

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 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	12
2	La gestion de l'information produit	15
2.1	L'information produit	15
2.2	Le processus associé	15
2.3	Le PIM (Product Information Management)	15
II	Les données	16
3	Le périmètre produit	17
3.1	Accessibilité de la donnée en fonction des branches	17
3.2	Les branches déployées	17
3.3	Les types de produit	17
4	Les données utilisables	18
4.1	Données structurées	18

4.2	Données non structurées	18
4.3	Pièces jointes	18
4.3.1	Fiches techniques fournisseur	18
4.3.2	Étiquettes produit	18
4.3.3	Fiches logistiques fournisseur	18
4.3.4	Fiches techniques et argumentaires Pomona	18
4.4	Récapitulatif de la complétude des données	18
4.5	Analyse qualitative des données	19
4.6	Les données « manuellement étiquetées »	19
III	Les objectifs de ce projet	20
5	Les cas d’usage	21
5.1	Objectifs : Qualité et productivité	21
5.2	La préalimentation d’information	21
5.3	Le contrôle à la saisie fournisseur	21
5.4	L’aide aux vérifications Pomona	21
5.5	Les contrôles en masse asynchrones	21
6	Les types de données à récupérer	22
6.1	La composition produit	22
6.2	Les données nutritionnelles	22
6.3	Les données logistiques	22
7	Le choix du cas d’usage	23
7.1	Les multiples formats	23
7.2	Les informations « spatialisées »	23
7.3	La complexité dans la représentation des données logistiques . . .	23
7.4	La moindre représentation des étiquettes	23

IV	Construction du modèle	24
8	Les principes généraux	25
8.1	Contenu du texte d'une liste d'ingrédients	25
8.2	Limitation à l'identification des listes d'ingrédients	25
8.3	Conversion de documents en texte	26
9	Construction d'un modèle simple « ouvert »	27
9.1	Extraction des données	27
9.2	Conversion en blocs de texte	28
9.3	Train/Test split	28
9.4	Entraînement du modèle	28
9.5	Calcul de la similarité	28
9.6	Illustration des résultats obtenus	28
10	Utilisation des données manuellement étiquetées	29
10.1	Chargement des données manuellement étiquetées	29
10.2	Train/Test split	29
10.3	Entraînement du modèle	29
10.4	Illustration des prédictions obtenues	29
11	Mesure de la performance	30
11.1	Calcul de la précision	30
11.1.1	Approche naïve	30
11.1.2	Avec du « text-postprocessing »	30
11.2	Calcul d'une <i>loss</i>	30
11.2.1	Distance de Levenshtein	31
11.2.2	Distance de Damerau-Levenshtein	31
11.2.3	Distance de Jaro	31
11.2.4	Distance de Jaro-Winkler	31
11.3	Cross-validation des modèles précédents	31
11.3.1	Modèle « ouvert »	31

11.3.2	Modèle entraîné sur les données étiquetées manuellement	31
12	Transfer learning	32
12.1	Principe du pré-entraînement	32
12.2	Illustration de l'impact sur la performance	32
13	Hyperparameter tuning	33
13.1	Les paramètres ajustables	33
13.2	Application d'une grid search	33
V	Travaux subséquents	34
14	Opérationnalisation de cette maquette	35
14.1	Client et sponsor métier	35
14.2	Définition des règles de gestion	35
14.3	Mise en place d'une organisation projet	35
14.4	Industrialisation du code	35
14.5	Monitoring de la performance du modèle	35
15	Extension des fonctionnalités offertes	36
15.1	Prise en compte de nouveaux types de pièces jointes	37
15.2	Utilisation d'outil d'OCR pour les pdf non structurés	37
15.3	Mise en place d'outil de spatialisation des textes	37
15.4	Construction d'outils d'extraction de données connexes à la com- position	37
15.5	Élargissement aux données nutritionnelles	37
15.6	Extraction « opportuniste » d'informations complémentaires	37
15.7	Évaluation de la performances sur d'autres familles de produits	37

VI	Figures et tableaux	38
VII	Bibliographie	41
VIII	Exemple de documents fournisseur	43
A	Fiches techniques	44
B	Étiquettes produit	45
IX	Le code utilisé	46
C	Extraction de données du PIM	47
D	Conversion des pièces jointes en textes	48
E	Identification des listes d'ingrédients	49

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[3].

