Mise en oeuvre d'une base de données accessible via une interface Web. Exemple sur des données économiques

Nadine Herrard, Magalie Houée Bigot

Cah. Tech. I.N.R.A., 2006 57, 31-46

Mise en œuvre d'une base de données accessible via une interface Web Exemple sur des données économiques

Nadine Herrard[1], *Magalie Houée-Bigot*[1]

Les travaux de recherche en économie appliquée ont la spécificité d'être basés, non pas sur des données expérimentales, mais sur des données observées. Prenons l'exemple du domaine agricole où les données peuvent porter sur les caractéristiques de l'exploitation, sur les individus, sur la situation financière, etc. Les données observées peuvent être de plusieurs types : données individuelles (microéconomiques), données macroéconomiques (croissance économique, taux de change, etc.), données agrégées au niveau d'un pays par exemple (production de blé en France pour une année).
Généralement deux dimensions caractérisent ces données : la périodicité (un trimestre, une année, …), et le type (une personne, une entreprise, un agrégat, un pays, un produit, etc.).
L'objectif ici est de décrire l'outil que nous avons développé pour répondre à une demande formulée par des chercheurs en économie. Cette demande est basée sur la mise à disposition d'une base de données homogène et cohérente avec la théorie économique et sur un accès aux données via un intranet. Un système d'information regroupant le traitement et l'exploitation des données est développé afin de faciliter le travail de recherche.
La première partie décrit la demande, les différents besoins des chercheurs et la deuxième partie présente les problèmes et les solutions pour répondre à cette demande.

Mots-clés

Système d'information, base de données, séries chronologiques, interface web, langage Perl, logiciel SAS

1. Contexte

1.1. Les besoins des utilisateurs

La demande formulée émane d'économistes et plus précisément de chercheurs en économie appliquée au domaine agricole. Les recherches menées dans ce domaine portent sur le développement d'outils d'aide à la décision ce qui nécessite d'étudier le comportement économique des individus. Les chercheurs utilisent des données observées afin d'étudier les comportements des individus ou le fonctionnement des marchés agricoles. Les données recherchées sont des séries temporelles (définies sur la période passée, par exemple 1960, et jusqu'à nos jours) et elles concernent plusieurs champs : un produit tel que le blé, le maïs ou le soja, un indicateur tel que la quantité produite, exportée, et un pays.

[1] INRA-ESR - 4 allée Adolphe Bobierre - CS 61103 - 35011 Rennes Cedex - Nadine.Herrard@rennes.inra.fr - Magalie.Houee@rennes.inra.fr - http://www.rennes.inra.fr/economie/index.htm

Les chercheurs ont des besoins de nature différente quant à l'utilisation de ces données. Les demandes peuvent être de la forme suivante :

- consultation de données brutes : retracer l'évolution d'une série ou bien connaître la valeur d'une donnée à une date particulière ;
- étude statistique des données : calculs d'indicateurs (par exemple calcul de la part de la production française de blé par rapport à la production mondiale à une date précise ou sur une période définie) ;
- récupérer les données afin de les utiliser pour des travaux de recherche.

Les données doivent avoir un sens du point de vue économique et donc vérifier, en plus des contraintes « classiques » de nettoyage des données (valeurs aberrantes, valeurs manquantes, etc.), des contraintes de nature économique. En effet, les chercheurs étudient les comportements des acteurs sur les marchés agricoles, et plus précisément les équilibres de marchés (équilibre entre l'offre et la demande). Les données doivent, par conséquent, être équilibrées au niveau national et au niveau du commerce international (par exemple, la somme des exportations doit être égale à la somme des importations).
Les utilisateurs ont également besoin d'échanger et de diffuser leurs résultats d'où la nécessité d'avoir un outil permettant de mettre en forme les données (convivialité de l'accès aux données).
Par ailleurs, un élément crucial à considérer est la pérennité de l'outil. Les données sont pour l'essentiel des données annuelles, elles seront donc renseignées ou complétées chaque année. Le système d'information doit prendre en compte le problème d'actualisation de la base de données sans que cela nécessite un remaniement de l'outil mis en place.

Chaque exigence ou besoin du chercheur est un élément du système d'information nécessitant un traitement particulier de l'accès aux données. Le système d'information doit permettre de stocker des données multidimensionnelles, d'homogénéiser les données (cohérence économique), de faciliter l'accès et le traitement des données selon l'objectif du chercheur (étude statistique, consultation de données brutes, etc.) et de permettre l'actualisation de la base de données et l'ajout de nouvelles données.

