Metaprogramme Did'it

La plateforme de données «Alimentation» pour des recherches pluridisciplinaires sur l'alimentation

Les recherches dans le champ de l'alimentation imposent de mobiliser de nombreuses données sur les consommations alimentaires et les caractéristiques économiques, nutritionnelles, environnementales, sanitaires… des aliments disponibles sur le marché. Il en est ainsi, par exemple, si l'on souhaite estimer les impacts économiques, de santé et environnementaux d'une évolution des régimes alimentaires, évaluer les impacts de politiques publiques de l'alimentation, ou encore identifier les marges de manœuvre des entreprises en matière de reformulation des produits et d'innovation.

Dans de nombreux travaux de recherche actuels ou à développer, la possibilité de prendre en compte simultanément des dimensions économiques, sociales, nutritionnelles et environnementales de l'alimentation s'avère ainsi primordiale. Or les données requises pour conduire ces travaux sont généralement dispersées, difficilement connectables entre elles, voire manquantes.

Ce constat a conduit à envisager la constitution d'une plateforme regroupant les données nécessaires à la conduite des recherches envisagées en matière de durabilité de l'alimentation. La démarche engagée vise ainsi à :

1. **Regrouper sur une même plateforme les bases de données existantes, de manière à en faciliter l'accès et rendre plus aisée leur mise en relation ;**
2. **Identifier les données non encore disponibles et préciser les modalités de leur acquisition ou de leur création ;**
3. **Définir les règles d'usage de ces données de façon à favoriser le développement de nouveaux programmes de recherche, tout en tenant compte des droits d'accès imposés par les fournisseurs de ces données ;**
4. **Développer les outils informatiques nécessaires à la mise à disposition des bases et à leurs appariements[1] entre elles.**

[1] Les appariements entre deux bases de données représentent les liens entre ces bases permettant leur interconnexion.

PÉRIMÈTRE DE LA PLATEFORME

L'originalité de la démarche est de mettre au centre de la plateforme les produits alimentaires disponibles pour le consommateur sur le marché final. Ces produits peuvent être décrits sur la base de nomenclatures (descriptions du produit) plus ou moins agrégées : certaines se situent au niveau de la référence produit du marché (au niveau de la marque), d'autres sont définies pour des catégories d'aliments plus agrégées (valeurs moyennes pour de grandes familles d'aliments).

L'objectif est de pouvoir associer aux produits alimentaires ainsi définis, un ensemble de variables utiles à mobiliser pour des recherches sur l'alimentation (voir graphique 1) : données économiques (prix, quantités achetées et consommées, caractéristiques sociodémographiques des consommateurs…), composition nutritionnelle, informations disponibles sur les emballages (labels, allégations, listes d'ingrédients…), données toxicologiques (teneurs en pesticides, métaux lourds…), données environnementales (impact CO2, consommation d'eau…), données sensorielles (perception des consommateurs pour les saveurs sucrée, salée…), données relatives aux procédés alimentaires (degré de transformation industrielle), données épidémiologiques (risques relatifs d'incidence de certaines pathologies…).
Ce périmètre a vocation à évoluer si d'autres dimensions autour du produit alimentaire s'avèrent importantes pour les recherches.

LES BASES DE DONNÉES CONCERNÉES

Ces données peuvent provenir de différentes sources. Certaines sont produites par des opérateurs privés comme Kantar WorldPanel pour les données d'achats des ménages, par des bureaux d'études pour l'impact environnemental des aliments. D'autres sont élaborées par des opérateurs publics :

- l'ANSES pour les données de consommation (INCA2, INCA3), de composition des aliments (CIQUAL), et relatives à la présence de contaminants (EAT2) ;
- l'Oqali pour les données d'étiquetage, de composition nutritionnelle et les listes d'ingrédients ;
- l'ADEME pour les données d'impact CO2 des aliments (FoodGES).

