles ne commercialisent pas les mêmes produits. Bien que nommées RHD, elles peuvent également vendre leurs produits à la grande distribution, mais généralement ces marchés sont verrouillés par les centrales d'achat des grandes enseignes.

Les branches spécialistes

Les branches spécialistes s'adressent aux clients dits spécialistes (cf. section 1.2.2 page 9) en France. Elles sont en mesure de commercialiser tout type de produit pour répondre aux besoins de leurs clients. En particulier, elles peuvent tout à fait commercialiser certains produits qui sont également vendus par les branches RHD. Elles se répartissent la clientèle spécialiste de la manière suivante :

Délice et Création : s'adresse aux *Boulangers et Pâtisseries*

Saveurs d'Antoine : s'adresse aux *Bouchers, Charcutiers et Traiteurs*

Relais d'Or : s'adresse à la *restauration indépendante nomade*

Comme pour les branches RHD, ces branches peuvent lorsqu'elles en ont l'opportunité vendre leurs produits à la GMS.

L'étranger

Bien que le Groupe Pomona soit une société dont l'essentiel de l'activité est faite sur le marché français, deux réseaux sont en cours de constitution sur des pays limitrophe. Ces branches sont susceptibles de travailler tout type de produit, à destination de tout type de client. Elles sont positionnées sur les marchés suivants :

Pomona Suisse : présente sur le marché Suisse

Pomona Iberia : présente sur le marché Espagnol

On peut synthétiser la répartition de l'activité par branche de la manière présentée à la figure 1.2.

Les produits non-alimentaires

Si l'essentiel des produits commercialisés par les branches du Groupe sont des produits alimentaires, comme évoqué précédemment une partie de l'activité commerciale se fait tout de même autour de produits non-alimentaires. Ces produits restent malgré tout destinés exclusivement aux professionnels des métier de bouche, et il s'agit de consommables (par opposition à des articles d'électroménager, de la vaisselle non-jetable, ...).

On distingue en général deux catégories de produits non-alimentaires :

- les produits dits « d'hygiène »
- les produits dits « de chimie »

Les produits de chimie regroupent les produits qui doivent faire l'objet d'une fiche de données de sécurité au sens du règlement Européen No 1907/2006 dit « REACH » (Registration, Evaluation, Authorisation and Restriction of Chemicals) [1].

Les produits d'hygiène regroupent tous les autres produits non-alimentaires. L'appellation « d'hygiène » est donc réductrice, dans la mesure où cette large famille regroupe les consommables de nettoyage (éponges, papiers absorbants, ...) mais également tout type d'autres consommables (serviettes de tables, gobelets en plastiques, pics à brochettes, boîtes de produits à emporter, ...).

La commercialisation de produits non-alimentaires existe au sein du Groupe, mais on se focalisera pour la suite sur les produits alimentaires qui reste le coeur de métier.

