NutriXConso : recherche et appariement de données d'achats et de données nutritionnelles

Olivier de Mouzon[1] *et Valérie Orozco*[1]

Résumé : *Cet article présente l'outil « NutriXConso » développé au GREMAQ*[2] *afin d'enrichir des données existantes avec des données externes. Les données d'achats de produits de grande consommation ont ainsi été complétées par des informations nutritionnelles. « NutriXConso » se présente sous la forme d'une interface conviviale permettant d'effectuer des recherches de produits dans les deux jeux de données et d'effectuer ensuite les appariements de produits similaires. Cette interface est en lien avec une base de données qui est alimentée via les appariements effectués.*

Mots clés : Base de données, interrogation, appariement, données nutritionnelles, données d'achats

Abstract: *This article presents « NutriXConso », a tool we developed in order to add some new information to available data. Indeed, the food products' purchases datascan has been completed with nutritional data. « NutriXConso » is a user-friendly interface allowing to search product categories into the two databases and to match similar product categories. This interface is linked with a database which registers the matchings.*

Keywords: database, query, matchings, nutritional data, purchases data

Remerciements : Les auteurs remercient Céline Langlais, Roland Mialoundama, Christophe Bontemps et Céline Bonnet de leur forte contribution à ce projet. Les auteurs remercient également Roland Chartier, Marc Roze et Philippe Breucker pour leur aide technique. Enfin, les auteurs remercient Jean-Michel Cohen pour la mise à disposition d'une partie de ses données dans le cadre de ce projet. Ce projet a été financé par l'UMR GREMAQ-INRA.

Introduction

L'équipe Inra du GREMAQ a pour problématique de recherche l'étude des marchés et des industries agroalimentaires. Les objectifs sont multiples : il s'agit d'analyser et de comprendre les mécanismes économiques qui président aux choix des consommateurs et aux stratégies des firmes (par ex. développement des marques de distributeurs, impact des labels…). Un autre objectif est de donner un éclairage sur l'impact des politiques publiques sur les industries agroalimentaires et la distribution (par ex. impact d'une taxe sur les produits gras et sucrés).

[1] UMR 1291 GREMAQ-INRA, Groupe de Recherche en Economie Mathématique et Quantitative - F-31000 Toulouse. Contact : Valerie.Orozco@toulouse.inra.fr.

[2] GREMAQ : Groupe de Recherche en Economie Mathématique et Quantitative.

Un grand nombre des recherches appliquées réalisées dans l'unité utilise des données d'achats (données Kantar WorldPanel)[3]. Ces données concernent l'ensemble des produits alimentaires achetés par un panel de plus de 20 000 ménages, observés sur 11 années (1998-2008) et comprennent des informations sur les produits achetés (prix, marque, lieu d'achat, caractéristiques du produit…) et des informations sur les ménages (taille du ménage, âge, CSP,…). Ces données sont très riches, mais certains projets de recherche nécessitent de coupler ces données avec d'autres données externes.

Le projet « NutriXConso » concerne l'appariement des données Kantar avec des données nutritionnelles renseignant sur la composition des aliments. La section 1 présente le cadre général du projet et motive les choix qui sont présentés dans les sections suivantes. Les choix techniques sont détaillés à la section 2. La section 3 décrit les fonctionnalités d'aide à la visualisation des données mises en place pour faciliter le travail d'appariement. Et enfin, les fonctionnalités associées aux appariements sont présentées à la section 4.

1. Cadre de « NutriXConso »

1.1 Objectifs et spécifications de « NutriXConso »

Le projet « NutriXConso » répond à une demande de chercheurs et d'ingénieurs afin de créer de l'information additionnelle dans les données d'achats Kantar déjà disponibles. Un cahier des charges a été défini ensemble. L'objectif principal est d'ajouter un ensemble d'informations nutritionnelles (macro- et micronutriments) à chaque produit acheté (présent dans les données Kantar).
Tout d'abord, il nous a fallu rassembler des données nutritionnelles provenant de différentes sources, essentiellement Ciqual, Cohen Serog et internet.

Ensuite, il fallait apparier les 342 539 produits Kantar aux données nutritionnelles. Pour y parvenir, il s'agit de repérer les produits similaires dans les données Kantar et dans les données nutritionnelles (plus de 14 000 références disponibles) en fonction des caractéristiques des produits dans les deux jeux de données[4] et de les apparier ensuite (cf. **Figure 1**).

Etant donné le très grand nombre de lignes dans les différents jeux de données, et compte tenu du cahier des charges (chaque produit Kantar devait être apparié à un produit Ciqual, et éventuellement apparié aussi à un produit Cohen Serog), il était nécessaire d'avoir un outil permettant de rechercher et visualiser des sous-ensembles de produits cohérents.

[3] Anciennement appelées données TNS ou données Secodip, ces données sont largement utilisées au sein du Département Sciences Sociales, Agriculture et Alimentation, Espace et Environnement (SAE2), du Département de Mathématiques et Informatique Appliquées (MIA), et du Département Alimentation Humaine (AlimH) (équipes : Aliss (SAE2), Gremaq (SAE2), Lerna (SAE2), Mét@Risk (MIA, AlimH)).

[4] Dans toute la suite du document, nous utiliserons le mot « jeu » de données. Le premier jeu de données étant l'ensemble des données d'achats, le deuxième l'ensemble des données nutritionnelles (provenant de diverses sources).

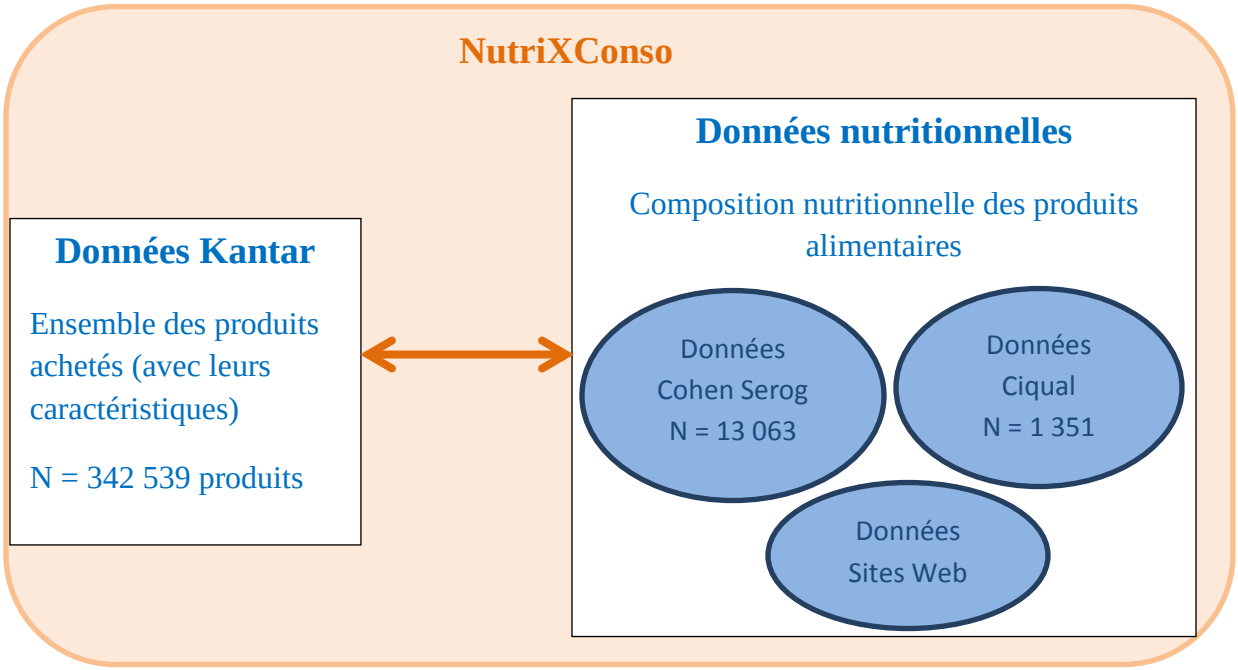


Figure 1 . *Appariement des deux jeux de données.*

De plus, l'outil devait permettre à une personne ayant des compétences nutritionnelles, mais non informaticienne, d'effectuer chaque appariement en quelques clics. Il devait également permettre des appariements groupés ou encore la création de moyennes nutritionnelles, aussi en quelques clics seulement.

Par ailleurs, dans un souci de traçabilité des appariements, chaque produit Kantar devait recevoir non seulement les données nutritionnelles, mais également la référence permettant de remonter à l'information source utilisée. De même, les actions importantes réalisées par l'utilisateur devaient être stockées (par ex. la création d'une moyenne de données nutritionnelles).
De plus, l'outil devait permettre de limiter les saisies à la main (et donc leurs erreurs inhérentes potentielles).

Le stockage de toutes les informations décrites ci-avant (données initiales, puis les appariements et principales actions à tracer) nous a bien sûr conduit à créer une base de données. L'outil « NutriXConso » a ensuite été conçu pour que ce travail d'appariement, à première vue répétitif et pénible, soit le plus aisé et agréable possible, en proposant une interface conviviale (entre la base de données et la personne qui effectue les appariements). Cette interface automatise un certain nombre de tâches, assure la traçabilité des traitements effectués et fait ainsi gagner du temps sur les appariements réalisés. L'outil « NutriXConso » va même au-delà d'une simple interface entre apparieur et base de données : non seulement il permet à une personne ne s'y connaissant pas en bases de données d'exécuter des requêtes (d'interrogation, d'insertion ou de modification) de la base de données, mais il propose également des fonctionnalités supplémentaires, permettant d'effectuer certains traitements directement sur les données extraites de la base, sans avoir à exécuter une nouvelle requête à chaque action de l'apparieur. La section

1.2 explique plus en détails la complexité du travail d'appariement et les implications au niveau des spécifications de l'outil.

Ainsi, « NutriXConso » est le nom du projet, mais c'est aussi le nom des deux piliers informatiques de ce projet : celui de l'outil créé ainsi que celui de la base de données croisant les données d'achats et les données de composition nutritionnelle.

1.2 Spécification détaillées de l'outil permettant un travail d'appariement complexe

Cette section présente la complexité du travail d'appariement et ses conséquences sur le développement de l'outil.

Ceci concerne tout particulièrement la phase de recherche de produits similaires dans les deux jeux de données. L'outil a notamment dû répondre aux difficultés suivantes :

- plusieurs nomenclatures peuvent exister pour définir une même catégorie de produits (par ex. « biscuits » ou « gâteaux »,…)
  → nécessité de recherche « souple », et possibilité d'avoir une recherche spécifique par jeu de données.

- à l'inverse, il peut aussi arriver qu'un même terme identifie deux catégories de produits différentes (par ex. le mot « pizza » va permettre d'identifier les pizzas, mais aussi les biscuits apéritifs en forme de mini-pizzas…)
  → nécessité de permettre une recherche précise et complexe (utilisation du SQL possible par l'utilisateur).

- la finesse de description du produit peut être différente dans les deux jeux de données impliquant des choix à effectuer dans les appariements (par ex. concernant les gâteaux au chocolat, dans les sources nutritionnelles la couleur du chocolat peut être connue : gâteaux chocolat noir, lait, blanc ; alors que dans Kantar non)
  → choix méthodologiques et développement de fonctionnalité (par ex. possibilité de créer des moyennes de produits à partir des produits du jeu de données nutritionnelles).

- certains produits Kantar peuvent n'être que partiellement identifiés (par ex. du lait sans mention « entier », « demi-écrémé » ou « écrémé » ou encore une pizza dont la composition ne serait pas précisée, etc.), alors que ces mentions existent sur d'autres produits Kantar
  → possibilité d'associer aux produits partiellement identifiés les moyennes pondérées par nombre d'achats des caractéristiques nutritionnelles des produits mieux identifiés (ce qui donne une autre façon de faire des moyennes de données nutritionnelles, mais en tenant compte de la probabilité d'acheter plutôt tel ou tel type de lait ou de pizza, par exemple).

- il est aussi possible qu'un produit Kantar ne soit pas présent dans les sources nutritionnelles
  → possibilité de rajouter à la main des informations provenant d'autres sources nutritionnelles.

Un autre aspect de la complexité du travail d'appariement provient directement des données très nombreuses et diverses, comme cela est présenté à la section 1.3.

1.3 Description des données

Cette section décrit plus en détails les principales données disponibles : Kantar, pour les achats (section 1.3.a), Ciqual et Cohen Serog, pour les données nutritionnelles (section 1.3.b).

1.3.a Données d'achats

Nous disposons des données Kantar WorldPanel qui recensent tous les actes d'achats effectués par plus de 20 000 ménages français représentatifs de la population française, sur une période assez longue (1998-2008) concernant près de 400 catégories de produits alimentaires.[5.]

Les données comprennent deux types informations : des informations sur les panélistes qui effectuent ces achats et des informations sur les produits achetés (prix, quantité achetée, marque, dénomination du produit, informations sur le conditionnement, arôme,…).

Chaque ligne d'un fichier d'achats correspond à l'ensemble des caractéristiques d'un achat effectué par un ménage. Le but du projet « NutriXConso » est d'ajouter sur chaque ligne la composition nutritionnelle du produit acheté.

Le fichier créé, utile pour le projet, correspond à l'ensemble des 342 539 produits distincts achetés sur la période 2001-2008 (Cf. **Tableau 1** pour un aperçu de ce fichier).[6]

Libellé produit	Marque[7]	Recette	Nombre d'occurrences sur 2001-2008
PIZZAS FRAICHES	M3	CHEVRE LARDON	10383
PIZZAS FRAICHES	M1	JAMBON FROMAGE	9873
PIZZAS FRAICHES	M1	JAMBON CHAMPIGNON	9338
PIZZAS FRAICHES	M1	CHORIZO	6058
PIZZAS FRAICHES	M2	CHEVRE LARDON	5555
PIZZAS FRAICHES	M3	3 FROMAGES	5421
PIZZAS FRAICHES	M1	CHEVRE LARDON	5208
PIZZAS FRAICHES	M1	JAMBON CHAMPIGNON	4710
……	….	……	……..

Tableau 1 . *Aperçu des données Kantar, exemple basé sur les achats de pizzas, N=1782 pizzas différentes (en termes de caractéristiques produits)*

[5] Le nombre de panélistes change sur la période puisque le panel est sans cesse renouvelé.

[6] Nous n'utilisons que les données à partir de 2001 car la dénomination des produits et les noms des variables ont changé en 2001. Pour constituer ce fichier, certaines variables qui n'ont aucun intérêt concernant la composition nutritionnelle ont été supprimées (conditionnement…).

[7] Par souci de confidentialité, le nom des marques n'est pas divulgué, et leur numéro ne correspond pas forcément à l'ordre de leurs parts de marchés.

1.3.b Données nutritionnelles

Il existe différentes sources de données relatives à la composition des aliments (Anses[8], livres, sites web, étiquettes des produits…) qui se distinguent selon les méthodes de recueil de données et la finesse des informations disponibles. Deux sources principales ont été utilisées :

- données Ciqual[9] (2008) : achetées auprès de l'Anses et disponibles sur http://www.afssa.fr/TableCIQUAL/. Elles concernent 1 351 aliments pour lesquels les informations relatives à 42 constituants sont disponibles. Les informations sont à un niveau assez agrégé des produits sauf pour l'eau pour laquelle les informations sont au niveau marque (Cf. **Tableau 2**).

- données Cohen-Serog : présentes dans le livre « Savoir manger, le guide des aliments 2008-2009 » de J.M. Cohen et P. Serog. Après avoir contacté un des auteurs en expliquant notre projet, ces données nous ont été transmises sous format numérique. Les informations sont au niveau marque (13 063 produits différents identifiés) et concernent majoritairement les macronutriments contenus dans les produits (Cf. **Tableau 3**).

Quelques sources complémentaires ont aussi été utilisées lorsque les deux premières sources ne permettaient pas de renseigner précisément la composition nutritionnelle d'un produit:

- site web idietetique (http://www.i-dietetique.com/)
- d'autres sites web variés (sites web des distributeurs, des marques…)

ORIGFDNM	Energie (kcal/100g)	Protéines (g/100g)	Glucides (g/100g)	Lipides (g/100g)	Sodium (mg/100g)	AG polyinsaturés (g/100g)	……
Pizza "spéciale"	229	11,3	23,4	9,98	597	1	
Pizza 4 fromages	235	11,3	24,8	10,1	426	1,2	
Pizza 4 saisons	170	7,08	20,4	6,62	361	1,1	
Pizza fromage	220	10,5	23,6	9,32	544	1,66	
Pizza jambon fromage	212	10,1	27,1	7,06	665	1,12	
Pizza jambon fromage champignons ou Pizza royale	203	10,4	24,8	6,87	466	0,5	
Pizza jambon fromage champignons ou Pizza royale, surgelée	208	9,06	26,6	7,26	560	1,12	

Tableau 2. *Aperçu des données Ciqual, exemple basé sur la composition nutritionnelle des pizzas, N=7 pizzas différentes (en termes de caractéristiques produits et nutritionnelles)*

[8] Agence Nationale chargée de la Sécurité Sanitaire de l'alimentation, de l'environnement et du travail.