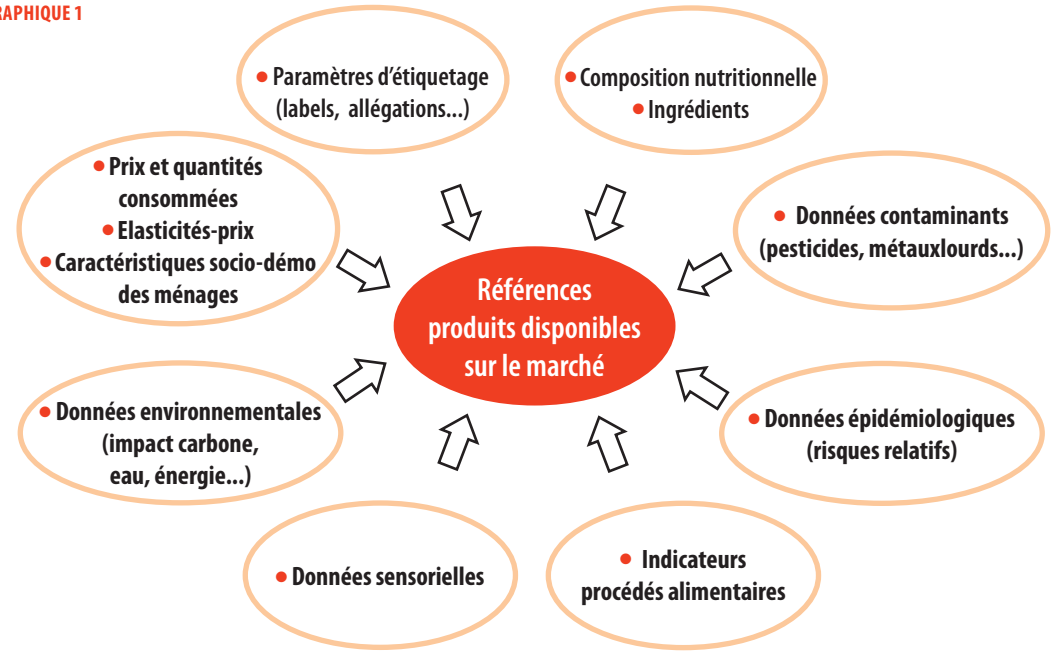
D'autres bases sont constituées par les équipes de recherche à partir de modélisations, d'expérimentations ou de revues de la littérature :

- données de biodisponibilité (UMR Nort),
- indicateurs de transformation industrielle (UMR GMPA),
- données sensorielles (ISG et UMR GMPA),
- valeurs des élasticités-prix (Unités Aliss et TSE).

A terme d'autres données pourraient être mobilisées comme les données épidémiologiques (risques relatifs) et les données sur les innovations-produits (bases Mintel-GNPD). Enfin, des liens avec des plateformes existantes pourront être développés (plateforme Means pour l'analyse de cycle de vie des produits ; ou plateforme PLASTIC pour la structures de produits de laboratoire).

Les droits d'accès à ces données varient selon les sources. Certaines sont publiques, d'autres sont soumises à des conditions particulières d'usage et peuvent nécessiter un accord du fournisseur. L'utilisateur est informé de ces droits d'accès sur la plateforme.

GRAPHIQUE 1



PRINCIPALES FONCTIONNALITÉS DE LA PLATEFORME

L'objectif principal de la plateforme est de mutualiser et de mettre à disposition des chercheurs ces bases de données (mise à disposition des données open source, ou orientation vers les fournisseurs pour les données soumises à des droits d'accès). Chaque jeu de données sera présenté par un ensemble d'informations permettant leur utilisation (Métadonnées, documentation, droits, licences). Des outils d'appariements des bases de données seront disponibles, permettant de lier les bases de données existantes entre elles et/ou de modifier des appariements existants entre deux bases en fonction des besoins de recherche. La mise en relation des données intégrées à la plateforme est un axe fort du système sur plusieurs aspects :

- Des standards de qualité seront mis en place à travers les outils permettant de relier les bases de données. Ils se traduiront notamment par un haut niveau de traçabilité des actions effectuées sur les données et leurs appariements, la mise en avant de procédures normalisées (incluant la mise en place de personnes expertes « valideurs » attestant la fiabilité des appariements ainsi créés), ainsi que la mise à disposition d'outils automatiques d'aide à l'appariement et à la vérification.