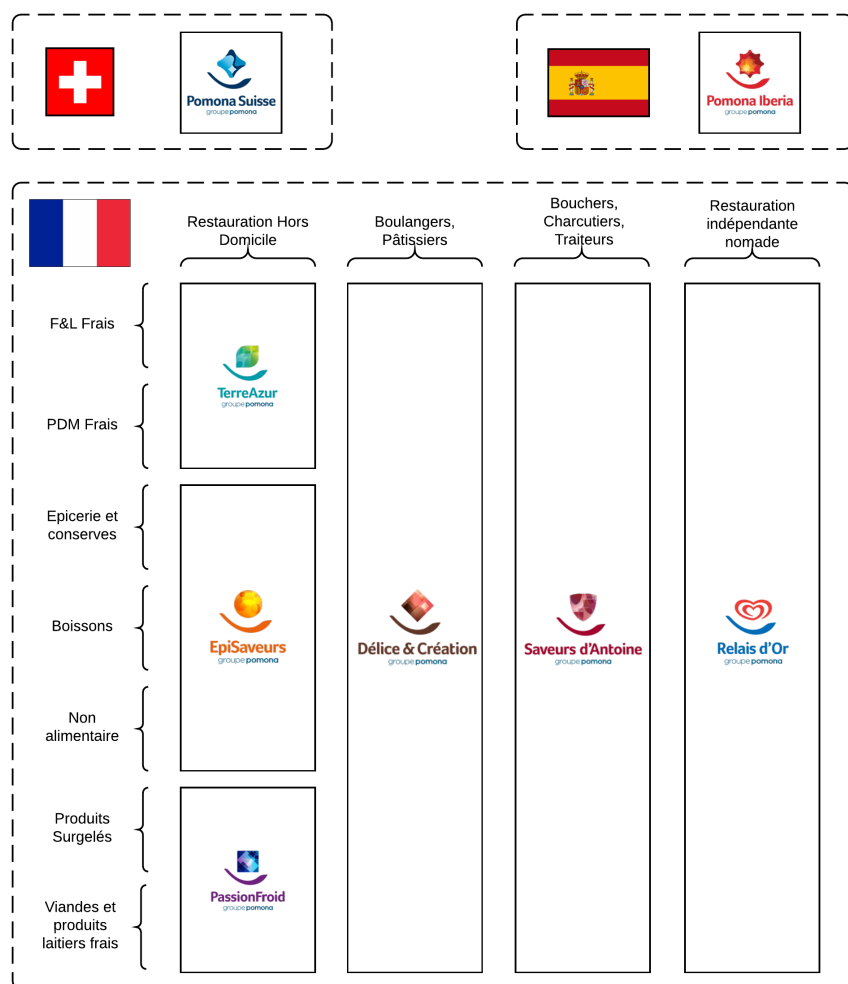


FIGURE 1.2 – La répartition de l'activité des branches

1.2.4 Le second niveau de décentralisation : les succursales

Chacune des branches est elle-même à son tour décentralisée en un réseau d'entrepôts régionaux : les succursales (parfois également appelées simplement « régions »). Ces succursales sont gérées comme des PME indépendantes, avec un directeur et un compte de résultat qui leur est propre. Si certaines négociation avec des fournisseurs ou des clients nationaux sont parfois menée par les branches, les succursales sont autonomes dans :

- la définition de leur assortiment, même si des contraintes s'appliquent
- la stratégie de développement commercial
- la négociation des prix d'achat
- la négociation des prix de vente
- la politique de rémunération de leurs employés

À ce titre, elles ont leurs propres équipes d'achat, leurs équipes commerciales (télévente et vente route), leurs équipes administratives et évidemment leurs équipes logistiques (essentiellement en entrepôt et les chauffeurs livreurs en charge des livraisons client).

Certaines activités restent de la responsabilité des équipes centrales des branches, comme : la négociation avec les clients ou les fournisseurs nationaux, la constitution de l'assortiment commun (les produits que toutes les succursales doivent détenir), la gestion des référentiels de données de base métier, ...

Un exemple de maillage régional est présenté en figure 1.3, sachant que ce maillage régional est différent pour chacune des branches.



FIGURE 1.3 – Le maillage régional de la branche ÉpiSaveurs

Chapitre 2

La gestion de l'information produit

Ce chapitre a pour vocation à éclairer les aspects métier en lien avec la gestion de l'information produit. C'est le seul processus métier qui sera détaillé dans la mesure où c'est uniquement celui qu'il est nécessaire de connaître pour comprendre les cas d'usage développés ultérieurement.

2.1 L'information produit

2.1.1 Utilisations de l'information produit

Conformité réglementaire

La gestion de l'information produit est essentiellement une contrainte réglementaire à satisfaire. Comme mentionné au préambule, la réglementation autour de l'information des consommateurs s'est sans cesse complétée au cours des dernières années. Un des textes centraux est le règlement n°1169/2011 dit INCO (INformation COnsommateur)[2][3]. C'est ce règlement qui définit l'ensemble des informations qui doivent être étiquetées sur le produit (liste d'ingré-

dients, tableau de données nutritionnelles, . . .), mais également affichée au client lors de commande en ligne sur les sites de e-commerce. Il s'agit principalement d'informations relatives à la sécurité alimentaire (ex : les allergènes) ou la santé (ex : informations nutritionnelles).

Attentes client

Les consommateurs finaux (les « convives ») étant de plus en plus sensibles au contenu de leur assiette, les clients du Groupe sont de plus en plus demandeurs d'information relatives aux produits qu'ils commandent. Ils demandent donc régulièrement des informations qui vont au-delà de ce qui est normalement prévu par la réglementation.

De plus, sur certains marchés pour lesquels des contrats couvrant sur de longues périodes - jusqu'à un an - sont établis (les marchés publics sont très concernés), il n'y a pas d'échantillonnage des produits. La seule manière pour ces clients d'évaluer la qualité des produits est de se référer aux documents contenant les informations produit, fournis par les distributeurs.

Gestion

Certaines informations relatives aux produits sont nécessaires pour des raisons de gestion administrative. Par exemple, la gestion des taxes sur les produits alimentaires est complexe :

- les taux de TVA sont variables en fonction du type de produit
- des taxes spécifiques s'appliquaient aux produits contenant de l'huile ou de la farine
- des règlements particuliers s'appliquent aux alcools
- ...

D'autres informations, comme la nomenclature douanière, sont nécessaires pour effectuer les déclarations auprès des douanes européennes.

Un autre type d'information capital pour la gestion des flux d'achat et de vente sont les informations logistiques, qui définissent par exemple le nombre

d'unités consommateur dans le colis, le nombre de colis sur une palette, ... Une gestion rigoureuse de ces informations est indispensable pour que les flux d'achat ou de vente soient correctement exécutés (que les quantités commandées soient les bonnes, que les montants facturés soient corrects, ...).

2.1.2 Des produits bruts aux produits transformés

Le niveau d'exigence en termes d'information produit est variable en fonction du niveau de transformation de ce produit. Par exemple, sur des fruits et légumes frais, à peu de choses près seul le pays d'origine doit être affiché au client. Sur une barre chocolatée, ou un plat cuisiné, il sera nécessaire d'afficher :

- une liste d'ingrédients (mettant en évidence les allergènes)
- un tableau de données nutritionnelles (protéines, glucides, ...)
- une dénomination réglementaire

2.1.3 Les grands types d'information

On se focalisera dans ce paragraphe sur les informations relatives aux *produits alimentaires*.

La composition

La première grande famille de données réglementaires sont les données de composition. Elles détaillent quels sont les ingrédients qui sont mis en oeuvre dans la fabrication des produits. Évidemment, la composition a en général plus de sens que pour les produits transformés que pour les produits bruts. Elle peut prendre la forme d'un texte listant la liste des ingrédients (l'étiquetage de ce texte est en général obligatoire sur les emballages des produits), ou d'un tableau.