[9] Centre d'Information sur la Qualité des ALiments, équipe de l'Anses.

Produits	Rayon	Marque	Pour 100g						
			Calories (kcal)	Protéines (g)	Lipides (g)	Glucides (g)	Sucres (g)	Fibres (g)	Sodium (g)
Pizza Provençale	Surgelé	M6	150	5,5	4,5	23			
Pizza margherita	Frais	M12	166	7,8	4,1	24,5			
Pizza royale [from, jamb, champi] cuite sur pierre	Surgelé	M7	177	10	5	23			
Pizza 4 shagioni	Surgelé	M8	181	8,5	4,8	26	2,4	1,6	0,5
Pizza 4 saisons cuite sur Pierre	Surgelé	M9	181	8,5	4,8	26	2,4	0,5	0,6
Pizza 4 saisons	Surgelé	M9	181	8,5	4,8	26	2,4	1,6	0,5
Pizza Kebab	Surgelé	M10	182	8,8	5,8	23,6	2,2	3,0	0,6
Pizza'O Printemps jambon Artichaut Tomate	Frais	M1	183	8,7	3,4	29,5			
……	……	……	……	……	……	……	……	….	….

Tableau 3. *Aperçu des données Cohen Serog, exemple basé sur la composition nutritionnelle des pizzas, N=309 pizzas différentes (en termes de caractéristiques produits et nutritionnelles)*

2. Développement de l'outil « NutriXConso »

L'outil développé a permis d'interroger et d'alimenter la base de données composée d'une table Kantar et des tables nutritionnelles. Il se présente sous la forme d'une interface d'affichage des données (résultats de requêtes), avec des fonctionnalités intuitives permettant d'affiner et d'améliorer l'affichage de ces recherches, et d'effectuer les appariements des deux jeux de données.

La partie d'extraction de données de la base vers l'interface et de manipulation de ces données dans l'interface est présentée en section 3. La partie traitant de l'appariement via l'interface (qui modifie en retour la base de données) est présentée section 4.

Ici, nous présentons succinctement les 2 piliers de « NutriXConso » : la base de données et l'interface.

2.1 La base de données

La base de données se compose de 14 tables, dont le détail complet est donné en Annexe 1.
Nous avons choisi MySQL comme système de gestion de bases de données, ce qui permet de se placer dans un environnement libre.

Pour chacune des 4 sources nutritionnelles (Ciqual, Cohen Serog, idietetique, Autre), il y en en effet 3 tables (soit 12 en tout) :

- une table (par exemple « ciqual ») qui contient les données nutritionnelles de la source : les données initiales, ainsi que les moyennes qui y ont été ajoutées (via l'interface) ;
- une table (par exemple « moyenne_ciqual ») qui permet de savoir de quelles lignes initiales chaque moyenne provient ;
- une table (par exemple « decision_ciqual ») qui permet de savoir quels éléments Kantar étaient affichés dans l'interface lors de l'appariement.

Ensuite, la table « kantar » contient les différents produits Kantar à apparier. En plus des données initiales de Kantar, cette table est complétée par les appariements faits entre un produit Kantar et une source nutritionnelle donnée. Quatre champs initialement vides (« id_ciqual », « id_cohen_serog », « id_idietetique » et « id_autre ») sont modifiés lors des appariements à partir de l'interface graphique.

Enfin une dernière table (« ponderation ») permet, à l'instar des tables moyennes, de définir les références Kantar pour lesquelles les données nutritionnelles seront issues des moyennes pondérées par le nombre d'achats d'autres références Kantar.

Par ailleurs, des variables « vmin » et « vall » ont été rajoutées aux 5 tables qui contiennent les données Kantar et les données nutritionnelles, afin de faciliter les interrogations complexes lors de l'extraction des données de la base vers l'interface. Ainsi, dans Kantar, « vmin » est la concaténation de 43 variables contenant des informations sur la catégorie de produits, et « vall » est la concaténation de toutes les variables permettant d'identifier un produit. Dans les sources nutritionnelles, ces 2 variables sont identiques et sont la concaténation de toutes les variables permettant d'identifier un produit. Cela permet d'écrire une même requête d'interrogation qui sera appliquée sur les 5 tables (cf. section 3 et tout particulièrement 3.1.a, 3.1.c et 3.2.a).

L'alimentation initiale de la base a été effectuée à partir des fichiers alpha-numériques disponibles pour chaque source, dans des formats différents et via importation dans MySQL.
Mais, dans les quelques cas de produits Kantar où les informations nutritionnelles contenus dans les tables Ciqual et Cohen Serog se sont avérées insuffisantes, une table « Autre » ou « Idietetique » a été alimentée manuellement dans la base de données, à partir de sources web et en utilisant l'interface graphique fournie par phpMyAdmin (cf. **Figure 2**).

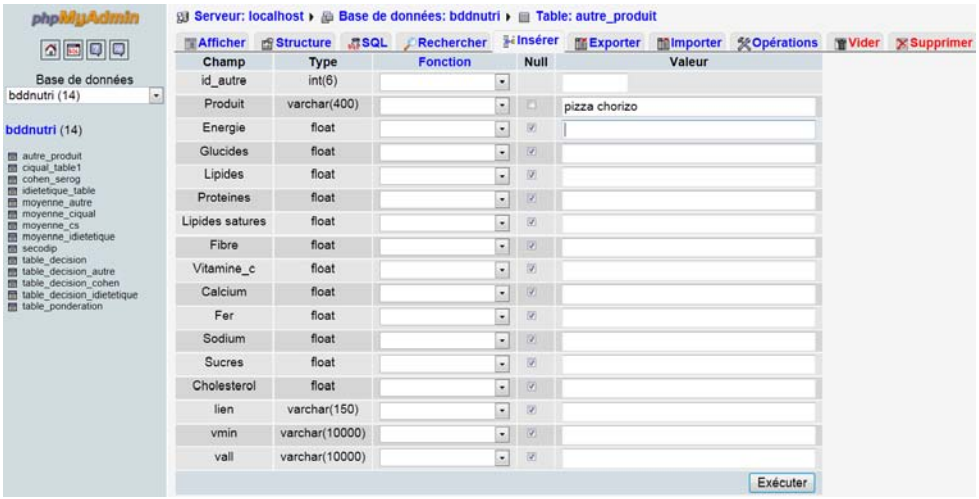


Figure 2 . *Alimentation de la table « Autre ».*

Les appariements s'effectuent ensuite normalement, dans l'interface « NutriXConso », une fois les nouvelles données chargées.

2.2 L'interface

L'interface graphique de l'outil « NutriXConso » est une interface web qui permet de réaliser toutes les opérations (extraction, appariement) sur la base de données sans avoir besoin de connaître SQL (si ce n'est quelques rudiments pour les extractions les plus complexes). Elle offre également des outils supplémentaires pour aider à visualiser les lignes à apparier (cf. section 3). Toutes ces fonctionnalités sont écrites en PHP (HTML et JavaScript).

L'outil « NutriXConso » permet ainsi l'interrogation des jeux de données via des requêtes SQL ainsi que via des requêtes dynamiques en Ajax (en utilisant le module jQuery).
Les requêtes dynamiques permettent d'affiner le résultat d'une première requête SQL en filtrant l'ensemble des données chargées à l'écran selon des mots souhaités. Ces dernières sont réalisées instantanément. Les lignes retournées peuvent également être triées par ordre croissant (ou décroissant) ou alphabétique d'une des variables affichées.

Les fonctionnalités développées pour interroger les données nutritionnelles sont généralisées à l'ensemble des sources (i.e. elles fonctionnent aussi bien pour les données Ciqual que pour les données Cohen Serog même si les variables diffèrent).

Par ailleurs, l'outil utilise des bibliothèques qui nécessitent de passer par Internet Explorer pour éviter des problèmes de compatibilité.

Au final, l'outil est donc à la fois une interface de recherche, une interface de visualisation, et une interface d'appariement entre deux jeux de données.

2.3 Implémentation et environnement de l'outil

« NutriXConso » (l'ensemble base de données et interface) a été mis en place sur un serveur pour des questions de rapidité d'exécution et d'accès à distance. Le serveur est en environnement Windows. Nous y avons installé WampServer qui offre une bonne intégration pour une base de données MySQL et une interface PHP.

Etant donné que le projet était un projet assez long (plus de 2 mois étaient nécessaires pour réaliser l'ensemble des appariements) et par souci de sûreté, nous avons mis en place une sauvegarde automatique quotidienne de la base de données, d'une part sur le disque dur du serveur et d'autre part sur un autre disque réseau (qui lui disposait d'une sauvegarde et sécurité en RAID 5). Pour cela nous avons utilisé le planificateur de tâches qui lançait chaque nuit un fichier batch faisant appel à la commande mysqldump correctement paramétrée.

3. « NutriXConso » : Une interface de recherche et de visualisation

L'interface développée a permis de rechercher et de visualiser des produits présents dans les deux jeux de données en vue des appariements.

3.1 Présentation de l'interface et de ses fonctionnalités basiques

Nous présentons ici les fonctionnalités simples (dans leur version basique) :

- La section 3.1.a montre l'extraction de données de la base et leur accès via l'interface.

- La section 3.1.b présente la sous-sélection dynamique sur les données déjà extraites.
- La section 3.1.c fait une synthèse des possibilités de sélection des données.
- Les sections 3.1.d à 3.1.f donnent quelques éléments simples supplémentaires d'aide à la visualisation.

3.1.a Extraction des données de la base vers l'interface

L'interface de recherche permet d'effectuer des recherches et de visualiser l'ensemble des produits (et leurs caractéristiques) présents dans les deux jeux de données.
La recherche d'un produit s'effectue via différents champs à remplir (cf. **Figure 3** ci-dessous).

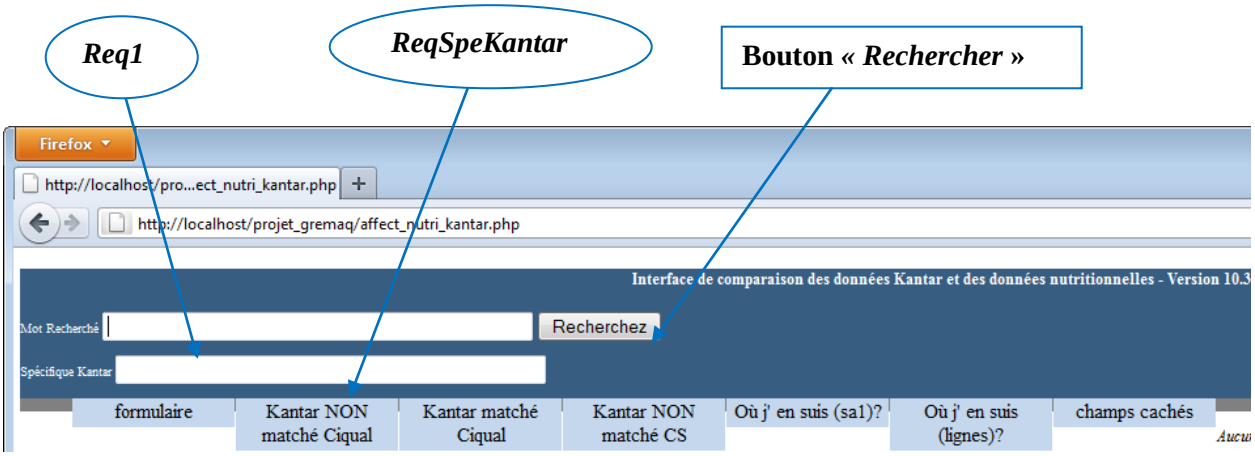


Figure 3 . *« NutriXConso », une interface de recherche.*

Le champ *« Req1 »* est un champ général, recherché à la fois dans les données Kantar et dans les données nutritionnelles. Il s'agit généralement de la catégorie de produit recherchée.
Comme les données Kantar disposent d'un très grand nombre de variables (caractéristiques des produits achetés), nous avons souhaité pouvoir affiner la requête en autorisant des champs supplémentaires spécifiques aux données Kantar. Le champ *« ReqSpeKantar »* sera ainsi recherché dans les données Kantar uniquement.

Après lancement de la recherche (bouton *« Rechercher »* cf. **Figure 3**), les produits correspondants s'affichent (cf. **Figure 4** et **Figure 5** ci-dessous) :

- A gauche : les produits Kantar correspondant à la recherche
- A droite : les informations nutritionnelles disponibles pour les produits recherchés. Plusieurs sources sont disponibles : Ciqual, Cohen Serog, idietetique et Autres. Les données des différentes sources sont visibles en cliquant sur l'onglet souhaité.

Remarque : le programme php permet de n'afficher que les variables non vides pour la recherche en cours (ceci est surtout utile pour les données Kantar pour lesquelles les caractéristiques disponibles des produits varient selon le produit).

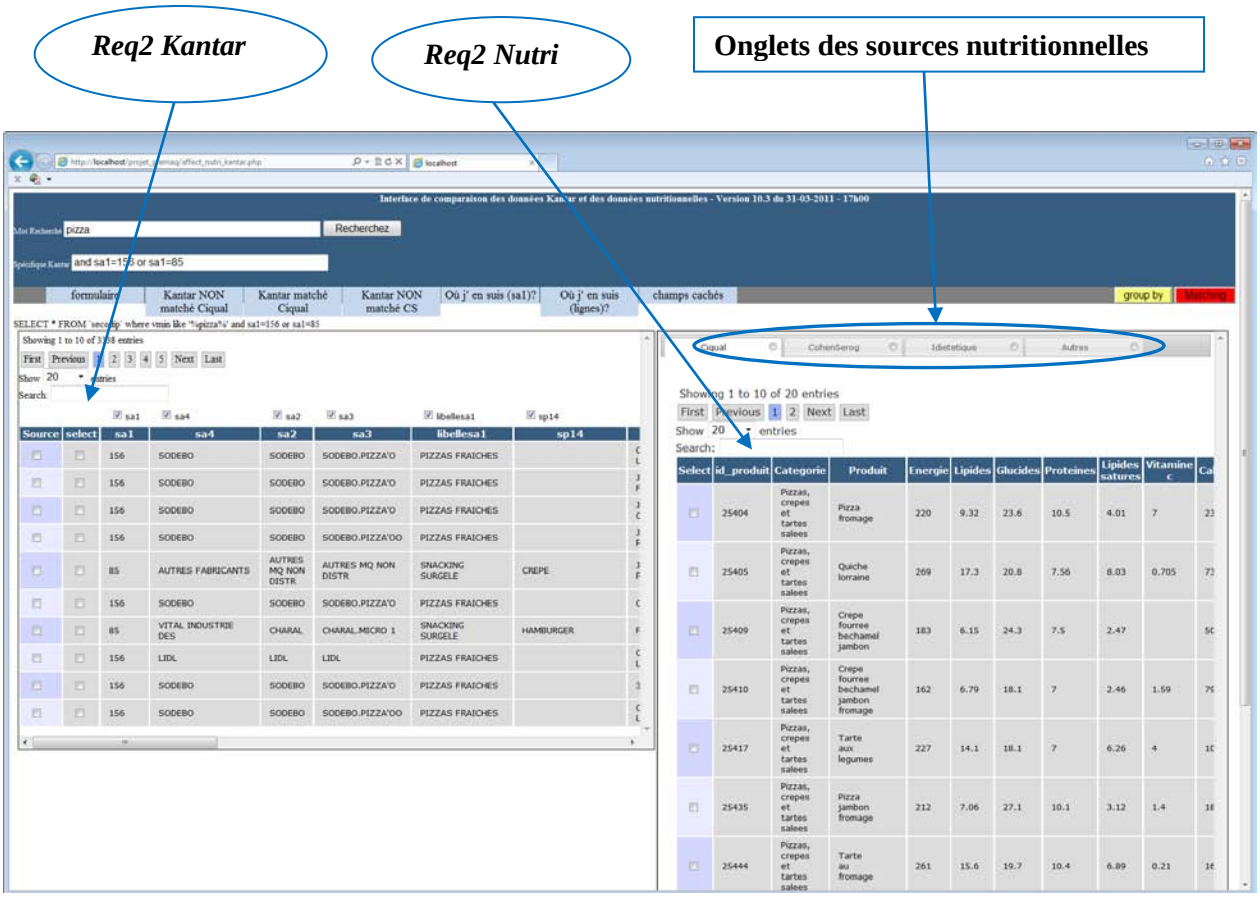


Figure 4 . *Résultat de recherche dans (Kantar, Ciqual).*

Si l'utilisateur clique sur l'onglet « *CohenSerog* », l'affichage des données nutritionnelles (à droite) correspondra aux données Cohen Serog (**cf. Figure 5**).