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

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

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

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

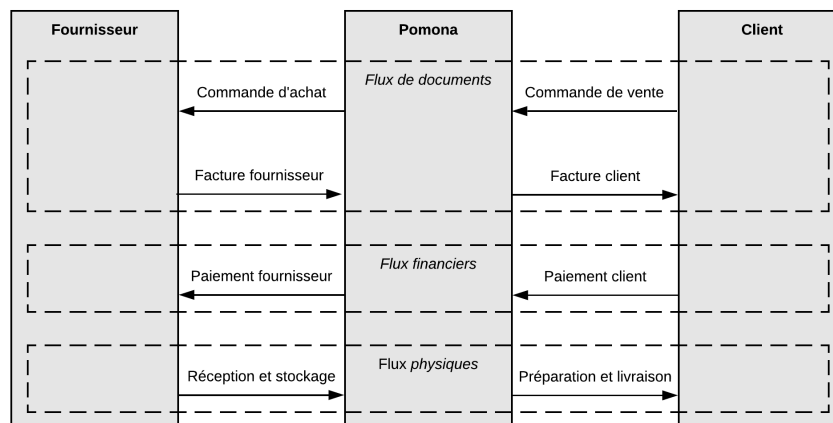


FIGURE 1.1 – Les flux métier avec les partenaires commerciaux

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. paragraphes suivants

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, ... 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 répartissent 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.

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

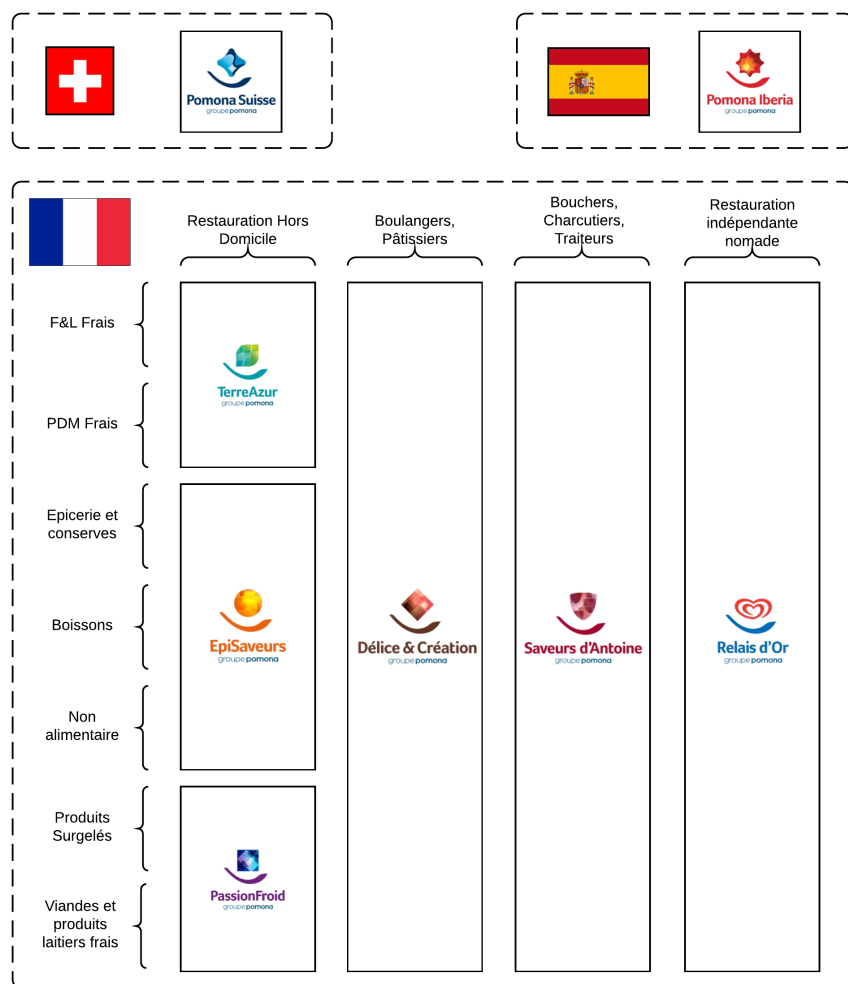


FIGURE 1.2 – La répartition de l'activité des branches



FIGURE 1.3 – Le maillage régional de la branche ÉpiSaveurs

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

Chapitre 4

Les données utilisables

4.1 Données structurées

4.2 Données non structurées

4.3 Pièces jointes

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

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 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

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

Parfois détaillé par phase, mais en général déconseillé.

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 Calcul de la précision

11.1.1 Approche naïve

11.1.2 Avec du « text-postprocessing »

11.2 Calcul d'une *loss*

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

Ex : on ne garde pas la distance de Hamming

11.2.1 Distance de Levenshtein

11.2.2 Distance de Damerau-Levenshtein

11.2.3 Distance de Jaro

11.2.4 Distance de Jaro-Winkler

11.3 Cross-validation des modèles précédents

11.3.1 Modèle « ouvert »

11.3.2 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 et tableaux

Liste des tableaux

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	14

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>.
- [3] Groupe Pomona. Site institutionnel du groupe pomona. <https://www.groupe-pomona.fr/>.

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