- Les données ainsi reliées permettront la génération de tout un nouveau champ d'information. En plus des appariements effectués, ces nouvelles facettes seront accessibles aux utilisateurs dans le respect de leurs droits individuels sur les données dont elles sont issues.

Il s'agit enfin de rendre possible l'intégration de nouveaux jeux de données au fil du temps. Le graphique 2 présente les bases retenues à ce stade et les connexions déjà réalisées.

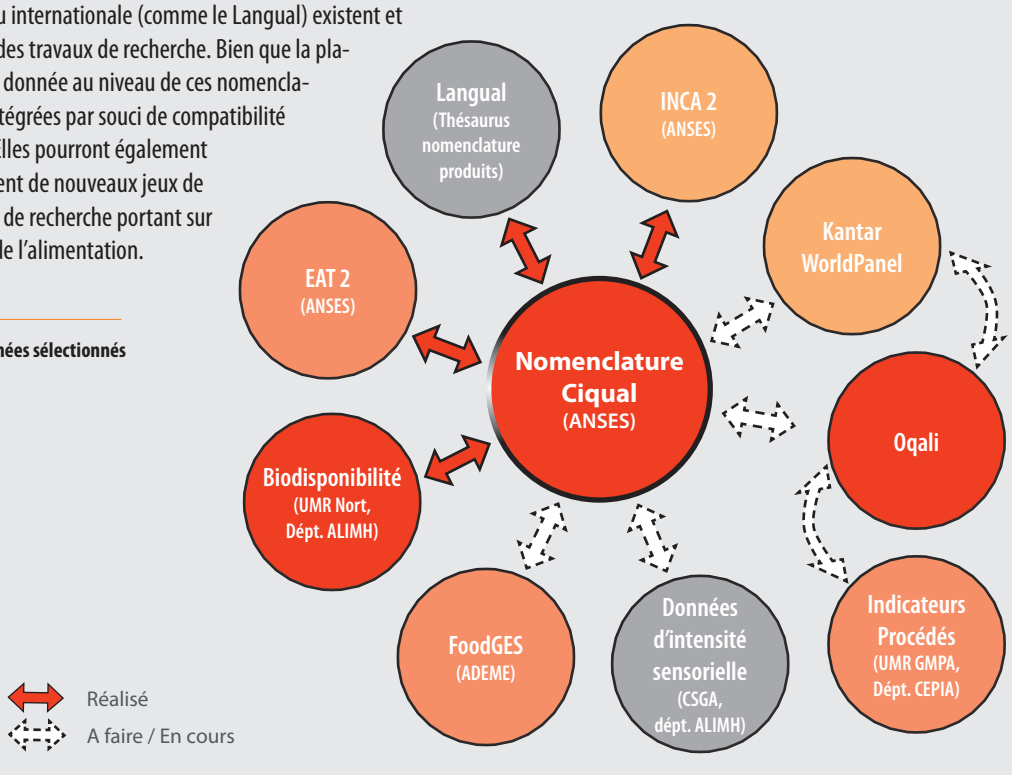
NOMENCLATURES

La difficulté dans la mise en relation de jeux de données provenant de sources différentes, se positionne au niveau de l'hétérogénéité de leurs nomenclatures aliments (différences de précision, d'agrégation et de hiérarchisation). Deux approches seront proposées aux utilisateurs concernant l'accès aux jeux de données et la manipulation de ces nomenclatures :

- La plateforme permettra par défaut d'accéder à l'ensemble des informations des bases de données par la nomenclature Ciqual (ANSES). Cette dernière représente l'ensemble de l'alimentation et est composée d'environ 1500 aliments (agrégés) proposant un bon équilibre entre la couverture de l'offre alimentaire et la précision de ses descriptions. Déjà très utilisée dans les travaux de recherche des différentes unités impliquées, elle a déjà fait l'objet d'appariements à d'autres bases (cf graphique 2) et apparait centrale dans les projets. L'ensemble des nomenclatures des autres jeux de données sera apparié sur cette dernière dans un système contrôlé par la plateforme.