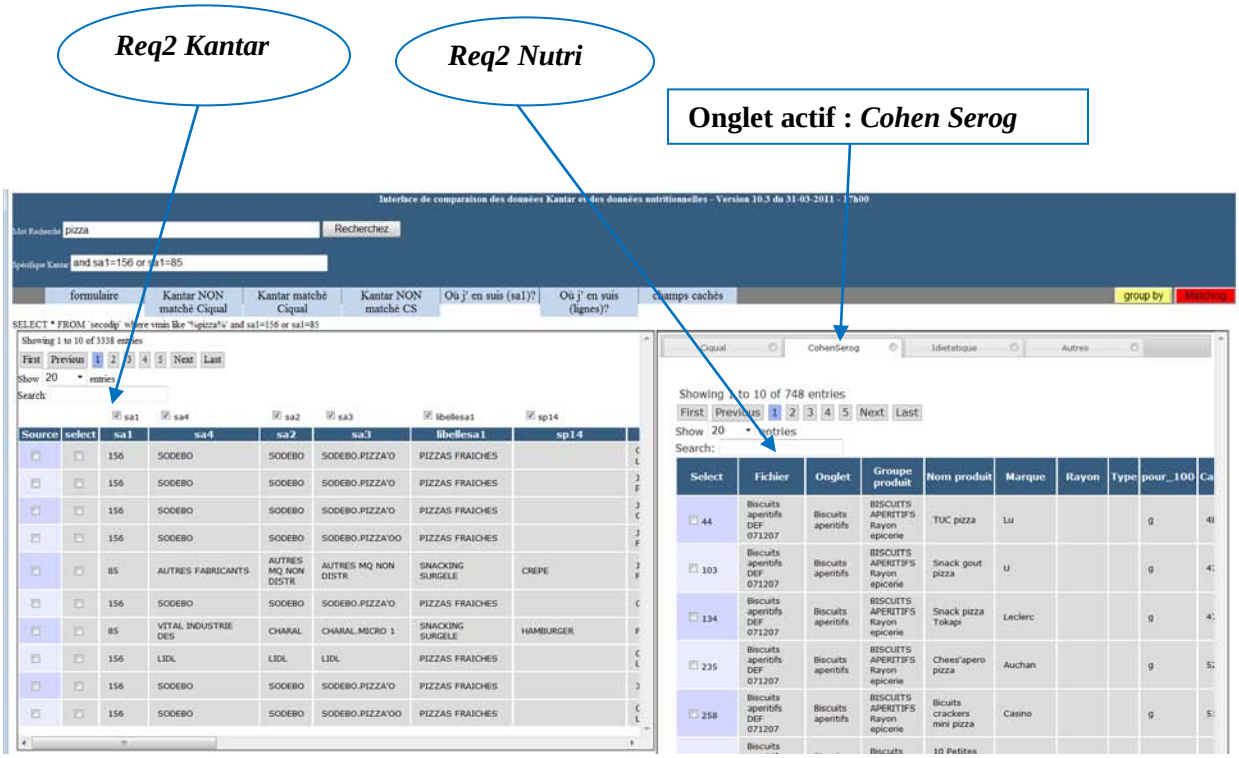


Figure 5 . *Résultat de recherche dans (Kantar, Cohen Serog).*

3.1.b Les requêtes dynamiques sur l'interface

De nouveaux champs de recherche apparaissent aussi : *« Req2Kantar »* et *« Req2Nutri »* (cf. **Figure 4** et **Figure 5** ci-dessus). Ce sont des champs spécifiques permettant d'affiner la recherche selon les caractéristiques des produits (respectivement dans les données d'achats ou les données nutritionnelles). *« Req2Kantar »* sera ainsi recherché dans les données Kantar uniquement alors que le champ *« Req2Nutri »* sera uniquement recherché dans les données nutritionnelles. Il s'agit de requêtes dynamiques qui interrogent instantanément l'ensemble des données affichées à l'écran (issues de la première requête, « *Req1* » et « *ReqSpeKantar* »). Elles sont plus rapides que les requêtes « *Req1* » et « *ReqSpeKantar* » qui **interrogent** la base de données.

L'exemple relatif aux pizzas (cf. **Figure 6** et **Figure 7**), montre comment fonctionnent ces requêtes. Dans la Figure 6, « *Req2Kantar* » vaut « chev la » et le résultat instantané qui apparait contient des pizzas « chèvre lardons », mais aussi des pizzas « chèvre, tomate classiques » ou « chèvre classiques ».

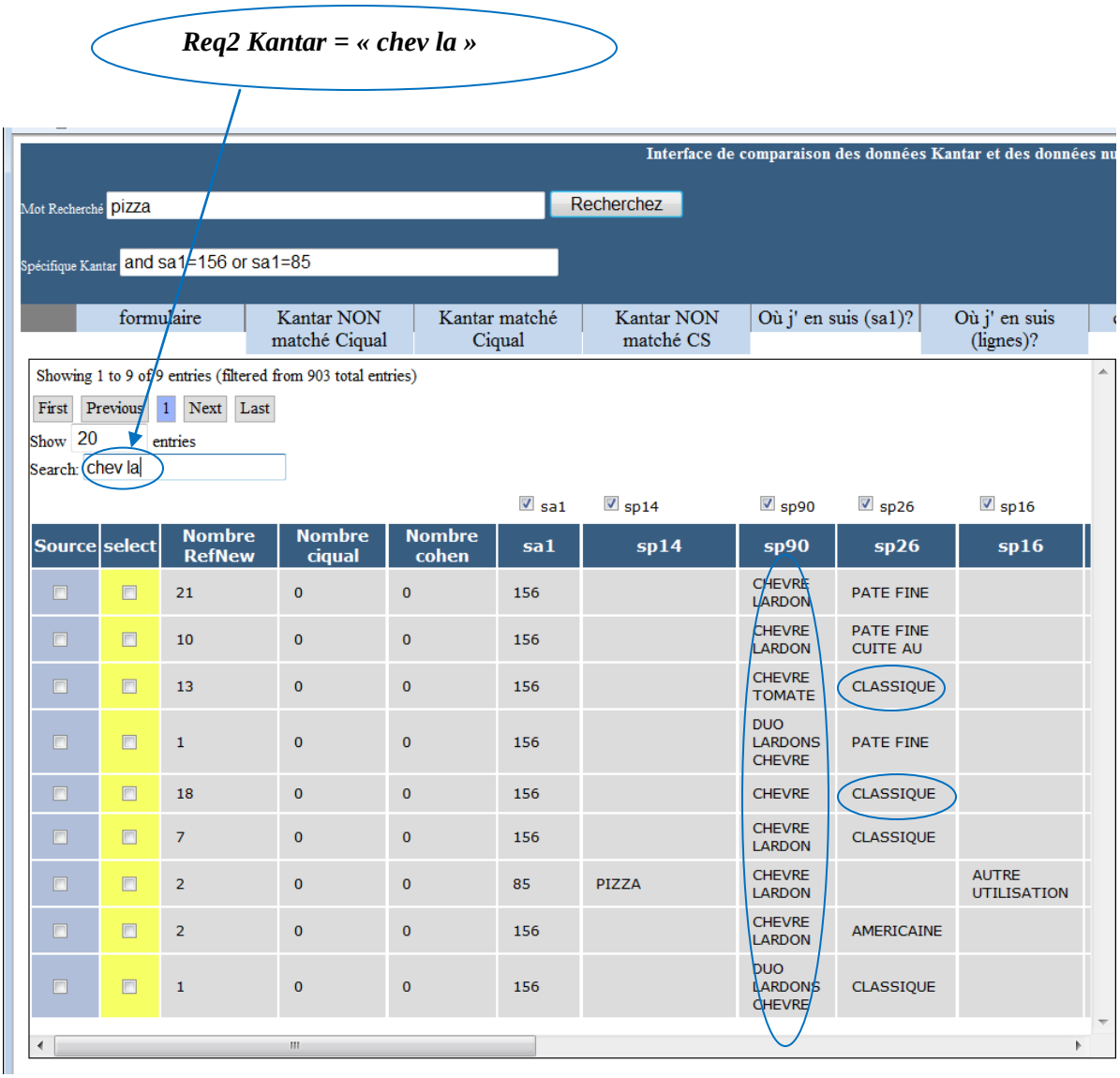


Figure 6 . *Résultat de recherche issue de requêtes dynamiques (données Kantar).*

Lorsqu'on complète le champ « *Req2Kantar* » en rajoutant un « r », les pizzas classiques disparaissent et seules les pizzas « chèvres lardons » sont affichées (cf. **Figure 7)**.

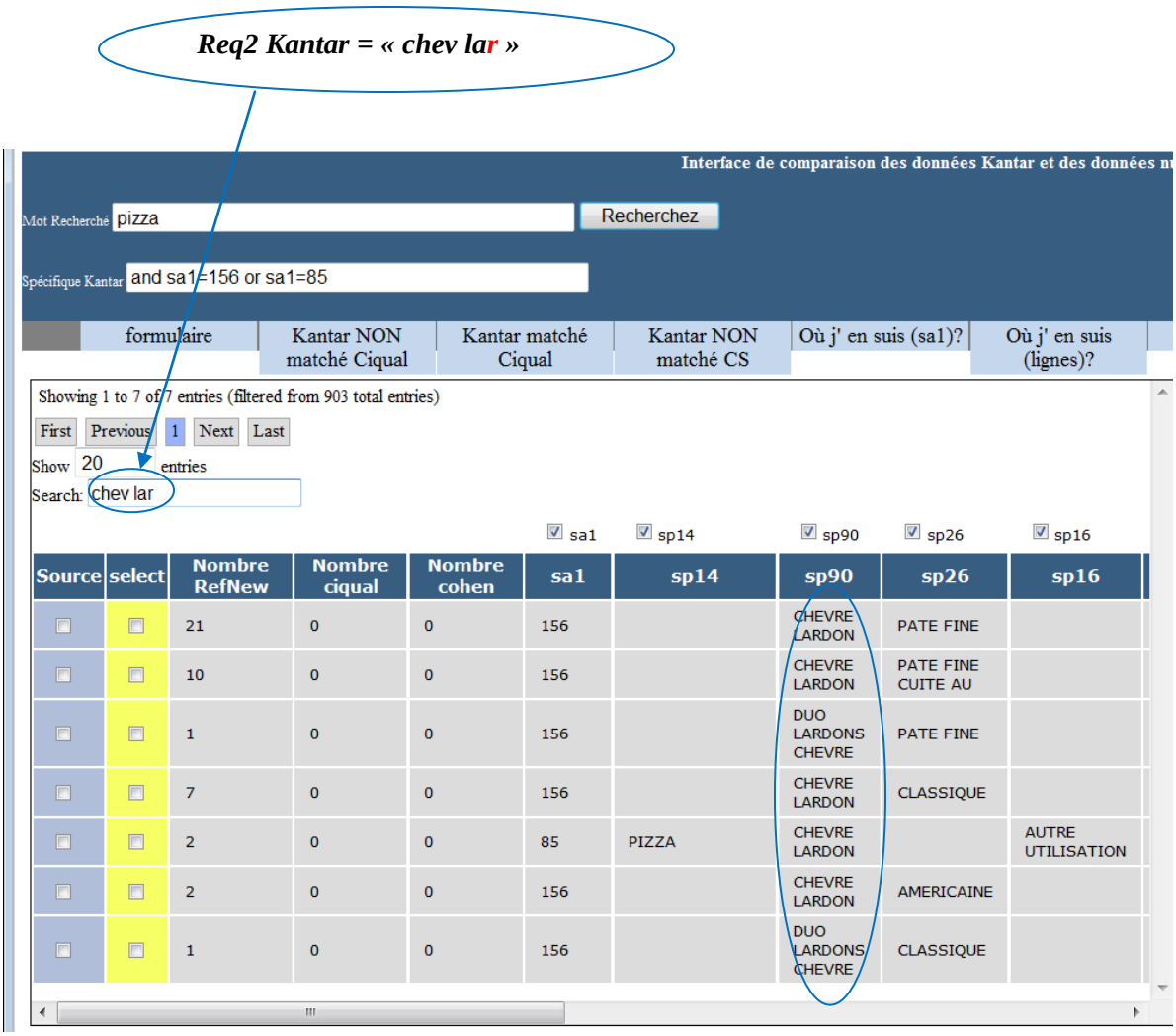


Figure 7. *Résultat de recherche issue de requêtes dynamiques (données Kantar).*

Si l'utilisateur change d'avis et remet le champ « *Req2Kantar* » à vide, on retrouve instantanément l'ensemble du résultat de la requête basée sur « *Req1* » et « *ReqSpeKantar* ».
« *Req2Nutri* » fonctionne exactement de la même façon mais agit sur les données nutritionnelles (cf. **Figure 8** et **Figure 9**). Il est d'ailleurs propre à chacune des sources nutritionnelles.

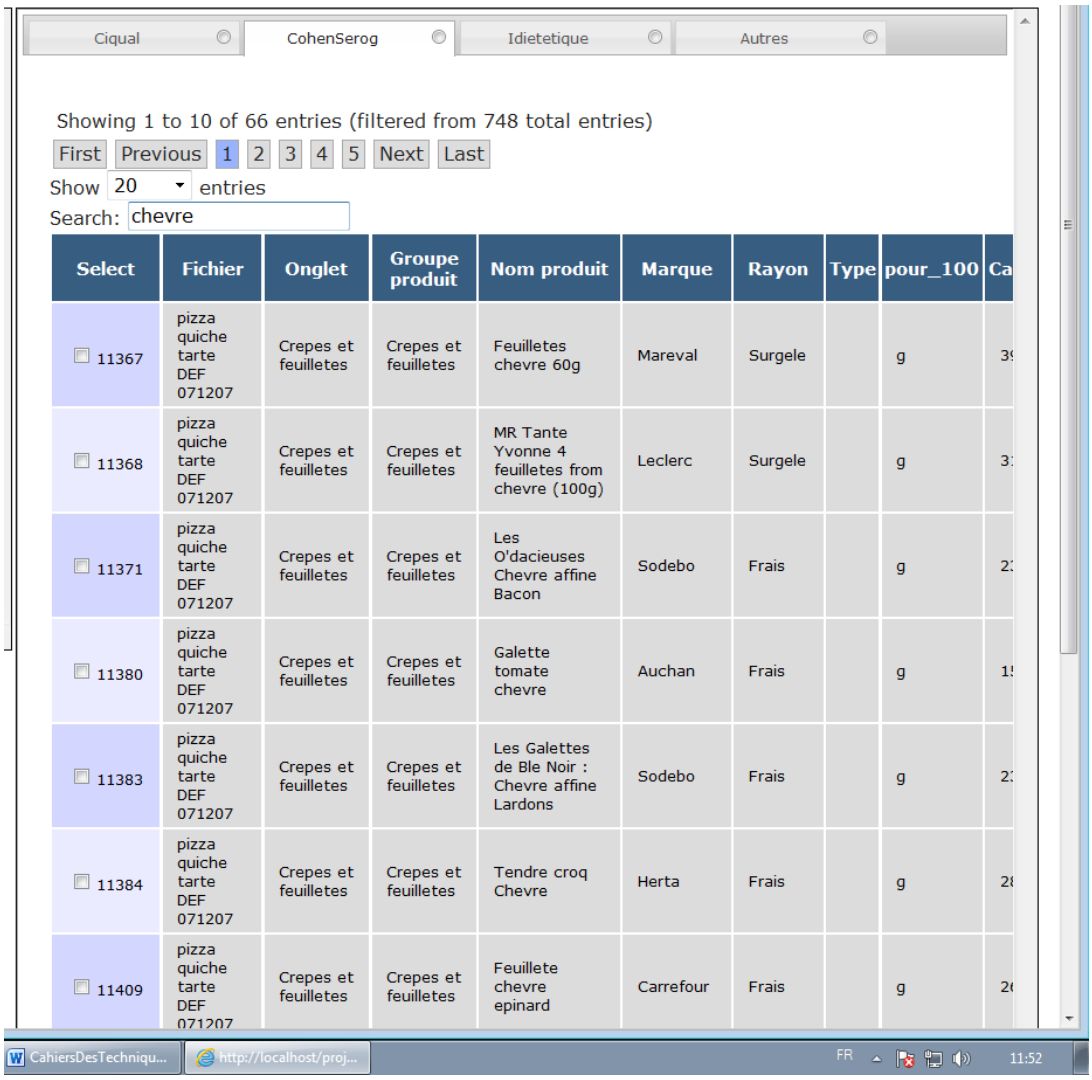


Figure 8. *Résultat de recherche issue de requêtes dynamiques (données Cohen Serog).*