Les ingrédients incluent également les additifs. Il s'agit de substances ajoutées à la recette pour répondre à des fonctions particulières (colorant, exhausteur de goût, émulsifiant). Elles ne représentent en général un pourcentage en masse négligeable dans la composition totale du produit.

Le pourcentage en masse est parfois inclus sur certains ingrédients. La réglementation l’oblige dans certains cas, par exemple quand l’ingrédient en question est mentionné dans la dénomination du produit (pour une *tarte aux framboises*, la proportion de framboise doit être mentionnée dans la composition).

Enfin, un aspect à la fois réglementaire et particulièrement important est la présence d’allergènes dans la composition. Le règlement INCO[2][3] impose de mettre en évidence les allergènes relevant d’une des 14 catégories suivantes :

1. Céréales contenant du gluten, à savoir blé, seigle, orge, avoine, épeautre, kamut ou leurs souches hybridées, et produits à base de ces céréales
2. Crustacés et produits à base de crustacés
3. Œufs et produits à base d’œufs
4. Poissons et produits à base de poissons
5. Soja et produits à base de soja
6. Lait et produits à base de lait (y compris le lactose)
7. Fruits à coque, à savoir : amandes, noisettes, noix, noix de cajou, noix de pécan, noix du Brésil, pistaches, noix de Macadamia ou du Queensland, et produits à base de ces fruits
8. Céleri et produits à base de céleri
9. Moutarde et produits à base de moutarde
10. Graines de sésame et produits à base de graines de sésame
11. Anhydride sulfureux et sulfites
12. Lupin et produits à base de lupin
13. Mollusques et produits à base de mollusques

Il peut y avoir deux niveaux de présence d’un allergène dans un produit (au-delà de la simple absence) :

intentionnellement mis en oeuvre : dans le cas où un ingrédient allergène fait volontairement partie de la recette. Ex : présence de moutarde dans un plat cuisiné.

contamination croisée : par exemple lorsque le produit fini est issu d'une chaîne de transformation qui traite un ingrédient allergène, mais que cet ingrédient ne fait pas partie de la recette. Ce cas de figure est en général mis en évidence par des mentions telles que *"Peut contenir des traces de soja"* ou bien *"Transformé dans un atelier processant également des fruits à coques et du sésame"*.

Les informations nutritionnelles

Une autre grande famille d'information produit sont les informations nutritionnelles. Elles détaillent la quantité des principaux nutriments contenus dans les produits. Certains d'entre eux sont rendus obligatoires par le règlement INCO[2][3] cf. l'exemple de tableau 2.1, et d'autres sont optionnels, comme par exemple la quantité de fer, de calcium, ...

Informations nutritionnelles	Pour 100g	Pour un biscuit	% des AJR pour un biscuit
Énergie	1674 kJ 398 kcal	209 kJ 50 kcal	3 %
Protéines	3.0 g	1.0 g	3 %
Matières grasses	13.0 g	1.6 g	2 %
dont acides gras saturés	5.8 g	0.7 g	4 %
Glucides	66 g	8.2 g	3 %
dont sucres	48 g	6.1 g	7 %
Fibres alimentaires	2.5 g	0.3 g	
Protéines	3.3 g	0.4 g	1 %
Sel	0.41 g	0.05 g	1 %

TABLE 2.1 – Exemple de tableau de données nutritionnelles

La réglementation rend obligatoire de mentionner les informations nutritionnelles de cette table pour 100g, ou 100mL de produit (pour les boissons).

Les informations nutritionnelles peuvent également se présenter sous forme d'allégations, qui ont des définitions précises dans la réglementation. Ces allé-

gations peuvent être : *sans sel, faible en sucres, riche en fibres, ...*

Les origines

Du fait de la complexification des opérations de transformation et de la complexification des flux d'échanges de marchandises, l'origine des produits alimentaires est une notion qui n'est pas définie avec précision dans l'absolu. Il n'y a donc pas non plus de réglementation précise sur le sujet, si ce n'est que l'information produit doit toujours être présentée de manière loyale au consommateur. On peut se donner une règle simple pour définir l'origine d'un produit alimentaire : plus il est brut, plus va compter l'origine de ses ingrédients ; plus il est transformé, plus va compter le lieu de dernière transformation.

Par exemple, sur des morceaux piécés de viande fraîche, on aura des origines multiples en fonction du pays de naissance, d'élevage ou d'abattage de la bête. Et à l'inverse, sur un steak haché cette information n'aura aucun sens dans la mesure où il est produit d'un assemblage de « minerais » pouvant provenir de multiples pays. L'industriel pourra choisir de communiquer sur le fait que la viande a été transformée en steak dans telle usine par exemple.