- La plateforme laissera aussi la possibilité d'accéder aux informations qu'elle contient à travers toutes les nomenclatures qui y sont intégrées, grâce à un système d'appariements bidirectionnels. Il sera donc possible d'exporter de l'information provenant de Ciqual à travers une autre nomenclature comme le Langual , ou encore de positionner des informations provenant de l'Oqali dans la nomenclature de Kantar WorldPanel (les informations ainsi obtenues décriront un niveau plus fin, pouvant aller jusqu'à la marque du produit).

Des nomenclatures européennes ou internationale (comme le Langual) existent et sont de plus en plus utilisées dans des travaux de recherche. Bien que la plateforme ne possède encore aucune donnée au niveau de ces nomenclatures, elles seront tout de même intégrées par souci de compatibilité avec les systèmes internationaux. Elles pourront également permettre d'intégrer plus simplement de nouveaux jeux de données ou d'effectuer des travaux de recherche portant sur des comparaisons internationales de l'alimentation.



GRAPHIQUE 2 : Jeux de données sélectionnés

MODALITÉS PRÉVUES D'USAGE DE LA PLATEFORME

Cette plateforme a vocation à être utilisée et disponible pour l'ensemble des agents l'INRA travaillant sur une ou plusieurs dimensions de l'alimentation durable.

Ainsi, la liste des données et des outils disponibles pour manipuler les données sera accessible par tout agent INRA disposant d'un LDAP INRA, ainsi que tout autre partenaire par l'intermédiaire du site web de la plateforme MAD. Cette liste variera selon l'utilisateur en fonction des droits qui lui sont accordés, les différentes sections du serveur étant gérées avec des droits d'accès qui changent selon le demandeur. Il sera ainsi possible de définir des droits par centre, par département, par unité et par utilisateur et cela pour chacune des bases de données et applications indépendamment des autres. Les utilisateurs pourront avoir un aperçu du contenu des données à travers un ensemble de métadonnées descriptives et un Wiki de présentation générale qui leur permettront de déterminer si les données correspondent ou non à leur besoin. Pour accéder aux données, des droits supplémentaires seront délivrés sur demande, par le gestionnaire de la plateforme ou le propriétaire des données.
La plateforme et ses services bénéficieront d'un feedback à double sens pour d'une part, en améliorer la qualité des informations qu'elle contient et d'autre part, optimiser leurs usages par les scientifiques. Les utilisateurs pourront à tout moment faire remonter des erreurs ou leurs remarques (sur les données et/ou appariements) et la plateforme aura la possibilité de diffuser de l'information ciblée.

EQUIPES ET FINANCEMENTS

L'ensemble de cette démarche s'appuie sur un travail initialement mis en place dans le cadre du Pôle Alimentation Parisien (PAP) et financé par le CPER Ile-de-France (2008-2012). Cette opération portée par l'unité ALISS avait mobilisé d'autres unités comme l'UREN, l'UMR Genial et Met@risk. La démarche s'appuie également sur des travaux préalables conduits par certaines unités (ALISS, TSE, NORT) sur des bases de données particulières et sur certains outils d'appariements, et développés dans le cadre de projets ANR et européens récents ou en cours (AlimInfo, OCAD, Susdiet…).

Les avancées récentes ont été conduites dans le cadre du métaprogramme Did'it qui a financé l'achat de certaines données et l'animation d'un groupe de travail visant à préciser les objectifs et les modalités techniques de fonctionnement de la plateforme.

Coordination : L.G. Soler
Coordination technique : C. Boizot-Szantai, N. Guinet, O. de Mouzon, V. Orozco, F. Stevenin
Unités INRA impliquées dans l'élaboration de la plateforme MAD : UR Aliss (Ivry-sur-Seine), UMR TSE (Toulouse), UMR GMPA (Grignon), UMR Nort (Marseille), CSGA (Dijon)