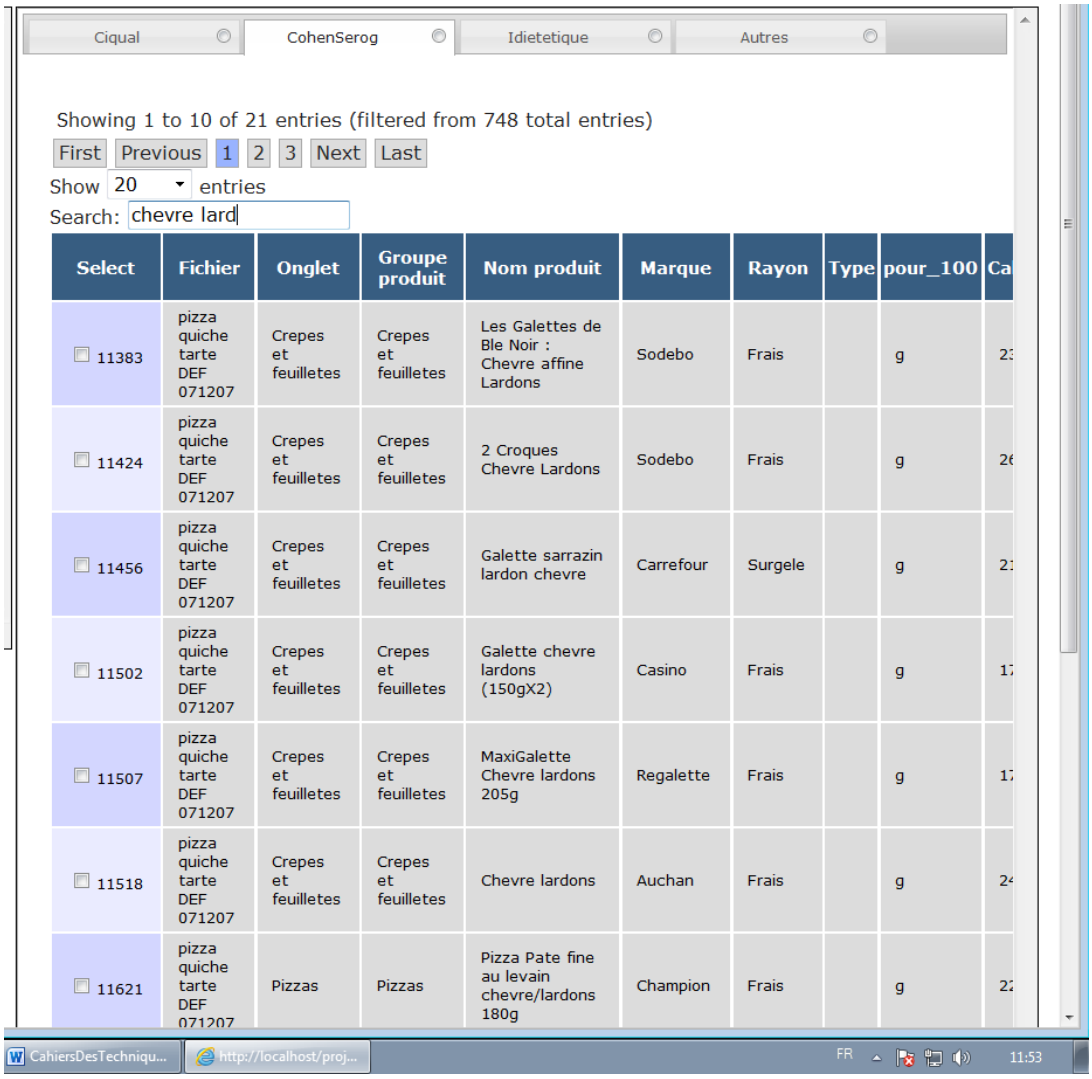


Figure 9. *Résultat de recherche issue de requêtes dynamiques (données Cohen Serog).*

3.1.c Propriétés des champs de recherche

Les différents champs de recherche (« *Req1* », « *ReqSpeKantar* », « *Req2Kantar* » et « *Req2Nutri* ») ont des propriétés différentes selon les jeux de données, les variables qu'ils interrogent, et selon la nature de la recherche. Le **Tableau 4** et le **Tableau 5** résument les différences entre ces 4 champs.

Requête	Agit sur :		Niveau de la recherche
	Kantar	les sources nutritionnelles	
*Req1**	X	X	Base de données
*ReqSpeKantar***	X		
*Req2 Kantar****	X		Données déjà extraites
*Req2 Nutri****		X	

***UNE** simple chaîne de caractères ou une requête codée en SQL (voir 3.2) ; **requête codée en SQL ;
***Requête dynamique (un ou plusieurs mots ou morceaux de mots)
Dans tous les cas, les mots devront être écrits SANS accent.

Tableau 4 . *Jeux de données interrogés selon les différents champs de recherche*

Requête	Agit sur :				Nature de la recherche
	vmin	vall	une variable précise	toutes les variables présentes à l'écran	
Req1	X (par défaut)	X*			SQL
ReqSpeKantar	X	X	X (par ex. sp90)		
Req2 Kantar (dynamique)				X (côté Kantar)	jQuery
Req2 Nutri (dynamique)				X (côté source nutritionnelle)	

*Voir aussi en 3.3 l'utilisation des expressions régulières utilisant « vall ».

Tableau 5 . *Variables interrogées selon les différents champs de recherche*

Par défaut, après toute nouvelle recherche basée sur « *Req1* » ou « *ReqSpeKantar* », la partie droite de l'interface correspond aux données Ciqual. Si l'utilisateur souhaite travailler sur une autre source nutritionnelle, il a la possibilité de cocher l'onglet de la source nutritionnelle en cours (par exemple Cohen Serog ; cf. **Figure 10**) afin de revenir systématiquement sur cette source :

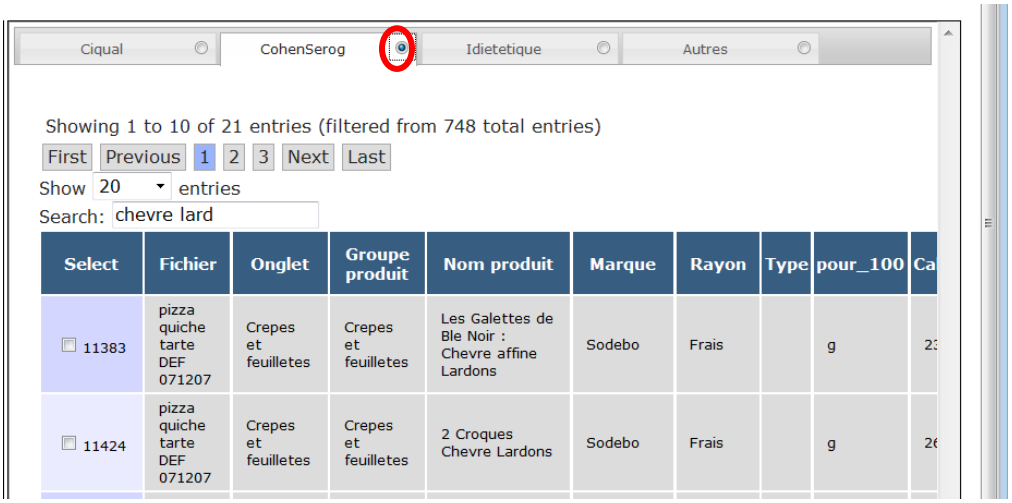


Figure 10 . *Onglet de la source nutritionnelle en cours.*

3.1.d Aide à la visualisation : le nombre de lignes par page

Afin de pouvoir adapter la visualisation des données extraites à différentes tailles d'écran et au besoin de visualisation, un menu déroulant propose le nombre de lignes par page souhaité (10 par défaut, comme sur la Figure 11). Ce choix est propre au jeu de données (Kantar ou nutritionnelle).

Figure 11 . *Choix du nombre de lignes affichées par page.*

3.1.e Aide à la visualisation : surlignage d'une ligne

En raison du grand nombre de colonnes, notamment dans les données Kantar, il peut s'avérer difficile de suivre une ligne jusqu'au bout (en faisant bouger l'ascenseur horizontal correspondant). C'est pourquoi nous avons inclus une fonctionnalité qui permet de distinguer une ou plusieurs lignes : un simple clic sur une ligne permet de basculer l'affichage en gras et italique. Un deuxième clic sur la même ligne rétablit l'affichage classique.

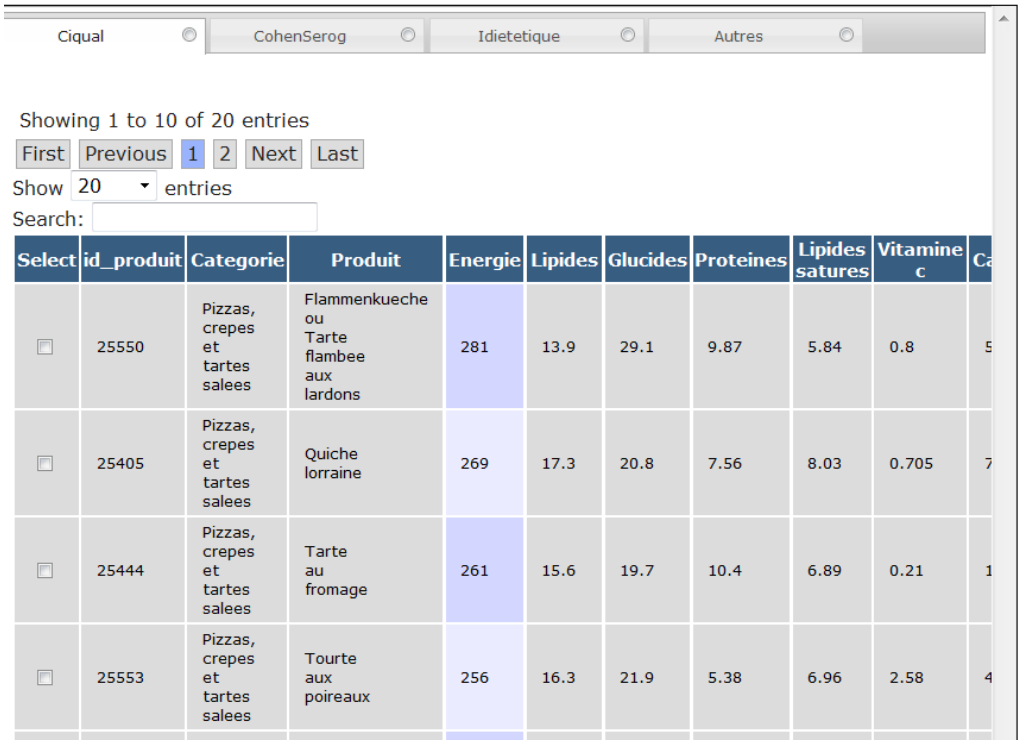
La **Figure 12** donne un exemple de 2 lignes mises en valeur par rapport à 3 autres. Cette fonctionnalité, illustrée ici sur l'onglet Cohen Serog, est disponible sur les deux jeux de données.

Select	Fichier	Onglet	Groupe produit	Nom produit	Marque	Rayon	Type	pour_100	Calo
☐ 11383	pizza quiche tarte DEF 071207	Crepes et feuilletes	Crepes et feuilletes	Les Galettes de Ble Noir : Chevre affine Lardons	Sodebo	Frais		g	233
☐ 11424	pizza quiche tarte DEF 071207	Crepes et feuilletes	Crepes et feuilletes	2 Croques Chevre Lardons	Sodebo	Frais		g	265
☐ ***11456***	***pizza quiche tarte DEF 071207***	***Crepes et feuilletes***	***Crepes et feuilletes***	***Galette sarrazin lardon chevre***	***Carrefour***	***Surgele***		***g***	***211***
☐ 11502	pizza quiche tarte DEF 071207	Crepes et feuilletes	Crepes et feuilletes	Galette chevre lardons (150gX2)	Casino	Frais		g	174
☐ ***11507***	***pizza quiche tarte DEF 071207***	***Crepes et feuilletes***	***Crepes et feuilletes***	***MaxiGalette Chevre lardons 205g***	***Regalette***	***Frais***		***g***	***171***

Figure 12 . *Exemple de surlignage dynamique effectué sur les données Cohen Serog.*

3.1.f Aide à la visualisation : le tri dynamique

Il arrive souvent que l'utilisateur ait envie de trier les données selon une caractéristique de son choix (caractéristiques du produit pour le jeu de données Kantar, caractéristiques produit ou nutritionnelles pour le jeu des données nutritionnelles). Un simple clic sur une colonne permet de la trier (ordre croissant/décroissant ou alphabétique selon la nature de la variable). Ce tri est donc un tri dynamique à partir des données chargées sur l'interface. La **Figure 13** présente un exemple d'utilisation de ce tri à partir des données Ciqual, sur la colonne « Energie », triée par ordre décroissant. La colonne triée apparaît ensuite dans des tons violet.



Select	id_produit	Categorie	Produit	Energie	Lipides	Glucides	Proteines	Lipides satures	Vitamine c	Ca
☐	25550	Pizzas, crepes et tartes salees	Flammenkueche ou Tarte flambee aux lardons	281	13.9	29.1	9.87	5.84	0.8	5
☐	25405	Pizzas, crepes et tartes salees	Quiche lorraine	269	17.3	20.8	7.56	8.03	0.705	7
☐	25444	Pizzas, crepes et tartes salees	Tarte au fromage	261	15.6	19.7	10.4	6.89	0.21	1
☐	25553	Pizzas, crepes et tartes salees	Tourte aux poireaux	256	16.3	21.9	5.38	6.96	2.58	4

Figure 13 . *Exemple de tri dynamique effectué sur les données Ciqual.*

3.2 Fonctionnalités avancées de « NutriXConso »

Souvent le résultat d'une recherche comprend trop de lignes et la visibilité est réduite. Par exemple, la recherche du mot « pizza » dans les achats Kantar (« *Req1* »= « *pizza* ») renvoie un résultat de 1 856 lignes. Cela s'explique par le grand nombre de caractéristiques des produits (marque, appellation, type de pizzas, ingrédients majoritaires, type de pâte,…).

Afin d'obtenir une présentation plus parlante, deux fonctionnalités avancées ont été implémentées. La première consiste à sélectionner seulement une partie des données en ajoutant des conditions à l'extraction. Cela permet de réduire le nombre de données extraites, donc les temps de réponse, et aussi de mieux visualiser ces résultats plus ciblés. La section 3.2 explique comment utiliser « *Req1* » et « *ReqSpeKantar* » pour passer d'une recherche « simple chaîne de caractères » à une recherche plus évoluée.

La deuxième fonctionnalité avancée permet de grouper un certain nombre de lignes ensemble, afin de réaliser les mêmes appariements sur plusieurs lignes en même temps. Ceci est tout à fait indiqué lorsque la source nutritionnelle comporte moins de détails que Kantar (par ex. la marque est connue dans Kantar, mais pas dans Ciqual et on peut donc masquer la variable marque pour ces appariements). C'est ce que présente la section 3.2.b plus en détails.

3.2.a Expressions régulières

Il est parfois utile d'aller plus loin qu'une recherche d'une simple chaîne de caractères, lors de l'extraction des données de la base. L'utilisation du langage SQL est dans ce cas très utile (cf. Encadré 1 pour quelques rappels sur le langage SQL).

SQL (Structured Query Language) est un langage de programmation informatique normalisé destiné à effectuer des opérations sur des bases de données. Il permet notamment de rechercher, d'ajouter, de modifier ou de supprimer des données dans les bases de données.