La base de données, cœur du système d'information, est composée de deux sous-ensembles : un ensemble céréales et un ensemble oléagineux.
De nombreux organismes diffusent des données économiques agricoles mais ces dernières ne sont pas toujours exploitables directement.Pour ces raisons de nombreux instituts ont investi dans la construction de bases de données spécialisées permettant ainsi aux utilisateurs de disposer de données homogènes, validées et définies sur une longue période. Dans ce cadre, le logiciel SAS[2] est souvent utilisé. Ce logiciel procure les outils essentiels à la réalisation des tâches d'une application de traitement de données : accès aux données, gestion, analyse et présentation des données.L'accès aux bases de données SAS nécessite de connaître le langage SAS ; c'est pourquoi une solution conçue pour s'intégrer aux nouvelles architectures informatiques avec des postes clients légers présente beaucoup d'intérêt. De plus les chercheurs de l'unité n'utilisent pas nécessairement les mêmes logiciels et le système d'information développé ne doit pas contraindre l'utilisateur à manipuler un nouveau logiciel.

[2]SAS *Statistical Analysis System*. SAS est un logiciel d'analyse statistique, économétrique et de recherche opérationnelle qui possède de puissants outils pour la gestion des données, le calcul matriciel et la programmation d'applications graphiques. A l'INRA, le Système SAS reste un outil statistique de référence même si actuellement l'éditeur se présente plus comme fournisseur de plates-formes décisionnelles d'entreprise.

Dans un premier temps, il s’agit de construire la base de données et de réaliser les programmes qui en assureront la cohérence. Ces programmes doivent être réutilisables afin d'être employés à chaque mise à jour, ce qui implique une étude préalable de l'ensemble des problèmes qui surviendraient lors des mises à jour.
Dans un deuxième temps, nous décrivons la mise en œuvre du système d’information permettant une interaction dynamique du système SAS avec le Web ; cette solution permet de mettre à disposition de tous les utilisateurs les fonctionnalités du système SAS les plus couramment utilisées tels que l’accès aux données, l’analyse statistique, l’exportation des données, et ce sans connaissances particulières du langage SAS.
Afin de faciliter la consultation de la base de données, une interface d'interrogation accessible via un intranet a été développée. Les utilisateurs souhaitent, en particulier, disposer de fiches synthétisant le bilan économique d'un pays ou d'une culture pour une période donnée.

1.2. Présentation générale du système d'information

Nous définissons le système d’information comme l’ensemble des moyens (organisation, acteurs, procédures, outils informatiques développés) nécessaires au traitement et à l’exploitation des données.
Les données sont au cœur des projets de recherche en économie. Ces données proviennent de différentes institutions et ne sont pas toujours utilisables directement par les chercheurs. L'intérêt du système d’information est de mettre à la disposition des utilisateurs des données cohérentes validées et de leur fournir une interface de consultation conviviale et facile d’utilisation.
Le procédé doit être automatisé dans la mesure où les données seront actualisées annuellement.