Les données logistiques

On appelle données logistiques essentiellement le plan de palettisation et de conditionnement du produit. Il s'agit de la définition de la « hiérarchie logistique » du produit. Cette hiérarchie se base d'abord sur la définition d'une « unité de base » qui est la plus petite unité légalement détaillable (i.e. qui porte l'ensemble des informations réglementaire pour sa commercialisation). Ces notions ont été standardisées par l'organisme international de standardisation GS1[5]. Deux exemples pour illustrer :

- pour un boîte de sachets de thé, l'unité de base est la boîte car les sachets de thé ne portent pas les informations nécessaires à leur commercialisation
- pour un paquet de barres chocolatées (comme celles qu'on peut trouver au

détail en boulangerie), l'unité de base est la barre car elle porte l'ensemble des mentions réglementaires sur son emballage

La hiérarchie logistique est à la fois :

- la définition des niveaux successifs d'emballage des produits : combien d'unités de base dans un paquet, combien de paquets dans un carton, combien de cartons sur une palette, ...
- la définition du contenu de l'unité de base (ex : le nombre de sachets de thé, le nombre de doses dans une boîte d'aides culinaires, ...)

Les données logistiques concernent également les durées de vie du produit (type de durée de vie : Date Limite de Consommation ou Date de Durabilité Minimale ; ainsi que la durée en jours entre la fin de production du produit et son expiration).

Parfois, certaines contraintes d'approvisionnement peuvent être mentionnées :

- unités commandables (ex : on ne peut commander que des cartons complets)
 - multiples de commande (ex : on ne peut commander les cartons que 10 par 10 pour des raisons de montage des palettes)
 - minimum de commande (ex : il faut commander au minimum 30 cartons)
- mais elles sont dépendantes d'un accord entre l'industriel et son client distributeur et ne sont donc pas à proprement parler des informations produit.

Les données administratives et financières

Les données dites administratives et financières regroupent le reste des informations de gestion pour lesquelles il existe des contraintes réglementaires. Il s'agit :

- du taux de TVA du produit
- de sa nomenclature douanière et du pays d'origine au sens de la déclaration d'échange de biens[4]
- de toute autre taxe applicable au produit

2.2 Le processus

2.2.1 Le fournisseur est propriétaire des informations produit

Comme présenté à la section 1.1, le Groupe Pomona n'a pas d'activité de fabrication ou de transformation de marchandises. À ce titre, l'ensemble des données produits ne peuvent être déterminées que par les fournisseurs de ces produits. L'ensemble des entités du Groupe s'appuient donc sur les données transmises par les industriels ou producteurs de marchandises.

Il peut arriver que certains produits soient achetés par Pomona à d'autres négociants non-producteurs. Dans ce cas, de la même manière que le Groupe Pomona a la responsabilité de collecter puis transmettre les informations produit à ses clients, ces fournisseurs négociants doivent eux-même aller chercher l'information produit et la transmettre à Pomona.

Dans tous les cas, c'est le fournisseur qui est responsable de produire et de transmettre l'information produit aux entités du Groupe Pomona.

2.2.2 La notion de produit et d'article

Un mot sur la modélisation des données adoptée est nécessaire pour comprendre les grandes lignes du processus. Comme il a été vu à la section 1.2.3 page 10, certains produits sont susceptibles d'être commercialisés par plusieurs branches du Groupe.

De plus, du fait que les systèmes d'information ne sont pas identiques entre les branches, certaines contraintes imposent parfois des différences de modélisation, des duplications volontaires de codes pour répondre à ces contraintes. Une illustration de ce point pour clarifier : la facturation client pour une canette de soda peut se faire au litre (permet de comparer les prix entre les différents conditionnement et les différentes marques) ou à la cannette (permet de se faire une idée du coût portion d'un produit). Or, la possibilité de pouvoir facturer un

même article dans plusieurs unités différentes n'est pas une fonctionnalité offerte par tous les systèmes d'information. En particulier, ÉpiSaveurs peut gérer dans ce cas un unique article et le facturer dans l'unité de son choix en fonction des demandes des clients. Mais Délice et Création (qui possède un système d'information différent) doit dupliquer cet article car une seule unité de facturation est possible pour un article donné.

Enfin, au-delà des contraintes liées au SI, certaines pratiques imposent de laisser aux branches une indépendance forte dans la gestion de leurs référentiels articles. Il faut savoir que commercialiser sous un même code article des produits qui sont similaires permet d'obtenir des gains de productivité à plusieurs niveaux :

- on économise des emplacements en entrepôt (un emplacement ne peut contenir qu'un article)
- on gagne du temps administratif dans la gestion des prix : le foisonnement d'articles impose de gérer plus de prix client
- ...

La contrepartie à adopter cette pratique est qu'il n'est alors plus possible de différencier ces produits similaires, par exemple pour leur appliquer des prix de vente distincts, ou bien offrir la possibilité à un vendeur de garantir au client la livraison d'un produit plutôt que l'autre. Néanmoins, en fonction de la clientèle adressée, certains produits pourront être considérés comme similaires, alors que pour d'autres ils ne seront pas interchangeables. Cet exemple est détaillé dans la figure 2.1 page 25.