Par exemple, écrire **:** **SELECT * FROM `kantar` where vmin like '%lait%'**

Signifie « Recherche et affiche toutes les lignes Kantar pour lesquelles la variable « vmin » comprend le mot « lait ».

Les mots clefs du langage SQL sont les suivants :

Caractéristique de la chaîne de caractères recherchée	**Mot clef**	**Exemple**
Contient « toto »	like	like '%toto%'
Ne contient pas « toto »	Not like	Not like '%toto%'
N'importe quelle chaîne de caractères	%	
Commence par « toto »	@	like '%@toto%'
Fini par « tata »	@	like '%tata@%'
Commence par « toto » et finit par « tata »	@..@	like '%@toto%tata@%'
Un et un seul caractère	_	like '%lait_e%'
Contient « l » suivi de 2 caractères et de « t »	_	like '%l__t%' pour lait, laitue, …

Encadré 1. *Rappel SQL*

Afin de pouvoir effectuer des requêtes précises, nous avons autorisé l'écriture de requêtes complexes grâce à l'utilisation possible d'expressions régulières écrites en langage SQL dans « NutriXConso ».

Que l'on écrive une requête simple (une seule chaîne de caractères) ou une requête plus complexe (requêtes SQL complexes avec possibilité d'utiliser des expressions régulières), l'outil fonctionne de la même façon. La chaîne de caractères écrite dans les différents champs de recherche vient s'insérer dans une requête SQL invisible pour l'utilisateur. Lorsque l'utilisateur remplit les

champs « *Req1* » et « *ReqSpeKantar* » et lance une recherche, les requêtes suivantes sont lancées et interrogent la base de données :

Champs remplis dans « NutriXConso »	**Requêtes SQL sous-jacentes***	
Req1 Ou *Req1* et *ReqSpeKantar*	Interrogation de la Table Kantar	**SELECT * FROM `kantar` where vmin like '%***Req1***%'** *ReqSpeKantar*
	Interrogation d'une Table de Données Nutritionnelles	**SELECT * FROM `nutri` where vmin like '%***Req1***%'**

*En vert est indiqué ce qui provient de l'interface « NutriXConso » (rentré par l'utilisateur). En rouge est indiqué ce que l'outil ajoute par lui-même.

Tableau 6 . *Utilisation de « Req1 » et « ReqSpeKantar » : Structure de la requête générale*

Lorsque « *Req1* » comprend un seul mot, il n'y a rien d'autre à préciser et il n'est pas nécessaire d'utiliser le langage SQL dans l'outil « NutriXConso » (pas besoin d'insérer de « % » ou de « _ »). Lorsqu'on a envie que « *Req1* » comprenne plusieurs mots ou des expressions plus compliquées, alors il est nécessaire d'utiliser le langage SQL. De même, l'utilisation de « *ReqSpeKantar* » nécessite la manipulation (même si elle peut être très légère) du langage SQL.

Utilisation de « ReqSpeKantar » vs « Req1 » :
Lorsque la requête porte sur des caractéristiques des variables Kantar, alors il faut utiliser « *ReqSpeKantar* » (car ces variables n'existent pas pour les données nutritionnelles). En revanche, « vall » et « vmin », même si elles ne représentent pas la même chose pour les données Kantar ou les données nutritionnelles, existent dans les 2 jeux de données. Elles peuvent donc être utilisées dans « *Req1* » sans problème, afin d'effectuer une même recherche dans Kantar et les données nutritionnelles.

Lorsqu'on lance la recherche avec « *Req1* »= « pizza » en espérant identifier les pizzas, le résultat Kantar comporte tous les produits dont au moins une caractéristique contient le mot « pizza ». On retrouve donc un certain nombre de produits ne correspondant pas au produit recherché (pâte à pizzas, mélange d'épices pour pizzas, …). Une solution est donc d'utiliser « *ReqSpeKantar* », les expressions régulières et le langage SQL afin de ne sélectionner que les pizzas (cf. **Figure 14**).

La requête générée est indiquée dans l'outil pour aider l'utilisateur en cas d'erreur de syntaxe.

Figure 14 . *Utilisation de « Req1 » et « ReqSpeKantar » pour n'afficher que les pizzas (N=1782)*

D'autres exemples d'utilisation de « *Req1* » et « *ReqSpeKantar* »

Produit recherché	Dans NutriXConso	Requête sous-jacentes pour Kantar
lait	*Req1* : lait	**SELECT * FROM `kantar` where vmin like '%**lait**%'**
Lait entier (de marque M1 pour Kantar ; toutes marques pour les sources nutri)	*Req1* : lait%' and vall like '%entier *ReqSpeKantar* : and sa2 like '%M1%'	**SELECT * FROM `kantar` where vmin like '%**lait%' and vall like '%entier**%'** and sa2 like '%M1%'
Lait entier (de marque M1 pour Kantar ET pour les sources nutri)	*Req1* : lait%' and vall like '%entier%' and vall like '%M1	Idem pour Kantar, mais affecte aussi les sources nutri : 0 ligne pour Ciqual (pas de marque pour le lait), 1 pour Cohen Serog
Tous les produits pizza (non) appariés avec Ciqual	*Req1* : pizza *ReqSpeKantar* : and id_ciqual is (not) null	**SELECT * FROM `kantar` where vmin like '%**pizza**%'** and id_ciqual is (not) null
Les pizzas ou les quiches (sans les biscuits apéritifs, les crêpes, nems, pâtes à tarte)	*Req1* : ' and (vmin like '%pizza%' or vmin like '%quiche%') and vall not like '%biscuit%' and vall not like '%crepes%' and vall not like '%nems%' and vall not like '%pates menageres	**SELECT * FROM `kantar` where vmin like '%**' and (vmin like '%pizza%' or vmin like '%quiche%') and vall not like '%biscuit%' and vall not like '%crepes%' and vall not like '%nems%' and vall not like '%pates menageres**%'**

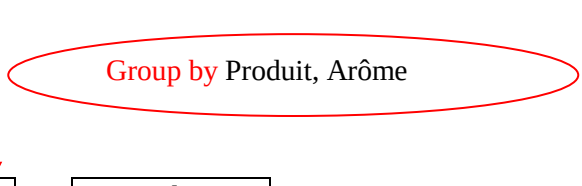
Tableau 7 . *Exemples d'utilisation de « Req1 » et « ReqSpeKantar »*

3.2.b Aide à la visualisation : le « *Group by* »

Une autre façon d'améliorer la visualisation d'une recherche sur les données Kantar est d'utiliser le « *Group by* ». Il permet de regrouper des lignes entre elles selon des caractéristiques choisies. Il modifie ainsi l'affichage des résultats de la requête en cours, en éliminant des variables non pertinentes et en réduisant le nombre de lignes affichées. Plus précisément, le « *Group by* » va n'afficher que les combinaisons distinctes des variables souhaitées (cochées).

L'exemple ci-dessous (**Figure 15)** présente 3 produits distincts avec des informations concernant le nom du produit, l'arôme, la marque, et la présence ou non de vitamines. Admettons que les informations sur la marque et les vitamines ne soient pas présentes dans le jeu des données nutritionnelles, il peut être intéressant d'enlever ces variables de l'affichage et de regrouper ainsi les 2 produits « A » « nature » :

Nom du Produit	Arôme	Marque	Vitamines
A	Nature	X	Non
A	Nature	Y	Oui
B	Vanille	Z	Oui



Produit	Arôme
A	Nature
B	Vanille

Représente
2 lignes
1 ligne

Figure 15 . *Exemple de « Group by ».*

Dans l'outil « *NutriXConso* », cela s'implémente simplement en décochant les champs jugés inutiles, soit parce qu'ils ne jouent pas dans le contenu nutritionnel du produit, soit parce que nous ne disposons pas de ces caractéristiques dans les sources nutritionnelles. Ensuite, il suffit de cliquer sur le bouton « G*roup by* ». Par exemple, dans les données Ciqual, les informations ne portent pas sur la marque du produit (à part pour les eaux). Il paraît donc pertinent de regrouper les produits Kantar de marque différente mais ayant les autres caractéristiques communes.

Dans notre exemple des pizzas, si l'utilisateur souhaite uniquement garder l'information sur le type de la pizza (fraîche ou surgelée) et sur les ingrédients, alors il décoche toutes les autres (cf. **Figure 16**). Après avoir cliqué sur le bouton « *Group by* », ces variables disparaissent de l'affichage (cf. **Figure 17**). Les lignes, avec des marques ou appellations ou fabricants différents, mais avec l'ensemble des caractéristiques que l'on a souhaité garder, identiques, seront regroupées en une seule.

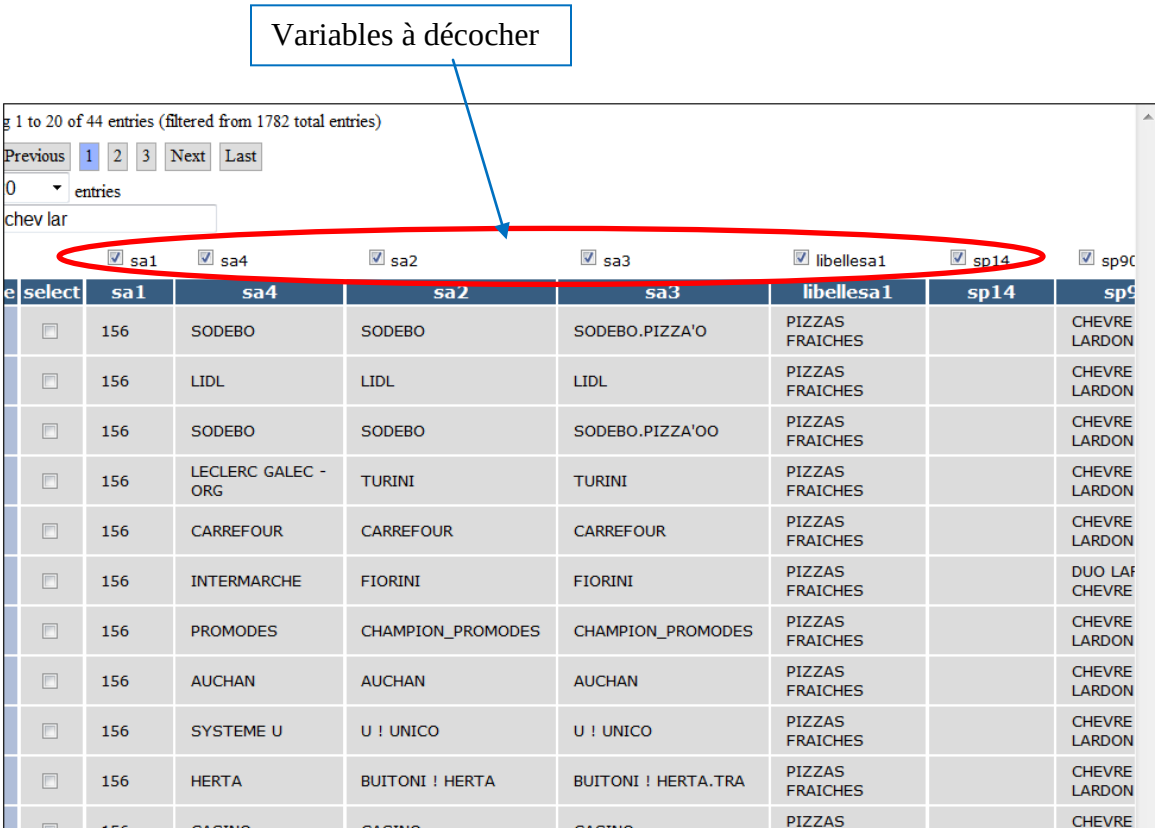


Figure 16 . *Avant « Group by » sur les pizzas → N= 1 782 lignes Kantar avec des caractéristiques distinctes (dont la marque, l'appellation, le fabricant,…).*

Après avoir décoché les variables et cliqué sur le bouton « *Group by* », les colonnes décochées disparaissent, certaines lignes sont regroupées et de nouvelles variables Kantar sont créées : « *Nombre RefNew* », « *Nombre Ciqual* » et « *Nombre Cohen* » (cf. **Figure 17**) :

- « *Nombre RefNew* » correspond au nombre de lignes (nombre de produits Kantar) représentées par cette agrégation,
- « *Nombre Ciqual* » (resp. « *Nombre Cohen* ») correspond au nombre de lignes pour lesquelles des données nutritionnelles Ciqual (resp. Cohen Serog) ont été affectées. Dans cet exemple, aucun appariement n'a encore été effectué pour ce produit.

Enfin, les variables renseignant sur le nombre d'occurrences d'achats des produits sont modifiées après le « *Group by* ». En effet, comme une ligne représente maintenant plusieurs produits, ce nombre d'occurrences correspond à la somme des occurrences d'achats de chaque produit représenté.

Cette ligne regroupe 40 produits qui sont des pizzas fraîches au chèvre/lardons (dont les caractéristiques décochées étaient différentes).

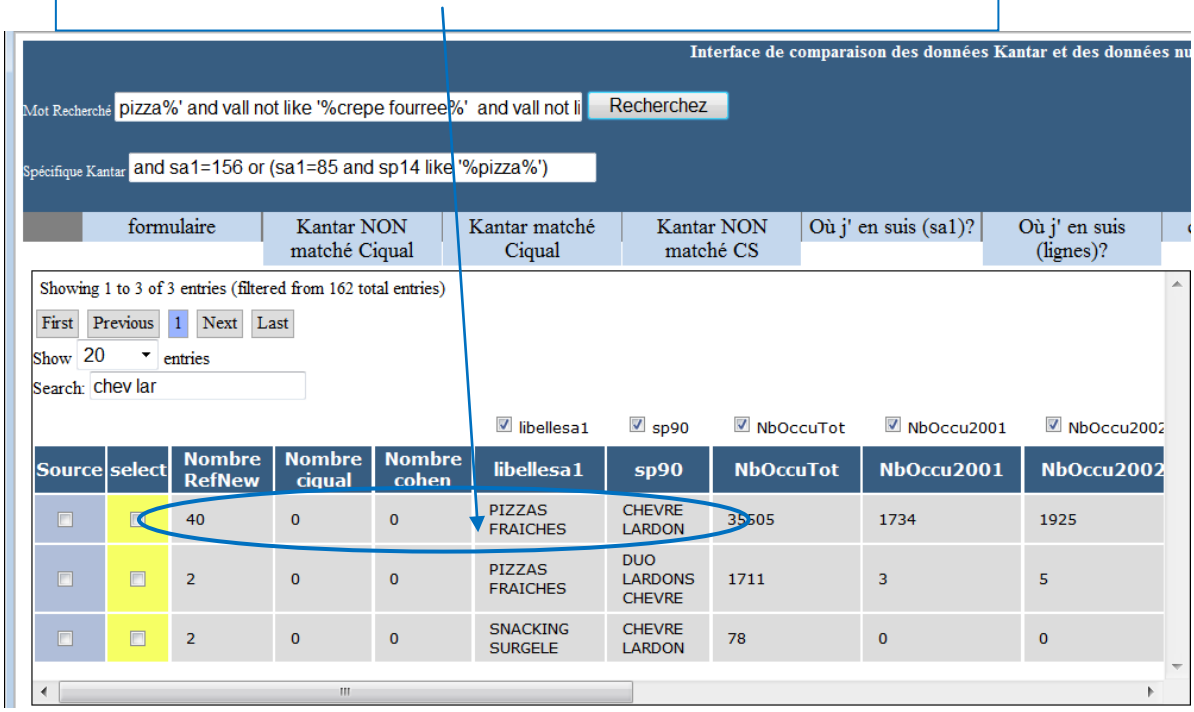


Figure 17 . *Après le Group by sur les pizzas → N= 3 lignes Kantar.*

Il est possible de voir les variables inutilisées en cliquant sur le bouton « *champs cachés* ». Les variables décochées lors d'un « **Group by** » apparaissent ainsi (cf. **Figure 18**).