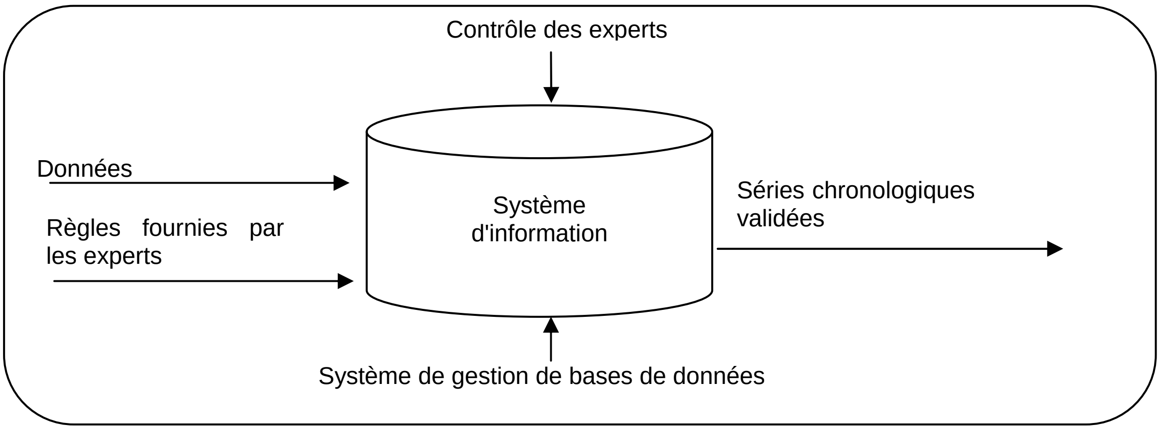


Figure 1. Flux d’information gérés par le système

Comme le montre la **figure 1**, les entrées de ce système d'information sont doubles ; d'une part, des données correspondant à une sélection des données publiées par diverses institutions et, d'autre part, des règles d'expertise correspondant aux manipulations que nous souhaitons appliquer à ces données pour en assurer leur cohérence. Les sorties correspondent aux données validées attendues par les utilisateurs.

La réalisation de ce système d'information s'appuie sur le système de gestion de bases de données SAS.

1.2.a. Les données

L'objectif consiste à mettre en place un système d'information dont la base de données est composée de deux ensembles : un ensemble céréales et un ensemble oléagineux

Les données correspondent à une sélection des données brutes publiées par diverses institutions. Elles sont issues de la base PS&D Online[3] du ministère de l'Agriculture des USA (USDA) ; Ces données sont accessibles via le site web http://www.fas.usda.gov/psd/ . Cette base économique de données décrit les faits relatifs à la production, à la consommation et aux échanges des produits issus de l'agriculture, de l'élevage, de la pêche de chaque pays. Elle couvre un large éventail de pays (environ 220 pays) ; cette caractéristique est une des raisons qui ont amené les experts à la choisir plutôt qu'une autre.

Les données importées sont représentées sous forme de séries chronologiques annuelles couvrant généralement les années 1960 à 2003. Une série chronologique est caractérisée par un identifiant unique, un calendrier définissant le rythme des mesures et une liste de valeurs successives. A ces caractéristiques de base, nous pouvons ajouter d'autres caractéristiques qui nous permettent de mieux définir la série : son unité d'expression des valeurs, sa date de saisie, son degré de fiabilité, etc.

Les séries chronologiques qui sont stockées dans la base de données sont identifiées par un triplet de valeurs : le nom de pays, le produit (maïs, blé, soja, …) et la caractéristique d'intérêt (quantité produite, importée, …). L'unité d'expression est renseignée dans le libellé des variables.

Actuellement les données intégrées au système d'information sont ventilées comme suit :

- 43 années (1960 à 2003)
- 245 pays, ou zones géographiques
- 16 produits (9 pour les céréales, 7 pour les oléagineux)

1.2.b. Règles fournies par les experts

La qualité des données est d'une importance primordiale en économie appliquée. Afin de s'assurer de la véracité des données utilisées, il importe de les vérifier avant leur utilisation effective dans les projets de recherche. On va donc appliquer des règles qui résultent de la connaissance experte des statisticiens économètres.

De manière très générale, pour définir la cohérence des données nous utilisons 2 critères : i) l'interprétation économique (nullité, négativité, apparition de nouveaux pays, etc.) ; ii) le respect des équilibres de marchés.

i) Interprétation économique

- La base de données doit stocker des séries chronologiques ne présentant pas de rupture de séries dans les années.
- Evolution géopolitique : certains pays connaissent un changement dans leurs frontières, soit en s'unissant avec d'autres pays, soit en se divisant en plusieurs pays. C'est le cas par exemple de la réunification de l'Allemagne en 1990 ou de l'éclatement de l'Union soviétique en 1986. Dans ces cas, de nouveaux pays sont créés, on doit donc s'assurer que les données sont présentes uniquement sur la période de leur existence et que variables liées les unes aux autres dans le temps sont cohérentes avec

[3] Production, Supply and Distribution, système de diffusion de données internationales sur la production et la distribution des produits agricole de base proposées par l'United States Department of Agriculture.

l'évolution de leur situation géopolitique. Par exemple, la somme des stocks finaux de la RFA et de la RDA de 1989 devra être égale au stock initial de l'Allemagne unifiée de 1990)

- Valeurs négatives : les variables numériques décrivent chacune un fait économique : quantité produite, importée, exportée… Ainsi une valeur négative correspond à une situation aberrante.
- Valeurs aberrantes : une valeur trop élevée ou trop faible peut s'avérer être une valeur aberrante. Nous avons besoin de détecter ces valeurs, sans pour autant les corriger nécessairement, c'est « l'expert », qui après son analyse décidera de la correction ou non.

ii) Respect des équilibres de marché

Cette étape permet de valider les données du point de vue économique avant leur mise à disposition à des fins opérationnelles. Cette vérification a lieu seulement après l'exécution de toutes les autres vérifications et la correction subséquente des données.