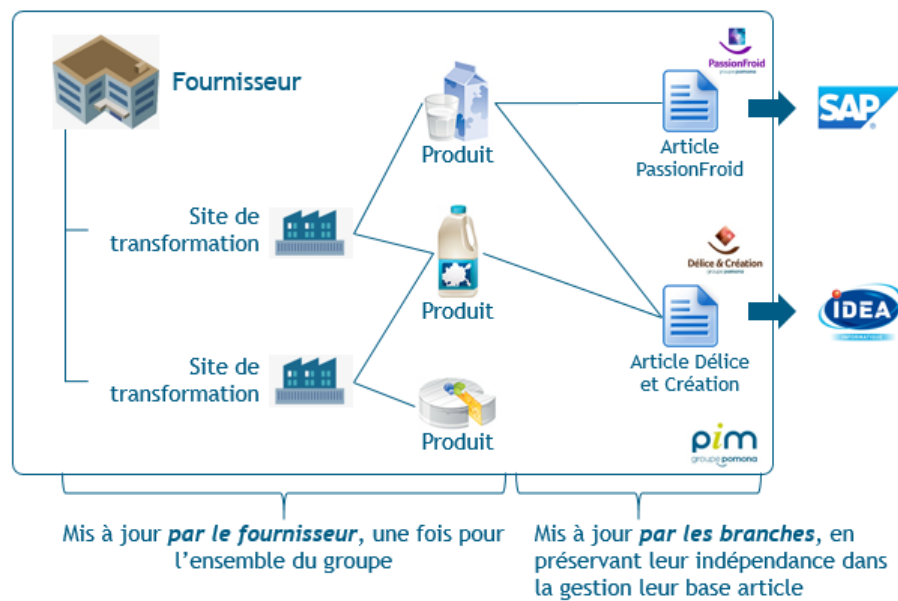
On différencie donc les deux notions suivantes :

les produits : ils représentent une marchandise physique produite par un fournisseur. Ce sont les produits qui portent les *informations produit* décrites à la section 2.1.3 page 18. Un produit ne peut appartenir qu'à un seul fournisseur. Le référentiel produit est unique pour l'ensemble du Groupe.

les articles : ce sont les objets qui sont gérés par les branches dans leurs

systèmes de gestion respectifs. Leurs attributs sont très liés au système d'information qui les porte. Chaque branche gère de manière autonome son référentiel article, incluant les liens qui sont faits entre produits et articles.

Cette modélisation permet de répondre à l'ensemble des contraintes présentées dans ce paragraphe.



Dans cet exemple fictif, le type de conditionnement du lait n'a aucun impact sur les clients de la branche Délice et Création. Elle commercialise donc sous un même code article deux produits distincts.

Pour des raisons de contrainte de conservation, la branche PassionFroid a quant à elle choisit de ne commercialiser qu'un seul produit - en brique. Elle pourrait choisir d'ouvrir un nouveau code article pour le lait en bouteille (non représenté sur le schéma ci-dessus).

FIGURE 2.1 – La distinction entre produit et article

2.2.3 Les contrôles

Il a été vu que le Groupe Pomona - dans sa qualité de distributeur - n'est pas en mesure de déterminer seul les informations produit sur les marchandises qu'il

commercialise. Néanmoins, comme cela a été vu dans la section 2.1.1 page 16, il est nécessaire que les différentes entités du Groupe soient en possession d'une information produit fiable. Or, avoir des données de qualité nécessite des efforts de la part des métiers, en particulier lorsque le processus n'est pas entièrement porté en interne dans la société. À ce titre, plusieurs étapes de contrôle ont été définies dans le processus de gestion du référentiel de données produit et article :

lorsque le fournisseur a saisi les données produit : la personne à l'origine de la demande de référencement (en général, un acheteur) doit contrôler la cohérence des données produit

lorsque le demandeur a demandé la création d'un article : le gestionnaire de référentiel valide à nouveau la cohérence des informations produit

après la création article, de manière asynchrone : le service qualité contrôle par échantillonnage les données d'une partie des produits et articles qui ont été modifiés pendant une période.

Le retour d'expérience montre que ces contrôles, loin d'être redondants, sont nécessaires pour avoir une qualité de données acceptable. Ces processus de contrôle sont décrit à la figure 2.2 page 27.

Les contrôles effectués à chacune des étapes sont les suivants :

contrôle de la complétude des données : vérification que les données transmises comportent l'ensemble des données attendues

contrôle de cohérence entre les données : vérification que les informations transmises sont cohérentes entre elles (ex : un allergène présent dans la liste d'ingrédients du produit a bien été signalé comme allergène par ailleurs)

contrôle de la cohérence avec les pièces jointes : en plus de données structurées, les fournisseurs transmettent également des fichiers portant des informations produit (ex : l'étiquette produit ou le visuel de l'emballage). Ces pièces jointes sont décrites à la section 4.3 page 34. La personne