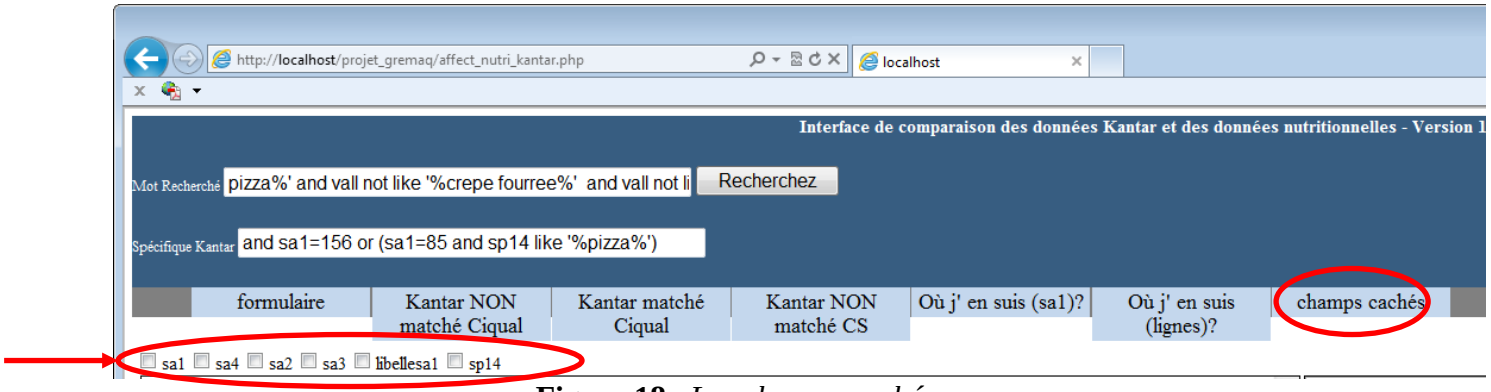


Figure 18 . *Les champs cachés.*

Il est aussi possible de faire autant de « *Group by* » que l'on souhaite, en décochant de nouvelles variables, ou en en réintégrant certaines (en les cochant après avoir cliqué sur « champs cachés »). Un mix des 2 (en décocher et en réintégrer) est aussi possible.

3.3 Synthèse des possibilités de visualisation

L'interface a donc été conçue avec nombre de fonctionnalités qui facilitent l'extraction et la visualisation des données :

- L'utilisation de l'ensemble des 4 champs de requêtes disponibles (« *Req1* » et « *ReqSpeKantar* », cf. section 3.1.a; « *Req2Kantar* » et « *Req2Nutri* », cf. section 3.1.b) permet d'affiner la recherche souhaitée. La possibilité d'utiliser des expressions régulières (voir section 3.2.a) dans le champ de recherche « *ReqSpeKantar* » ou encore « *Req1* » ajoute de très nombreuses possibilités pour générer des requêtes très précises.
  Dans l'exemple des pizzas, on a pu affiner la recherche en utilisant « *Req2Kantar* » et « *Req2Nutri* » (par exemple « chevre lar »).

- L'outil permet également de supprimer certaines variables sans intérêt et de regrouper des produits similaires (en termes de caractéristiques produit) de l'affichage Kantar (cf. section 3.2.b). Cela s'avère très utile étant donné le nombre très important de variables disponibles dans Kantar.

- D'autres fonctionnalités plus modestes sont malgré tout intéressantes pour une meilleure visualisation des données, comme :
  - Le tri dynamique, mentionné plus haut (section 3.1.f), peut permettre de repérer des variables sans intérêt (notamment concernant les données Kantar) ou de classer les produits selon une caractéristique souhaitée (par exemple les plus caloriques pour le jeu des données nutritionnelles).
  - le fait de pouvoir choisir de visualiser les lignes en une seule ou plusieurs pages (cf. section 3.1.d) ;
  - le fait de pouvoir mettre une ligne en valeur par rapport aux autres (cf. section 3.1.e) est également très intéressant pour les longues lignes (en particulier sur la source Kantar).

4. NutriXConso : un outil d'appariement

4.1 Fonctionnement

Une fois les produits similaires identifiés dans les deux jeux de données, il est très facile de réaliser les appariements. Il suffit de sélectionner le ou les produits à apparier à gauche (données Kantar), le ou les produits avec lesquels on souhaite les apparier (données nutritionnelles) et de cliquer sur le bouton « *Matching* » (cf. **Figure 19**).

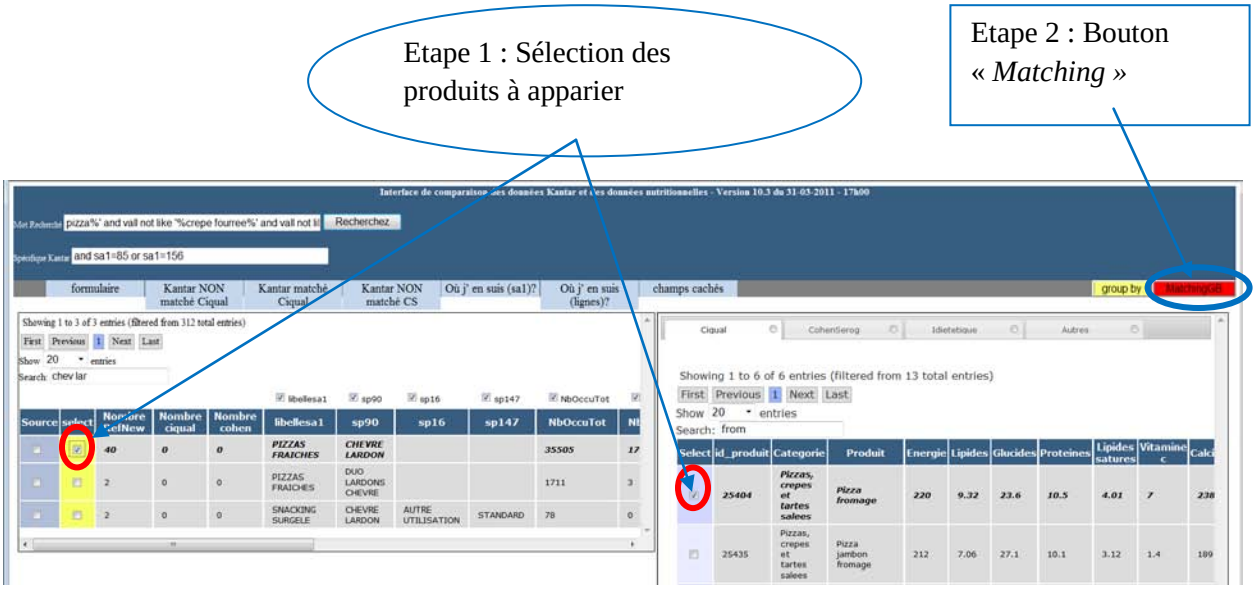


Figure 19 . *Exemple d'appariement Kantar-Ciqual : pizzas chèvre lardon.*

Une fois qu'on clique sur le bouton « *Matching* », les appariements sont réalisés et apparaissent maintenant en jaune, et la variable « *Nombre Ciqual* » qui valait 0 vaut maintenant 40 (signifiant que les 40 lignes ont été appariées), cf. **Figure 20**.

Figure 20 . *Après appariement Kantar-Ciqual : pizzas chèvre lardon.*

Ici (**Figure 20**), l'ensemble des pizzas chèvre lardons Kantar (regroupement de 40 produits) a été apparié avec un produit Ciqual (pizza fromage). Il peut ensuite être intéressant de rechercher les produits dans les données Cohen Serog (cf. **Figure 21**). Si le contenu nutritionnel diffère de celui de la source Ciqual, cela permettra d'affiner la composition nutritionnelle affectée au produit apparié.

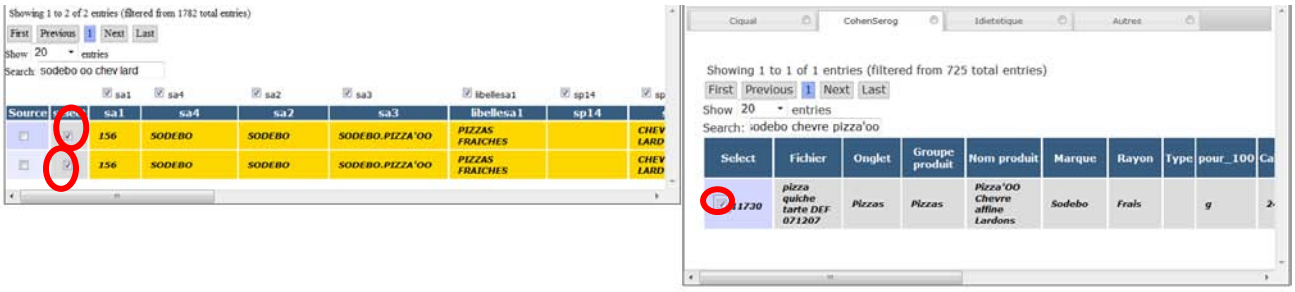


Figure 21 . *Exemple d'appariement Kantar-Cohen Serog : pizzas chèvre lardon.*

Une fois qu'on clique sur le bouton « *Matching* », les appariements sont réalisés et apparaissent maintenant en vert pour Kantar, en bleu pour Cohen Serog (cf. section 4.4 pour plus d'explications). La variable « *id_cohen* » correspond maintenant à l'identifiant du produit Cohen Serog apparié (cf. **Figure 22**).

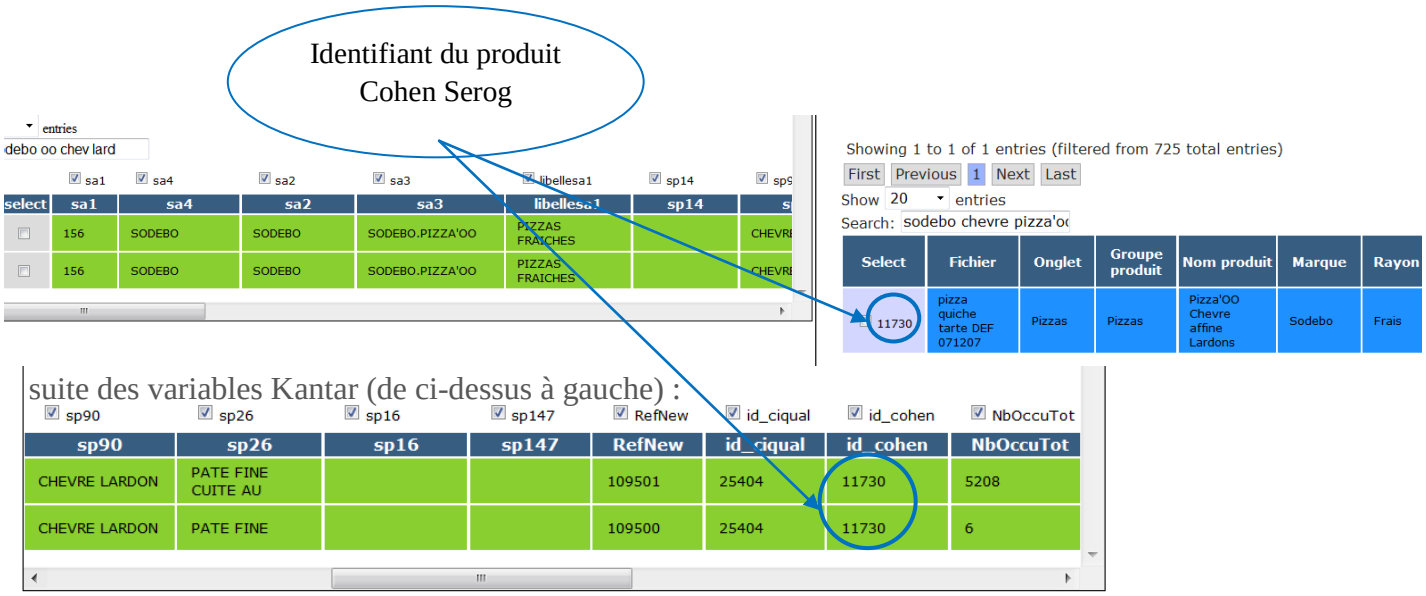
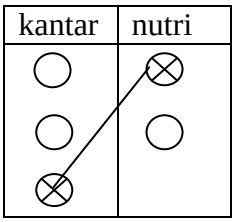


Figure 22 . *Après appariement Kantar-Cohen Serog : pizzas chèvre lardon.*

4.2 Différents types d'appariement

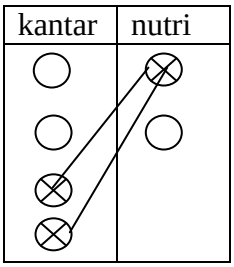
4.2.a Cas Standard

Le cas le plus simple correspond à un appariement simple 1:1 ou multiple N:1 :



Appariement 1:1

ou



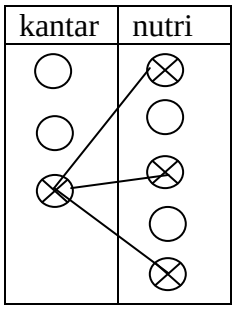
Appariement N:1 (ici 2:1)

La **Figure 21** illustre un exemple d'appariement simple. Les appariements réalisés après un « *Group by* » correspondent à un appariement multiple N:1 (cf. **Figure 19**).

4.2.b Cas Moyenne simple

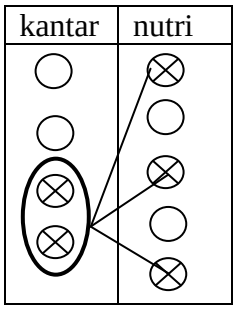
Nous avons ajouté une fonctionnalité très utile : dans le cas où aucun produit des sources nutritionnelles ne correspondrait exactement au produit Kantar que l'on souhaite apparier, il est possible de créer un nouveau produit moyen Ciqual (ou Cohen Serog). Ce nouveau produit correspondra à une moyenne des informations nutritionnelles de produits Ciqual déjà existants sélectionnés. Les N lignes sélectionnées dans la table nutritionnelle permettront de créer un nouveau produit (composé des moyennes des nutriments des N lignes sélectionnées). Ce nouveau produit sera automatiquement inséré dans la table des sources nutritionnelles, et affecté au produit Kantar sélectionné. Notez que le nouveau produit créé, aura pour nom, l'ensemble des noms des produits utilisés dans la moyenne (séparés par des @).

Ce cas correspond à un appariement multiple 1:N ou N:N :



Appariement 1:N (ici 1:3)

ou



Appariement N:N (ici 2:3)

Dans notre exemple des pizzas, on remarque qu'il n'y a pas de pizzas au chorizo dans les données Ciqual. Une idée peut être d'affecter aux pizzas au chorizo Kantar une moyenne de l'ensemble des pizzas Ciqual. Il suffit simplement de réaliser l'appariement 1:N souhaité et un

nouveau produit moyen apparaîtra dans la table Ciqual. Après avoir sélectionné les produits Kantar (ici un seul) et les produits Ciqual (ici 7) une seule action suffit : en appuyant sur le bouton « *Matching* », le nouveau produit moyen sera créé et l'appariement réalisé (cf. **Figure 23** et **Figure 24**).