- La première contrainte économique consiste à équilibrer l'offre et la demande : pour chaque occurrence (pays, produit, année), la somme du stock initial et des quantités produites et importées doit être égale à la somme du stock final et des quantités exportées et consommées.
- La seconde contrainte consiste à présenter une situation équilibrée entre les quantités mondiales importées et exportées.

1.2.c. Utilisation du logiciel SAS

Dans le cadre de la mise en œuvre de ce système d'information, nous avons donc besoin d'un gestionnaire de base de données, d'un outil statistique, d'un outil de programmation pour automatiser les mises à jour de la base et d'un outil d'édition dans le cadre de la publication de rapport. Nous avons choisi le système SAS. Il permet de gérer efficacement un grand volume de données. Ses fonctionnalités répondent à nos besoins en termes de manipulation de données, d'édition de rapports, d'analyse de données, de visualisation interactive des résultats, sans oublier les fonctionnalités client/serveur. SAS est par ailleurs considéré comme un logiciel statistique de référence.

Pour la création de la base de donnée, nous utilisons donc le module SAS/Base. Ce module permet de lire les données à partir d'un grand nombre de types de fichiers, y compris les fichiers à enregistrements de longueur variable, en format libre et comportant des données manquantes. Ses procédures de gestion de données font que SAS est aussi un véritable langage de programmation, ainsi, on peut manipuler les données de manière très efficace. SAS inclut des fonctionnalités SQL[4]. La procédure SQL est un outil d'interrogation et de manipulation des données permettant d'exprimer des requêtes d'interrogation et de gestion des données au sein des programmes SAS dans un langage normalisé. Ce langage va nous permettre d'effectuer le traitement sur les données de manière sélective au moyen de spécifications basées sur des valeurs ou des plages de valeurs caractérisant les données (attributs).

SAS dispose également d'un module macro (SAS/Macro). A l'aide d'un langage basé sur des macro-commandes (commandes paramétrables à l'aide de chaînes de caractères dont le

[4] "Structured Query Language" est un langage d'interrogation normalisé de base de données relationnelles.

contenu peut varier en fonction du contexte d'application), on va pouvoir automatiser les programmes de mise à jour, générer du code SAS et paramétrer efficacement les programmes afin de généraliser leur utilisation.

2. Création du système d'information

2.1. La base de données

La base de données est au cœur du système d'information et les données stockées devront être validées. De façon générale, pour permettre le passage de l'acquisition des données à leur utilisation effective dans différents types d'études, on distingue les étapes suivantes : acquisition, traitement, contrôle et validation.
Plusieurs programmes SAS ont été écrits dans le but de nettoyer et d'équilibrer la base de données. Des macros SAS ont été réalisées pour corriger les données présentant une anomalie. Ces macros présentent deux avantages :
- appliquer la même correction aux observations détectées par les programmes de nettoyage et d'équilibrage dans le même ensemble de données ;
- être réutilisables pour chaque ensemble de données et faciliter la création d'autres ensembles de données (Il est prévu d'ajouter les données PS&D relatives à l'élevage et aux produits laitiers dans la base).

Chacune des étapes décrites ci-dessous fait ainsi l'objet d'une macro SAS. L'ensemble des traitements et corrections effectuées sur les données est consigné dans des fichiers « journaux » (fichier texte)

2.1.a. <u>Acquisition de l'information</u>
L'acquisition de données consiste à collecter l'information à partir des fichiers fournis par les différents organismes statistiques. Ces organismes changent quasiment tous les ans le format des fichiers, aussi, il n'est pas envisagé d'automatiser cette étape. Dans le cadre des mises à jour annuelles, nous pensions pouvoir récupérer uniquement les trois dernières années pour chaque série, il s'avère que des données sont modifiées au delà. Pour cette raison, nous avons décidé de récupérer tous les ans, l'ensemble des données.