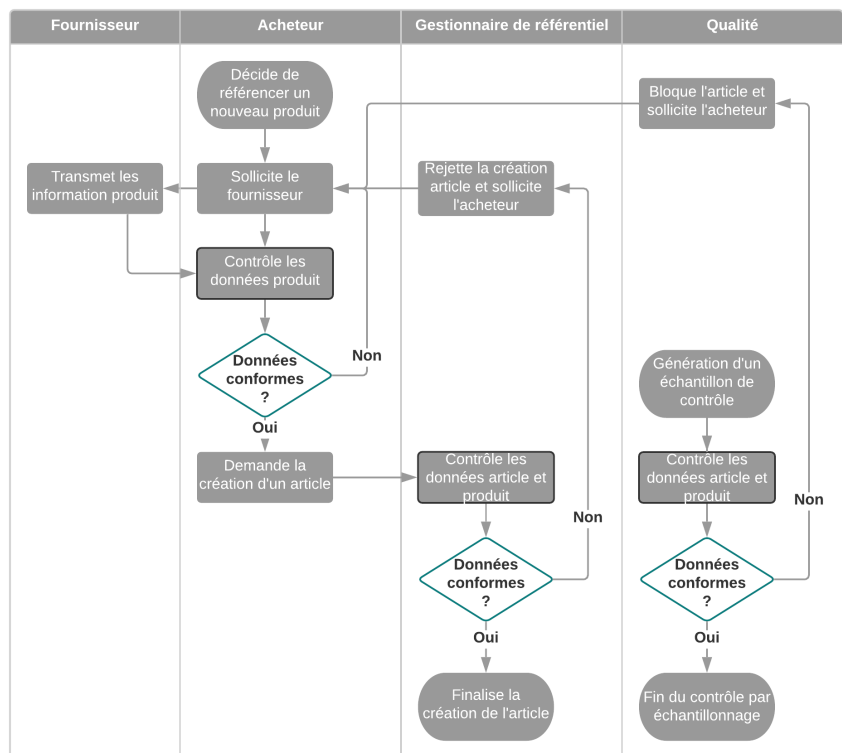


FIGURE 2.2 – Le processus de création article

en charge du contrôle vérifie que les données transmises sont cohérentes avec ces documents.

2.3 Les outils informatiques associés

Comme vu dans la description des branches du Groupe (voir section 1.2.3, page 10), les outils informatiques ne sont pas tous les mêmes sur l'ensemble des branches. Ainsi, les outils utilisés pour la gestion de l'information produit ne sont pas les mêmes.

2.3.1 Les branches faiblement outillées

Les branches étrangères (Pomona Suisse, Pomona Iberia), spécialistes (Déllice et Création, Saveurs d'Antoine) et la branche TerreAzur sont aujourd'hui faiblement outillées. Cela signifie que l'information produit est en général stockée uniquement sous la forme de fichiers (essentiellement les pièces jointes, décrites à la section 4.3 page 34). L'ensemble des échanges avec les fournisseurs se font par mail, et les articles sont créés directement dans les systèmes de gestion par les gestionnaires de référentiel. Les liens entre les articles et les informations produit ne sont pas matérialisés dans les systèmes informatiques.

2.3.2 Le GIP

Le GIP (Gestion de l'Information Produit) est utilisé sur la branche Passion-Froid. C'est un système de gestion de l'information produit qui est maintenant obsolète et en cours de remplacement. Il a toutefois le mérite de permettre le stockage dans une application des données et des pièces jointes relatives aux produits, avec la possibilité d'accéder aux informations produit à partir des identifiants des articles. C'est ce système qui a permis de pouvoir alimenter les sites de e-commerce PassionFroid et ÉpiSaveurs avec les informations produit. Il s'agit d'une application qui n'est pas ouverte aux utilisateurs externes au Groupe, et les échanges avec les fournisseurs passent donc par des échanges de

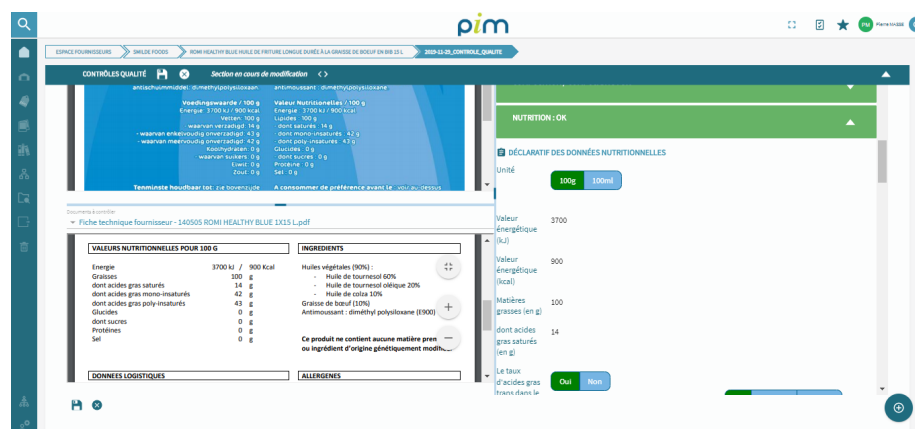
mails.

2.3.3 Le PIM

Le PIM (Product Information Management) est un système de gestion de l'information produit qui a été mis en production en mai 2019, pour la branche ÉpiSaveurs.

Description générale de l'outil

C'est un système qui porte l'ensemble du processus de gestion de l'information produit, tel que décrit à la figure 2.2 page 27. Il est accessible aux fournisseurs du Groupe, qui viennent directement mettre à disposition les données et les pièces jointes. Cet outil porte entre autres les fonctionnalités de contrôle des informations, comme illustré à la figure 2.3 page 29.



Cette capture d'écran montre l'outil de contrôle des données produit. Sur la partie gauche, le contenu des pièces jointes est affiché (visuel de l'emballage en haut, fiche technique en bas), sur la droite les données qui ont été transmises par le fournisseur.