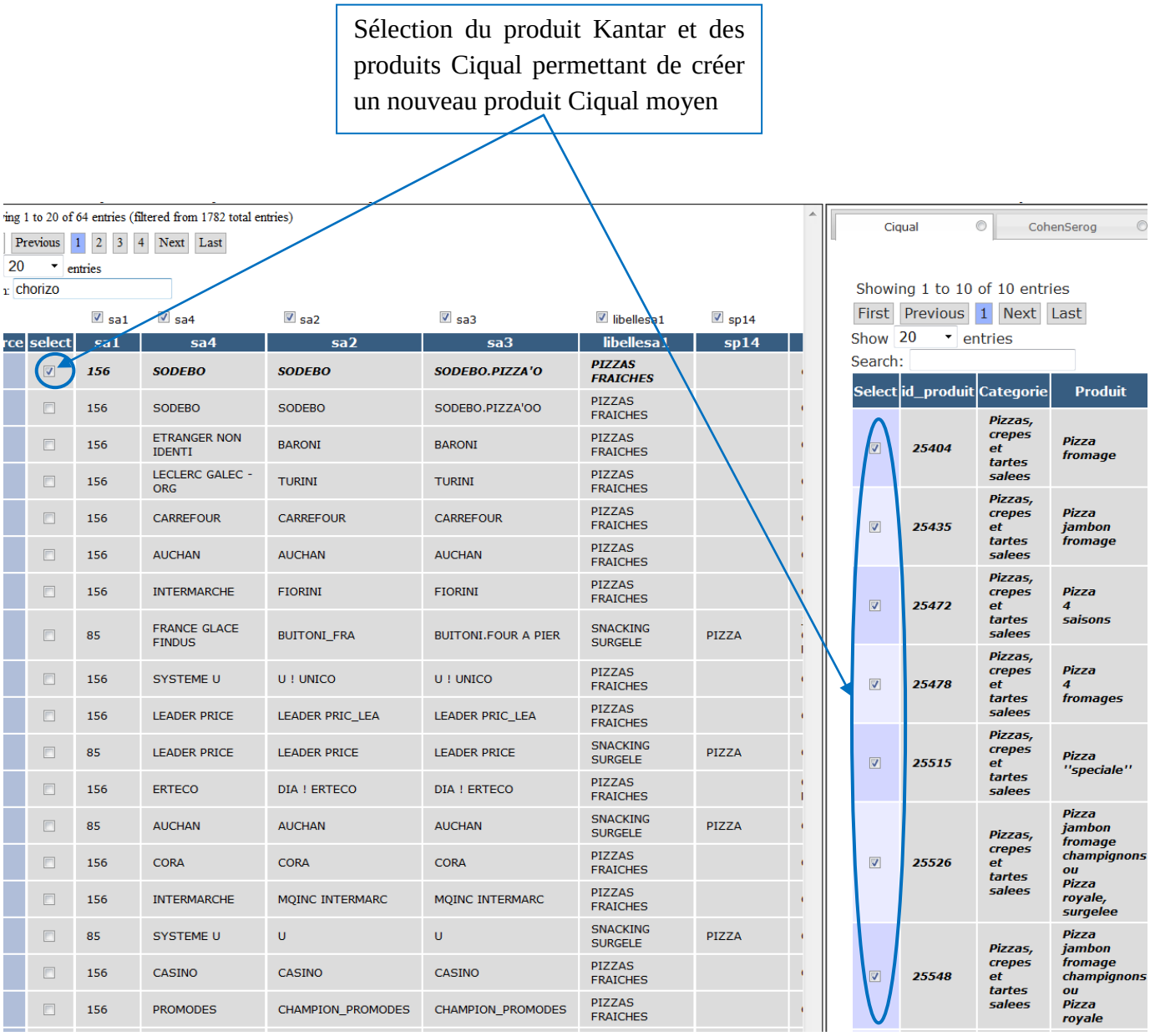


Figure 23 . *Génération d'une moyenne et appariement: pizzas chorizo.*

Et après avoir cliqué sur le bouton « *Matching* », la création du nouveau produit et l'appariement sont réalisés, d'ailleurs visible en jaune (cf. **Figure 24)**.

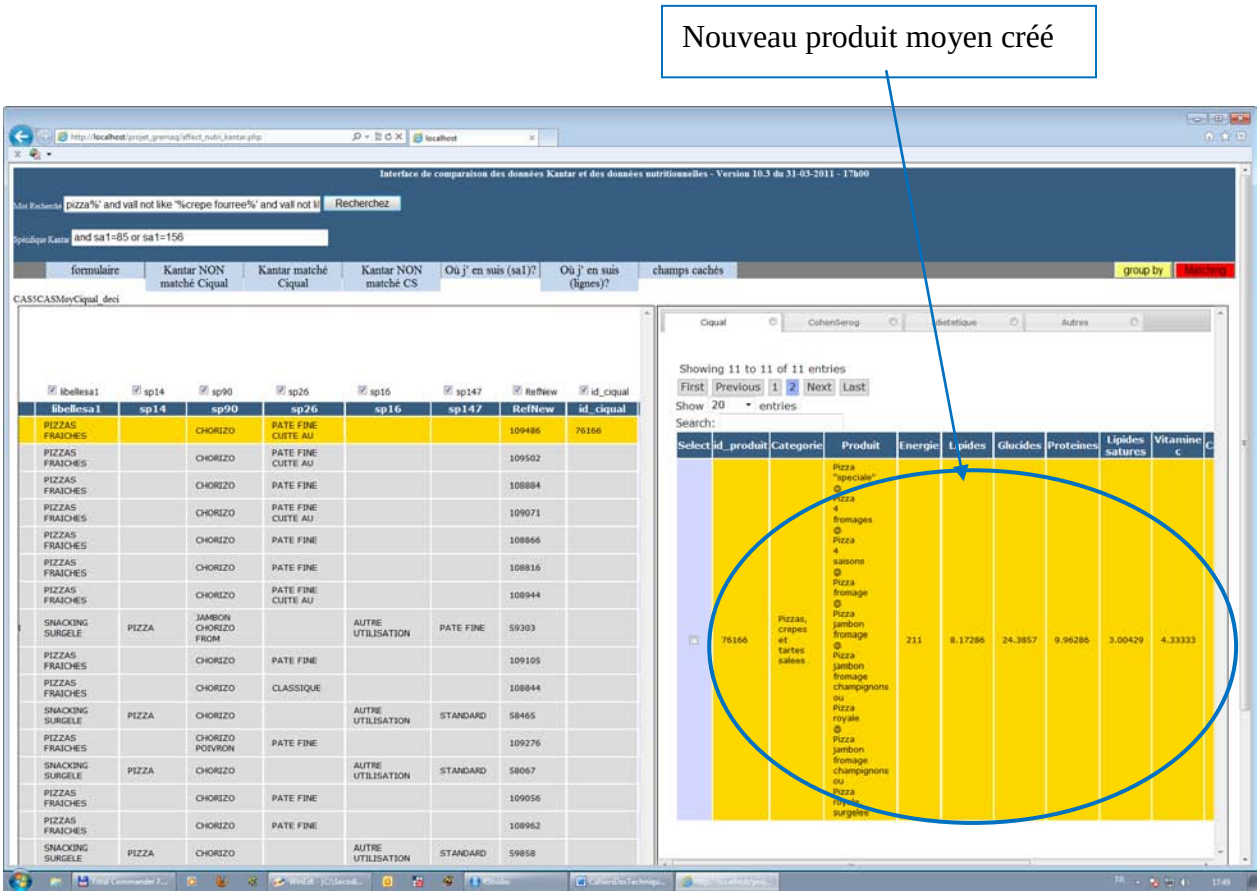


Figure 24 . *Nouveau produit moyen créé et appariement réalisé.*

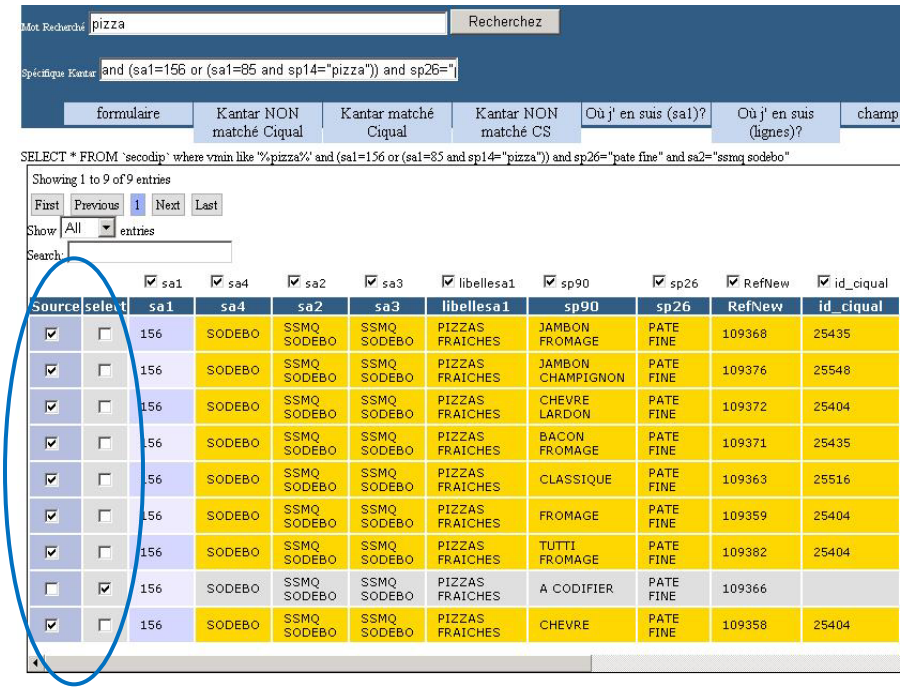
4.2.c Cas Moyenne Pondérée

Pour certains produits Kantar, une ou plusieurs variables, habituellement renseignées sur le type de produit considéré, peuvent être manquantes. Par exemple, **Figure 25**, la variable « *sp90* » qui donne habituellement le type de pizza indique « à codifier » pour la ligne grisée (et non « jambon fromage » ou autres recettes, comme sur les autres lignes, **Figure 25**).

Dans ce cas, plutôt que d'apparier cette pizza Kantar inconnue avec une pizza connue des sources nutritionnelles, il est possible d'indiquer que les données nutritionnelles de cette pizza inconnue seront calculées à partir des données nutritionnelles des pizzas connues, mais par une moyenne pondérée par le nombre d'achats. Ainsi, dans la **Figure 25**, nous avons coché la case « *select* » de la ligne grisée et coché la case « *source* » des autres lignes (jaunes, car elles sont déjà appariées avec une référence Ciqual, cf. section 4.4). Un clic sur le bouton « *Matching* » réalisera cet appariement par la moyenne pondérée. Il sera enregistré dans la table « ponderation » de la base (cf. Annexe 1).

Pour résumer :

- colonne **select** : produit(s) Kantar à affecter ;
- colonne **source** : produit(s) Kantar dont on se sert pour l'affectation (moyennes pondérées sur les données nutritionnelles de ces lignes ; pondération par le nombre d'achats).



Source	select	sa1	sa4	sa2	sa3	libellesa1	sp90	sp26	RefNew	id_ciqual
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	JAMBON FROMAGE	PATE FINE	109368	25435
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	JAMBON CHAMPIGNON	PATE FINE	109376	25548
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	CHEVRE LARDON	PATE FINE	109372	25404
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	BACON FROMAGE	PATE FINE	109371	25435
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	CLASSIQUE	PATE FINE	109363	25516
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	FROMAGE	PATE FINE	109359	25404
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	TUTTI FROMAGE	PATE FINE	109382	25404
☐	☑	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	A CODIFIER	PATE FINE	109366	
☑	☐	156	SODEBO	SSMQ SODEBO	SSMQ SODEBO	PIZZAS FRAICHES	CHEVRE	PATE FINE	109358	25404

Figure 25 . *Exemple de moyenne pondérée possible.*

Cette fonctionnalité que nous avions anticipée, n'a finalement pas été utilisée. En effet, la personne qui a réalisé les appariements a pu choisir une référence nutritionnelle particulière ou préférer créer un produit moyen à partir des données nutritionnelles sources.

4.3 Changement d'appariement

En cas d'erreur d'appariement, il faut savoir qu'une nouvelle affectation viendra écraser la précédente.

4.4 Visualisation des appariements effectués

Pour aider à la visualisation de l'avancement du travail sur les appariements, nous avons établi un code couleurs permettant à l'utilisateur de repérer très rapidement les appariements restants à réaliser. Des boutons permettent également de lister les produits Kantar non appariés (resp. ceux appariés).

4.4.a Un code couleurs

Des couleurs indiquent les lignes Kantar appariées (à des produits Ciqual et/ou Cohen Serog), cf. **Tableau 8**. Le jaune (resp. bleu) signifie que tous les produits (1 ou plusieurs dans le cas d'un « *Group by* ») sont appariés à Ciqual (resp. Cohen Serog). Le vert signifie que les produits

Kantar sont appariés à Ciqual et à Cohen Serog. Des variantes plus claires interviennent dans le cas du « *Group by* » lorsque seulement certaines lignes sont appariées (mais pas toutes).

	Ciqual		***Cohen Serog***
3 couleurs basiques :			
	toutes	et	aucune
	aucune	et	toutes
	toutes	et	toutes
Et des plus claires (cas group by) :			
	certaines	et	aucune
	aucune	et	certaines
	certaines	et	certaines
	certaines	et	toutes
	toutes	et	certaines

Tableau 8 . *Code couleur des appariements*

Notez que les 2 premières couleurs sont aussi affectées aux lignes des tables des sources nutritionnelles (resp. Ciqual et Cohen Serog) lorsqu'un produit de ces sources est affecté au moins une fois à un ou plusieurs produits Kantar.

La **Figure 21** illustre le cas où un produit Kantar est apparié à un produit Ciqual. De même, la **Figure 20** illustre le cas où tous les produits représentés par cette ligne sont appariés à un produit Ciqual. **La Figure 22** présente un exemple de cas où les produits sont appariés à Ciqual et à Cohen Serog.

L'objectif final du projet était bien sûr que toutes les lignes soient jaunes, ou vertes.

4.4.b Requêtes additionnelles

Nous avons ajouté des boutons permettant de visualiser très rapidement l'état d'avancement des appariements : sur le produit en cours ou de manière générale sur l'ensemble des produits Kantar (cf. **Figure 26**).

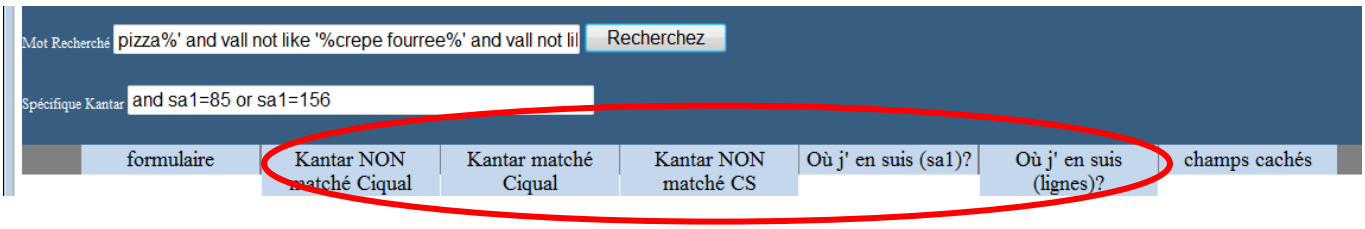


Figure 26 . *Les requêtes additionnelles.*

Les boutons permettant d'exécuter ces requêtes et leurs fonctionnalités sont les suivants :

- *« Kantar NON matché Ciqual »* : ouvre une nouvelle fenêtre avec l'ensemble des lignes Kantar, relatives au produit recherché en cours, non appariées à Ciqual.
- *« Kantar NON matché CS »* : idem mais Cohen Serog
- *« Kantar matché Ciqual »* : affiche les lignes appariées
- *« Où j'en suis (sa1) »* (cf. **Figure 27**) : ouvre une nouvelle fenêtre avec un tableau récapitulatif de l'état d'avancement des appariements, par numéro de produit Kantar. Par produit Kantar, on y trouve le nombre total de lignes (produits différents), le nombre de lignes déjà appariées à Ciqual (resp. Cohen Serog) et le pourcentage correspondant, ainsi que le nombre de produits Ciqual (resp Cohen Serog) différents utilisés dans ces appariements.
- *« Où j'en suis (lignes)»* (cf. **Figure 28**) : idem mais sur l'ensemble des produits

First Previous 1 2 3 4 5 Next Last
Show 20 entries
Search:

libellesa1	sa1	NbLignes	NBlignes_Ciqual_OK	NBlignes_CS_OK	Pourc_Ciqual_OK	Pourc_CS_OK	Distinct_Ciqual	Distinct_CS
PIZZAS FRAICHES	156	861	41	2	4.7619	0.2323	2	1
CHEWING GUM	36	461	0	0	0.0000	0.0000	0	0
RHUBARBE FRAICHE	321	1	0	0	0.0000	0.0000	0	0
SAINT NECTAIRE EAN	256	40	0	0	0.0000	0.0000	0	0
HERBES ET AROMATES FRAIS	487	149	0	0	0.0000	0.0000	0	0
PLATS ET ENTREES SURGELEES	122	3389	0	0	0.0000	0.0000	0	0
PORC ENTIER CARCASSE SANS EAN	412	2	0	0	0.0000	0.0000	0	0
CHOCOLAT EN TABLETTE	31	3068	0	0	0.0000	0.0000	0	0
PISSENLITS FRAIS	316	1	0	0	0.0000	0.0000	0	0
MIMOLETTE EAN	251	83	0	0	0.0000	0.0000	0	0

Figure 27 . *Exemple de résultat de la requête « Où j'en suis (sa1) ».*

Par exemple (**Figure 27**), le bouton « *où j'en suis (sa1)* » nous indique que, pour les pizzas fraîches (première ligne grisée de la Figure), sur les 861 produits Kantar différents, 41 ont été appariés avec Ciqual (soit près de 5%) et 2 avec Cohen Serog (soit 0,2% environ). En plus de ces informations, les deux dernières colonnes de la figure indiquent que ces appariements impliquent 2 références Ciqual distinctes et 1 référence Cohen Serog.