2.1.b. <u>Traitement préalable des données : nettoyage des données</u>
L'objectif de cette étape est de tester la validité des données et de les corriger si nécessaire.
Le critère de cohérence appliqué à ce stade est l'interprétation économique ; cette première étape va consister à « nettoyer » les données pour rendre pertinente leur interprétation économique. Lors de cette opération, on introduit des indices de qualité de la donnée ainsi que des indices indiquant que celle-ci est reconstituée, calculée, manquante, définie par les experts (statisticiens économètres, économistes), etc.

-***Les séries chronologiques*** ne doivent pas présenter de rupture de série dans les années. Une macro SAS a donc été développée afin de créer les observations correspondantes aux années manquantes que l'on initialise à une valeur non renseignée (symbolisée par le point sous SAS).

-***Evolution géopolitique*** : pour chaque pays ayant connu une évolution géopolitique on vérifie la cohérence des variables avec une macro SAS.

-***Détection et correction des valeurs négatives :*** Les valeurs négatives sont remplacées par la moyenne arithmétique des valeurs en i-1 et i+1 si ces dernières sont positives. Dans le cas contraire on se réfère aux valeurs en i-2 et i+2, sinon l'expert analyse la série et propose une correction. Une macro SAS appliquée à chaque observation décelée a été développée selon l'algorithme de correction défini par l'expert. Cette macro est paramétrable pour pouvoir être réutilisée quelle que soit la série chronologique concernée.

Les paramètres de la macro de correction sont : i) le nom de la structure de données (« dataset ») contenant les données à corriger ; ii) le nom de la structure de données contenant les données corrigées ; iii) le code pays ; iv) le code produit ; v) l'année ; vi) le nom de la variable.

- ***Détection et correction des valeurs aberrantes*** : Pour repérer des valeurs aberrantes, il existe plusieurs méthodes statistiques. La méthode que nous avons retenue consiste à filtrer les données selon le critère de Tuckey. Ce critère considère comme aberrantes, toutes les données en dehors de l'intervalle

$$[Q1 - 1.5 * (Q3 - Q1); Q3 + 1.5 * (Q3 - Q1)]$$

On applique ce critère sur la série exprimée en variation annuelle. Pour accompagner cette méthode de repérage, on visualise les graphiques correspondant aux séries contenant au moins une valeur détectée aberrante par la méthode décrite ci-dessus grâce au module SAS/GRAPH. L'expert choisit ensuite de la modifier ou non. En cas de modification, la macro SAS appelée dans le cadre de la correction des valeurs négatives est appliquée à l'observation concernée.

2.1.c. <u>Contrôle et validation des contraintes d'équilibres de marchés</u>

Cette étape permet de valider les données avant leur mise à disposition à des fins opérationnelles.

La première contrainte économique consiste à équilibrer l'offre et la demande : la somme du stock initial et des quantités produites et importées doit être égale à la somme du stock final et des quantités exportées et consommées. En cas de déséquilibre, on utilise les stocks comme variable d'ajustement.

Chaque modification est contrôlée par l'expert de façon à s'assurer que les changements apportés n'entraînent pas de trop grandes variations. Dans certains cas (déterminés par l'expert), on modifiera une autre variable : la consommation totale.

La seconde contrainte économique consiste à présenter une situation équilibrée entre les quantités mondiales importées et exportées.

Pour chaque année et chaque produit, on calcule la différence entre la somme des importations et la somme des exportations de l'ensemble des pays. Il y a déséquilibre si cette différence est non nulle. Dans ce cas, on crée un pays fictif « reste du monde » avec des valeurs permettant d'annuler cette différence tout en respectant les autres contraintes liées à l'ensemble de données. Ce pays fictif n'est pas mis à la disposition des utilisateurs, son seul rôle est d'équilibrer la base.

2.2. L'interface de consultation

L'objectif est de créer un outil accessible via un client léger (navigateur web) dont les principales fonctions sont :

- de faciliter l'interrogation de la base de données par les scientifiques, notamment de pouvoir créer ses propres agrégats (pays ou produits)
- d'offrir des fiches statistiques synthétisant l'information sur un pays ou sur un produit.

2.2.a. <u>Environnement de travail et outils utilisés</u>

L'unité dispose d'un serveur de calcul scientifique sur lequel le logiciel SAS est installé et d'un serveur web à partir duquel toutes les applications intranet sont accessibles.