FIGURE 2.3 – Une capture d'écran du PIM

La GDSN

La GDSN (Global Data Synchronization Network) est un réseau d'échange de données produit entre industriels, distributeurs, restaurateurs, ... Ce réseau est exploité par des opérateurs privés, mais le format et la chorégraphie des échanges a été standardisé par l'organisme de standardisation GS1. Son schéma de principe est décrit dans la figure 2.4 page 30. Sans rentrer dans le détail, au sein du Groupe Pomona l'utilisation qui en est faite est de récupérer les informations depuis ce réseau d'échange, afin de préalimenter les données produit pour les fournisseurs. Cette fonctionnalité permet de faire gagner du temps aux fournisseurs pour leur éviter une partie de ressaisie, mais également de limiter les erreurs. Toutefois, cette fonctionnalité ne permet pas à elle seule de garantir une parfaite qualité de données. Il s'agit uniquement d'un « tuyau », si les données en entrée ne sont pas correctes, elles ne seront pas correctes en sortie. Pour aller plus loin dans la compréhension de ce réseau, il est possible de consulter les ressources mises en ligne par GS1[6][7].

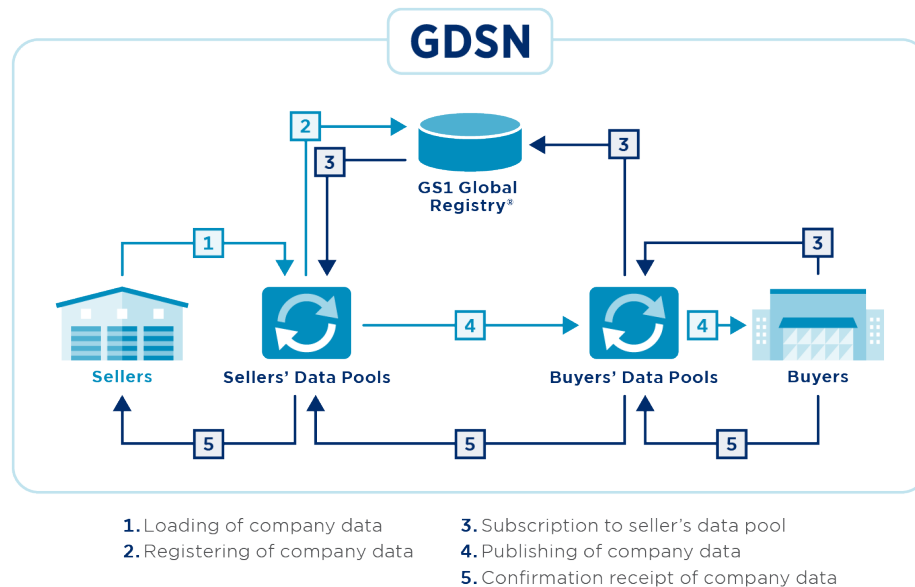


FIGURE 2.4 – Schéma de principe de la GDSN

L'identification des objets dans le PIM

Un dernier point à connaître à propos du PIM, est la manière d'identifier l'ensemble des objets en son sein. Chaque objet géré (ainsi que toute version archivée) porte un identifiant unique, nommé *uid* qui est totalement univoque. Pour la suite, on se basera sur ces uid pour faire référence à des produits stockés dans le PIM.



L'uid du produit affiché est mis en évidence dans l'url

FIGURE 2.5 – L'uid d'un produit

Les API

Un des aspects intéressants du PIM pour l'exploitation en masse des données produit, est qu'il expose des API permettant d'aller requêter l'ensemble de son contenu. Cela concerne à la fois les données dites structurées, mais également les pièces jointes. Cela rend les données produit de la branche ÉpiSaveurs bien plus simplement accessibles que celles des autres branches.

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

Chapitre 4

Les données utilisables

4.1 Données structurées

4.2 Données non structurées

Les listes d'ingrédients juste une liste ordonnées d'ingrédients triés par ordre décroissant de quantité mise en oeuvre.

Parfois détaillé par phase, mais en général déconseillé.

4.3 Pièces jointes

Dans chacune des sections, mentionner la volumétrie de données accessibles (avec les facettes migration, statuts, & compagnie) et tout

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 Récapitulatif de la complétude des données

Mettre ici un ou plusieurs tableaux récapitulatifs illustrant les données possédées quantitativement.

4.5 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

Dire ici qu'il y a finalement beaucoup de pdf qui possèdent des textes extractibles vs. uniquement des images.

4.6 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

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

Les types de données à récupérer

6.1 La composition produit

6.2 Les données nutritionnelles

6.3 Les données logistiques

Chapitre 7

Le choix du cas d’usage

7.1 Les multiples formats

7.2 Les informations « spatialisées »

7.3 La complexité dans la représentation des données logistiques

7.4 La moindre représentation des étiquettes

blablabla

Au vu des différentes contraintes listées plus haut, on s’attachera à extraire les listes d’ingrédients des produits alimentaires de la branche EpiSaveurs depuis les fiches techniques fournisseur, en se basant sur le contenu textuel de ces documents.

Quatrième partie

Construction du modèle

Chapitre 8

Les principes généraux

8.1 Contenu du texte d'une liste d'ingrédients

En général, chaque ingrédient sera présent une seule fois dans la liste.

Le calcul d'embeddings via des modèles tels que SVD ou Word2Vec fait peu de sens.

l'extraction des textes se fait au format Bag Of Words, sans utiliser de notion d'IDF. L'utilisation de TF semble également sujette à caution.

8.2 Limitation à l'identification des listes d'ingrédients

On est sur une taxonomie d'informations limitée dans les fiches techniques.