Showing 1 to 1 of 1 entries
First Previous 1 Next Last
Show 20 entries
Search:

NbLignes_TOT	NBlignes_Ciqual_OK	NBlignes_CS_OK	Pourc_Ciqual_OK	Pourc_CS_OK	Distinct_Ciqual	Distinct_CS
184617	41	2	0.0222	0.0011	2	1

Figure 28 . *Exemple de résultat de la requête « Où j'en suis (lignes) ».*

La **Figure 28** est le pendant de la **Figure 27**, mais pour l'ensemble des produits (bouton « *où j'en suis (lignes)* ». On retrouve exactement les mêmes résultats qu'auparavant (avec des pourcentages plus faibles : 0,02% pour Ciqual et 0,001% pour Cohen Serog), puisque dans cet exemple aucun autre produit n'a été apparié. Ce bouton a toutefois permis d'avoir une vue générale de l'avancée des appariements tout au long du travail.

L'objectif du projet a été d'atteindre, pour chaque produit Kantar, 100% d'appariements à Ciqual.

4.5 Alimentation de la base de données

La base de données ayant été initialisée avec les données Kantar et Cohen Serog, la principale source d'alimentation de la base de données provient bien sûr des appariements effectués.

Chaque appariement entre une ligne (i.e. un produit) Kantar et une ligne nutritionnelle vient :

- renseigner la référence nutritionnelle correspondante dans la table « kantar » : *« id_ciqual »*, *« id_cohen_serog »*, *« id_idiet »* ou *« id_autre »*, pour le produit Kantar apparié ;
- alimenter (d'autant de lignes qu'il y a de variables Kantar) la table de décision correspondante (« decision_ciqual », « decision_cohen_serog », « decision_idiet » ou « decision_autre ») qui permet notamment de retrouver les variables Kantar du produit apparié qui étaient masquées (cf. « *Group by* », section 3.2.b) lors de l'appariement ;
- et éventuellement ajouter une ligne dans la table nutritionnelle correspondante (« ciqual », « cohen_serog », « idiet » ou « Autre ») dans le cas d'un nouveau produit moyen créé. Dans ce cas, la table moyenne correspondante (« moyenne_ciqual », « moyenne_cohen_serog », « moyenne_idiet » ou « moyenne_autre ») est également alimentée d'autant de lignes qu'il y avait de références nutritionnelles sélectionnées.

L'unique autre source d'alimentation de la base pendant la période d'appariement est l'insertion « manuelle » (via l'interface graphique de phpMyAdmin, cf. **Figure 2**, section 2) de références nutritionnelles dans les tables « idietetique » et « autre ».

Conclusions

L'outil « NutriXConso » a permis d'effectuer l'ensemble des appariements nécessaires et ainsi d'ajouter de l'information dans les données d'achats déjà existantes. Chaque produit acheté dispose maintenant d'informations relatives à sa composition nutritionnelle (macro et micronutriments).

342 539[10] références produits Kantar (produit 2001NF-2008) ont ainsi été appariées à des produits des sources nutritionnelles. **2797** références nutritionnelles distinctes ont été utilisées (*RefNutri* qui correspond à *id_ciqual*, *id_cohen*, *id_idietetique* ou *id_autre*).

A partir de cette base de données, nous avons donc complété les fichiers d'achats Kantar en combinant les différentes informations nutritionnelles disponibles pour chaque produit (cf. Annexe 2).

Ces données sont déjà disponibles aux chercheurs Gremaq-INRA qui peuvent ainsi utiliser ces données dans leurs travaux de recherche[11]. L'outil « NutriXConso » a été conçu afin de répondre à une demande précise et ponctuelle. Il n'a donc pas vocation à être utilisé par la suite. Les codes

[10] Une fois les variables Kantar non pertinentes pour la composition nutritionnelle, supprimées, 184 618 produits ont été identifiés et appariés.

[11] Evidemment, seuls les chercheurs du GREMAQ ayant le droit d'utiliser les données Kantar ont la possibilité de disposer de ces données « NutriXConso ».

des programmes sont diffusables mais nécessiteraient des améliorations et des adaptations à toute nouvelle problématique.

Limites

Tout d'abord, au niveau des données, il est important d'avoir conscience que les données nutritionnelles utilisées correspondent aux données récoltées au moment du projet, même pour des données Kantar antérieures. Elles ne tiennent donc pas compte de l'évolution possible des recettes dans le temps.
Ensuite, la finesse d'information des produits appariés varie suivant les produits Kantar. Parfois les informations nutritionnelles correspondent exactement au produit Kantar (niveau marque), d'autres fois le produit apparié correspond à un niveau plus agrégé.

Ensuite, au niveau technique, les choix informatiques effectués font également apparaître des limites. Même si l'outil a permis de réaliser parfaitement l'ensemble des appariements, il présente certaines anomalies techniques, principalement dues au nombre de lignes à afficher parfois trop important dans l'outil : ainsi, le tri par colonne (cf. section 3.1.f) n'est parfois pas correctement effectué ; parfois le volume des données extraites (et à mettre en mémoire dans l'interface) est tel que l'outil s'arrête et affiche un message d'erreur correspondant.
Tous ces bugs peuvent cependant être contournés, par exemple en effectuant les recherches par sous-produit. C'est donc ainsi que nous avons procédé : en segmentant les plus grands groupes de produits en plus petits sous-groupes.

Enfin, si l'outil était à reprendre, nous avions pensé à mettre en place un certain nombre d'éléments supplémentaires pour le rendre plus robuste, mais que nous n'avons pas pu entreprendre faute de temps. Par exemple, nous avons simplement autorisé les changements d'appariement, afin de permettre de corriger de façon très simple une erreur d'appariement précédente. Une amélioration possible serait de prévenir l'utilisateur et lui faire valider chaque changement. Nnous aurions également pu modifier la base de données pour qu'elle puisse conserver chaque appariement avec son horodatage.

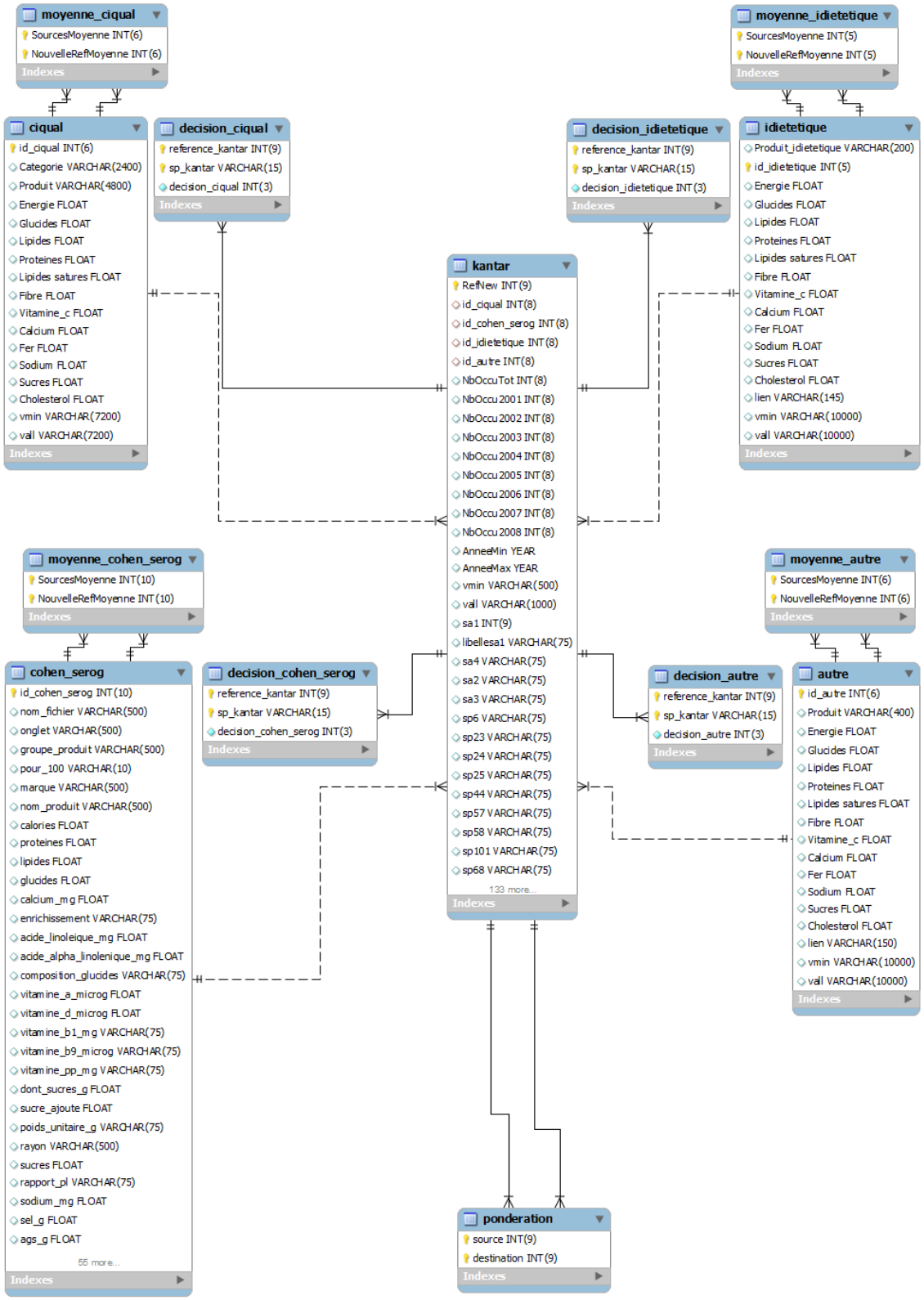
Références bibliographiques

"Savoir manger, Le guide des aliments 2008-2009", Cohen J.M, Serog P.

Table de composition nutritionnelle des aliments Ciqual 2008, Agence française de sécurité sanitaire des aliments

« L'équilibre nutritionnel. Concepts de base et nouveaux indicateurs: Le SAIN et le LIM », Darmon M., Darmon N, 2008

Annexe 1 : NutriXConso : une base de données



Annexe 2 : Choix effectués, méthodologies des appariements

A2.1 Variables et finesse des informations

Dans les fichiers d'achats disponibles, les informations nutritionnelles fournies correspondent aux informations du cahier des charges. Les informations nutritionnelles minimales sont:

- macronutriments : calories, lipides et lipides saturés, glucides, protéines ;
- micronutriments : fibre, vitamine C, calcium, fer, sodium, sucre ajoutés, cholestérol.

Les produits Kantar ont tous été affectés à un produit Ciqual.[12] Il arrive cependant que le produit Ciqual ne représente pas exactement le produit Kantar. C'était par exemple le cas pour la farine qui est absente de Ciqual (appariement avec le blé dur précuit sec de Ciqual).

Dans ces cas-là, une source additionnelle (Cohen Serog, idietetique, autre) vient compléter Ciqual. Les informations Cohen Serog permettent généralement d'affiner un produit Ciqual trop agrégé ou éloigné du produit Kantar. Les tables « idietetique » et « Autre » ont aussi permis d'ajouter des produits non présents dans Ciqual ni dans Cohen Serog (N=95 produits rajoutés, le contenu exact de la farine est par exemple renseigné dans la table « Autre »).

Les produits Cohen Serog ont aussi été examinés. Pour un même type de produit, si des différences de macronutriments entre marques (comparaison à la valeur Ciqual du produit moyen) sont constatées, les marques « hautes » ou « basses » (en termes nutritionnels) sont appariées à Kantar. Exemples de produits pour lesquels la finesse d'information est au niveau de la marque : chocolat en tablette, barres céréalières, céréales petit déjeuner, potages, barres chocolatées, compotes, eaux minérales…
Dans les fichiers délivrables, seules les informations de la source la plus précise sont conservées. Par exemple, si un produit Kantar est apparié à Ciqual et à Cohen Serog, toutes les informations nutritionnelles seront celles de Cohen Serog. Dans le cas de manquants pour certaines variables Cohen Serog, les informations Ciqual permettent de compléter (même si Ciqual correspond à un produit moins précis ou à un produit d'un niveau plus agrégé).

A2.2 Moyennes créées

Des moyennes ont parfois été créées dans les sources nutritionnelles (Ciqual, Cohen Serog…) afin de rajouter de l'information et d'affiner au mieux les appariements avec Kantar.
Par exemple pour les données Cohen Serog, certaines moyennes de plusieurs marques ont été générées afin d'affiner les appariements par rapport aux données Ciqual (ex. chips à l'ancienne absentes de Ciqual donc intéressant d'utiliser une moyenne de plusieurs marques Cohen Serog,…).

Ainsi : 86 nouveaux produits « moyens » Ciqual, 31 Cohen Serog et 3 Autres ont été créés.

[12] Sauf quelques produits non alimentaires présents par erreur dans Kantar (ex. bougies d'anniversaire, aliments pour animaux…) qui n'auront aucun identifiant Ciqual relié mais auront des nutriments nuls.

A2.3 Produit cru/cuit ?

Pour les légumes et les féculents, les nutriments sont ceux du produit cru (quand il y avait le choix) :

- car le poids connu du produit (dans Kantar) est le poids acheté du produit (cru) ;
- car on n'a pas correspondance poids cru/cuit pour chaque produit ;
- et les macronutriments ne vont pas changer entre cuit et cru. Par contre, garder en tête que les micros peuvent changer (ex. Vit C…).

A2.4 Acides gras saturés (AGS)

Dans le cas où un produit Kantar est apparié à une source additionnelle à Ciqual (Cohen Serog, idietetique, autre) et dans le cas où l'information sur les AGS est manquante dans cette source mais les lipides présents, on complète la variable AGS en utilisant les informations Ciqual. La procédure consiste juste à calculer la part des AGS dans les lipides Ciqual et à appliquer ce rapport aux lipides de la source additionnelle.

A2.5 Sucres ajoutés

Cette information est absente de Ciqual (dans Ciqual, on a les sucres totaux). Nous avons essayé de créer cette variable à partir de l'information des autres sources et en prenant certaines hypothèses :

- utilisation de l'information éventuellement présente dans une autre source appariée à un produit Kantar ;

- utilisation de l'information sur les sucres totaux (information présente dans Ciqual) :
  - Pour certains produits, les sucres ne sont que des sucres ajoutés. On peut donc facilement renseigner l'information des sucres ajoutés.
    - C'est le cas pour les produits classés dans le groupe « *Produits sucrés, salés et gras* » (sauf les « *Jus et nectars* », les *« Matières grasses* » et « *Graines oléagineuses et châtaignes* »),
    - Idem pour les groupes « *Boissons alcoolisées* », « *Charcuteries et salaisons* », « *Boissons sans alcool* », « B*oulangerie-viennoiserie* », « *Compléments alimentaires* », « *Céréales petit déjeuner et barres céréalières »*, « *Denrées destinées à une alimentation particulière* »
  - Si les sucres totaux sont nuls, cela implique que les sucres ajoutés sont nuls ;

- on complète également certains produits pour lesquels on sait que les sucres ajoutés sont nuls.

A2.6 Information manquante dans une source additionnelle à Ciqual

Dans les cas où une variable est manquante dans une source additionnelle à Ciqual, nous utiliserons la valeur de la même variable pour la source Ciqual (sauf AGS). Il faut donc bien avoir conscience que les différentes sources sont « combinées » afin d'obtenir les informations les plus complètes et précises possibles.

- **Information manquante dans Ciqual**

Même si un certain nombre de manquants ont été corrigés (certains manquants étaient des « 0 »), il reste certains manquants dans les variables du cahier des charges (micronutriments). Ils ne sont pas nombreux mais il faut garder cela en tête :

Variables micronutriments	***Nb de produits Kantar (RefNew) avec variable manquante***	***Nb de produits Ciqual utilisés avec variable manquante***
Fibres	537 soit 0.29%	10
AGS	1878 soit 1.02%	16
VitC	2384 soit 1.29%	20
Fer	2259 soit 1.22%	22
Calcium	1123 soit 0.61%	17
Sodium	328 soit 0.18%	3
Cholestérol	1578 soit 0.84%	17
Sucres ajoutés	**37168 soit 20.13%**	274

- **Des erreurs corrigées dans les sources nutritionnelles**

Certaines erreurs présentes dans les sources nutritionnelles ont été détectées et corrigées (notamment dans les fichiers Cohen Serog). Ces erreurs pouvaient être des fautes de frappes (par ex. 1106g de protéines au lieu de 11.06) ou des problèmes d'unités (par ex. calories pour 1 repas ou 1 barre et non pas pour 100g de produit). Une moulinette a été mise en place pour essayer de détecter le maximum d'erreurs. Il est cependant possible que certaines erreurs demeurent.