L'interface d’interrogation est codée en langage HTML. Le traitement des requêtes est effectué à partir du langage Perl. L’interrogation de la base de données s'effectue grâce au langage SAS et les résultats attendus par l'utilisateur sont stockés dans des pages HTML que SAS génère grâce à la fonctionnalité ODS (« Output Delivery System »). ODS est un ensemble de commandes SAS qui permet de gérer automatiquement les sorties SAS dans différents formats (HTML, postscript, etc.). L’affichage de statistiques sous forme graphique est réalisé à partir du pilote graphique Activex du module SAS/GRAPH. Le pilote ActiveX génère une page HTML contenant l’appel à Activex. Pour que le code soit correctement interprété par les navigateurs, il faut que le composant Control ActiveX SAS/GRAPH soit installé sur le PC. Ce composant va permettre d'insérer des graphiques interactifs dans des pages Web, il est utilisé en conjonction avec l'ODS.

2.2.b. Fonctionnement de l’application

La **Figure 2** ci-dessous représente l’architecture et le fonctionnement global de l’application développée.

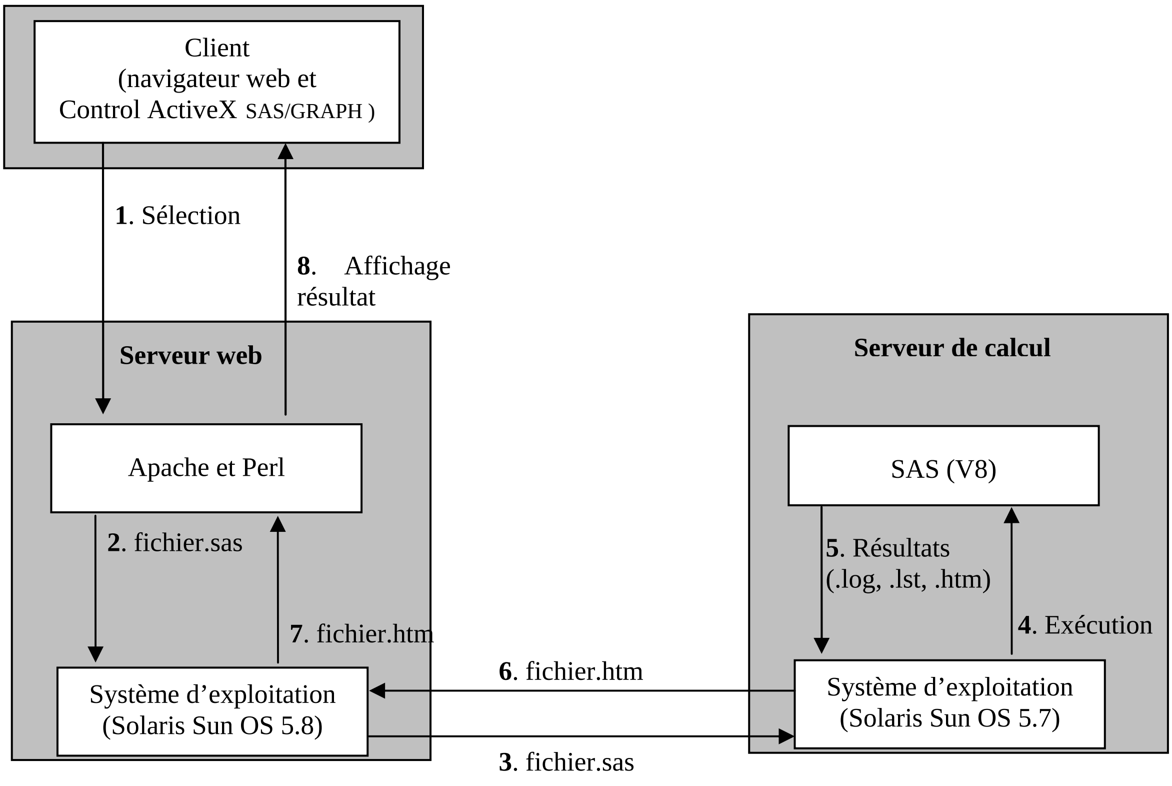


Figure 2. Configuration générale de l’application de consultation

1. Le programme PERL qui va générer l’interface web au format html est stocké sur le serveur web.
2. Lorsque l’on valide la sélection, un autre programme PERL va générer un programme sas qui sera alors envoyé sur le serveur de calcul. Le serveur www demande alors l’exécution de ce programme à distance (commande « rsh ») et récupère les fichiers générés par SAS (**.**log et **.**lst et **.**htm) et les affiche sur le client.
3. Un utilisateur avec le login « bdweb » a été créé sur le serveur de calcul pour pouvoir exécuter les programmes SAS. Pour que les deux serveurs puissent communiquer, on

utilise les fonctionnalités Unix du fichier **.**rhost de l'utilisateur « bdweb » pour permettre à l'utilisateur www l'exécution des programmes à distance.