On pourrait envisager de classifier l'ensemble des textes présents dans les fiches techniques.

Mais l'absence de données étiquetées rend cette tâche impossible. La charge d'étiquetage d'un nombre représentatif de blocs de texte de fiches techniques

est trop importante pour être mise en oeuvre dans le cadre de ce projet.

8.3 Conversion de documents en texte

dire ici qu'on utilise principalement pdfminer vs. d'autres outils d'OCR.

De plus, on partira dans un premier temps sur une transformation basique d'un document en texte, sans passer par une analyse de la localisation des textes sur le document.

Chapitre 9

Construction d'un modèle simple « ouvert »

Expliciter le principe de ce modèle avec un schéma simple.

Pas de mesure possible de la performance

9.1 Extraction des données

Ne garder que produits d'épicerie et boissons non alcoolisées

- 9.2 Conversion en blocs de texte
- 9.3 Train/Test split
- 9.4 Entraînement du modèle
- 9.5 Calcul de la similarité
- 9.6 Illustration des résultats obtenus

Chapitre 10

Utilisation des données manuellement étiquetées

Expliciter pourquoi on ne peut pas faire tourner (référence parties précédentes) sur l'ensemble des données

10.1 Chargement des données manuellement étiquetées

10.2 Train/Test split

10.3 Entraînement du modèle

10.4 Illustration des prédictions obtenues

Chapitre 11

Mesure de la performance

11.1 Précision

11.1.1 Approche naïve

11.1.2 Avec du « text-postprocessing »

11.2 Similarité cosinus

11.3 Fonction de *loss* spécifique

Expliciter les diverses distances, et pourquoi certaines sont plus pertinentes que d'autres.

Ex : on ne garde pas la distance de Hamming

11.3.1 Distance de Levenshtein

11.3.2 Distance de Damerau-Levenshtein

11.3.3 Distance de Jaro

11.3.4 Distance de Jaro-Winkler

11.4 Cross-validation des modèles précédents

11.4.1 Modèle « ouvert »

11.4.2 Modèle entraîné sur les données étiquetées manuellement

Chapitre 12

Transfer learning

12.1 Principe du pré-entraînement

Expliquer qu'il s'agit d'une approche hybride des 2 modèles précédents

12.2 Illustration de l'impact sur la performance

Chapitre 13

Hyperparameter tuning

13.1 Les paramètres ajustables

13.2 Application d'une grid search

Cinquième partie

Travaux subséquents

Chapitre 14

Opérationnalisation de cette maquette

14.1 Client et sponsor métier

14.2 Définition des règles de gestion

14.3 Mise en place d'une organisation projet

14.4 Industrialisation du code

Prochaines étapes : opérationnalisation via API
Documentation

14.5 Monitoring de la performance du modèle

Chapitre 15

Extension des fonctionnalités offertes

- 15.1 Prise en compte de nouveaux types de pièces jointes
- 15.2 Utilisation d'outil d'OCR pour les pdf non structurés
- 15.3 Mise en place d'outil de spatialisation des textes
- 15.4 Construction d'outils d'extraction de données connexes à la composition
- 15.5 Élargissement aux données nutritionnelles
- 15.6 Extraction « opportuniste » d'informations complémentaires
- 15.7 Évaluation de la performances sur d'autres familles de produits

Sixième partie

Figures, tableaux et bibliographie

Liste des tableaux

2.1	Exemple de tableau de données nutritionnelles	20
-----	---	----

Table des figures

1.1	Les flux métier avec les partenaires commerciaux	8
1.2	La répartition de l'activité des branches	13
1.3	Le maillage régional de la branche ÉpiSaveurs	15
2.1	La distinction entre produit et article	25
2.2	Le processus de création article	27
2.3	Une capture d'écran du PIM	29
2.4	Schéma de principe de la GDSN	30
2.5	L'uid d'un produit	31

Bibliographie

- [1] Conseil de l'Union Européenne. Règlement n°1907/2006 dit REACH, dec 2006. <https://bit.ly/2Jm05v9>.
- [2] Conseil de l'Union Européenne. Règlement n°1169/2011 dit INCO, nov 2011. https://www.senat.fr/europe/textes_europeens/ue0120.pdf.
- [3] 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>.
- [4] Direction Générales des Douanes et Droits Indirects. Notions essentielles sur la Déclaration d'Échanges de Biens. <https://www.douane.gouv.fr/notions-essentielles-sur-la-declaration-dechanges-de-biens>.
- [5] GS1. GDSN Trade Item Implementation Guide, nov 2019. https://www.gs1.org/docs/gdsn/tiig/3_1/GDSN_Trade_Item_Implementation_Guide.pdf.
- [6] GS1 France. Le réseau GDSN, le canal pour l'échange d'informations produits. <https://www.gs1.fr/Notre-offre/Le-reseau-GDSN-le-canal-pour-l-echange-d-informations-produits>.
- [7] GS1 Global. Global Data Synchronisation Network. <https://www.gs1.org/services/gdsn>.
- [8] Groupe Pomona. Site institutionnel du groupe pomona. <https://www.groupe-pomona.fr/>.

Septième partie

**Exemple de documents
fournisseur**

Annexe A

Fiches techniques

Annexe B

Étiquettes produit

Huitième 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