4. Le programme SAS est exécuté sur le serveur de calcul

5. Le résultat de la requête est stocké dans un fichier suffixé par htm

6. Ce fichier htm est récupéré sur le serveur web

7. Le serveur web le rend disponible pour le client

8. Le fichier est chargé sur le poste client.

2.2.c. Contrôle de la concurrence d'accès

L'application ne pose aucun problème de concurrence d'accès car les programmes générés sont suffixés par l'heure de connexion. Ainsi la précision de l'heure de connexion dans les noms de fichiers joue un rôle très important. En effet, imaginons qu'un utilisateur se connecte à 8h50min33s et qu'un deuxième utilisateur souhaite utiliser l'interface en se connectant à 8h50min53s, soit 20 secondes plus tard. L'heure de connexion étant différente, la requête du premier utilisateur n'interfère pas avec celle du second, les fichiers de requête et de résultat du premier utilisateur ne sont pas écrasés par ceux du deuxième utilisateur. Cette solution a évidemment ses limites lorsque deux utilisateurs se connectent à la même heure. L'application actuelle n'est pas utilisée à une telle fréquence si bien que le problème ne se pose pas. Une évolution possible en cas d'augmentation de la fréquence consisterait à accroître la précision de l'heure à la milliseconde près.

2.2.d. Description de l'interface

La page d'accueil de l'interface est présentée par la **figure 3**.

Un premier choix est imposé à l'utilisateur : exploiter l'ensemble céréales ou l'ensemble oléagineux.

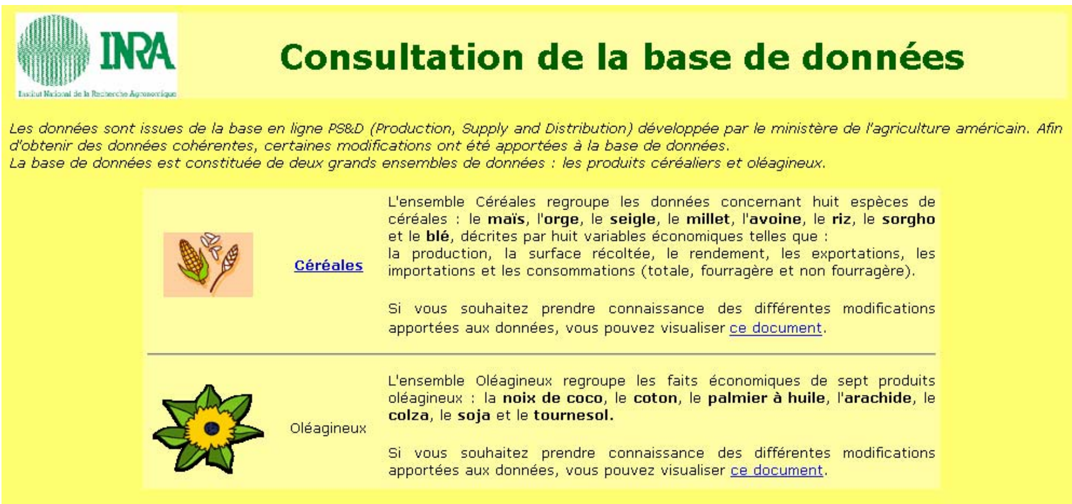


Figure 3 : *Page d'accueil de l'interface*
(http://eco2.roazhon.inra.fr/Intranet/Unites/RennesESR/bdstat/pagePrincipale.htm)

L'aide offerte sur cette page HTML (accessible en cliquant sur « *ce document* ») est composée de deux documents Word expliquant et présentant les modifications apportées aux données lors de la phase de nettoyage.

Lorsque l'utilisateur a choisi l'ensemble sur lequel il souhaite travailler, une autre fenêtre s'ouvre **(Figure 4**) et propose les différents traitements qu'il est possible d'effectuer.

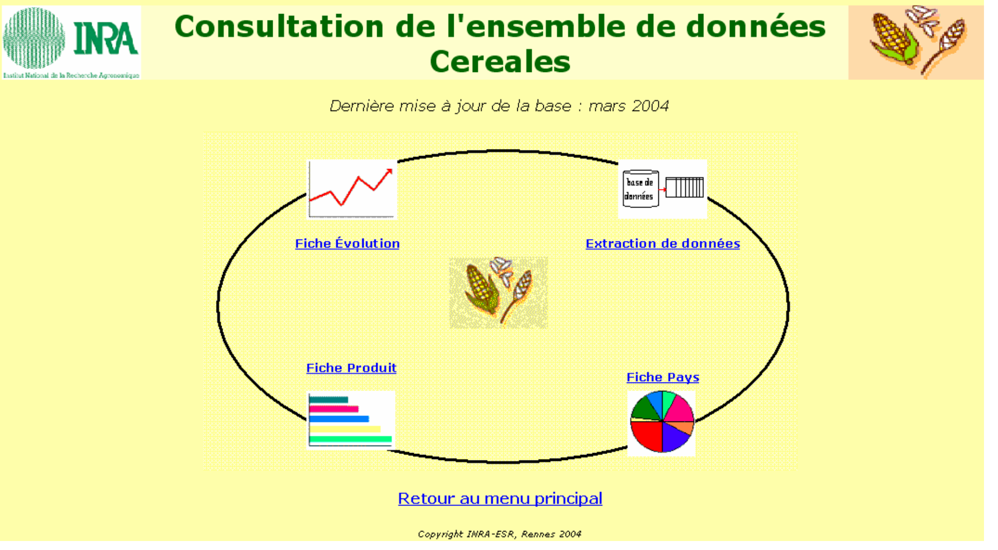


Figure 4 : *Menu de l'ensemble Céréales*

1. Le lien « *extraction de données* » permet, comme son nom l'indique, d'extraire un ensemble de données de la base suivant les besoins de l'utilisateur. L'utilisateur n'aura pas besoin de connaître le langage SQL puisqu'il choisira les pays, les produits, les variables économiques et la période dans des listes prédéfinies.
2. Le lien « *fiche pays* » permet d'obtenir une synthèse de l'état économique des produits céréaliers d'un pays sur une période.
3. Le lien « *fiche produit* » permet d'obtenir une synthèse de l'état économique d'une céréale dans le monde sur une période.
4. le lien « Fiche évolution » permet de suivre l'évolution d'une série pour un pays donné, un produit donné et sur une période donnée à partir d'un graphique.

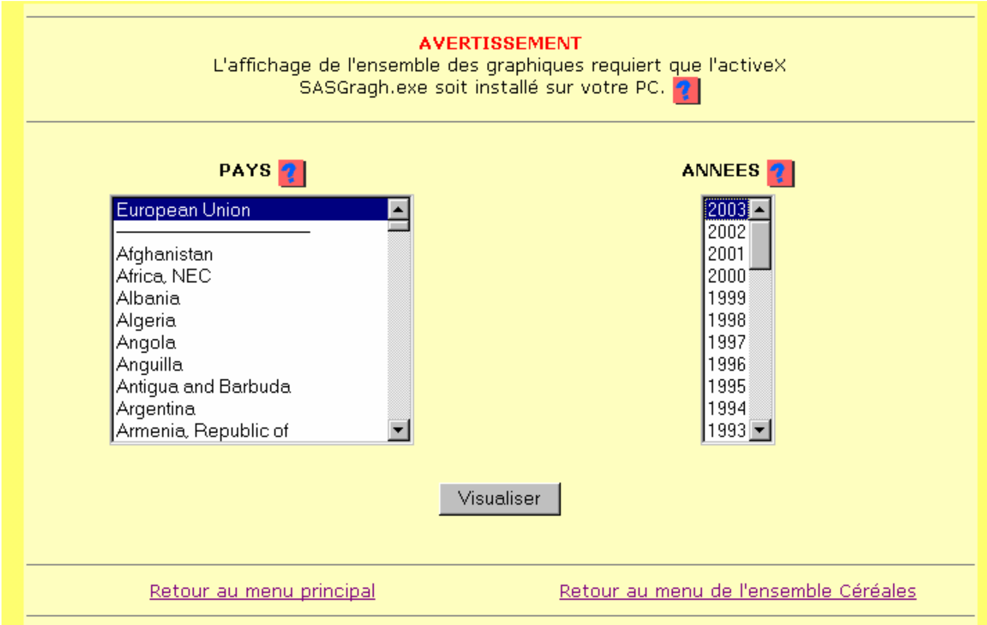
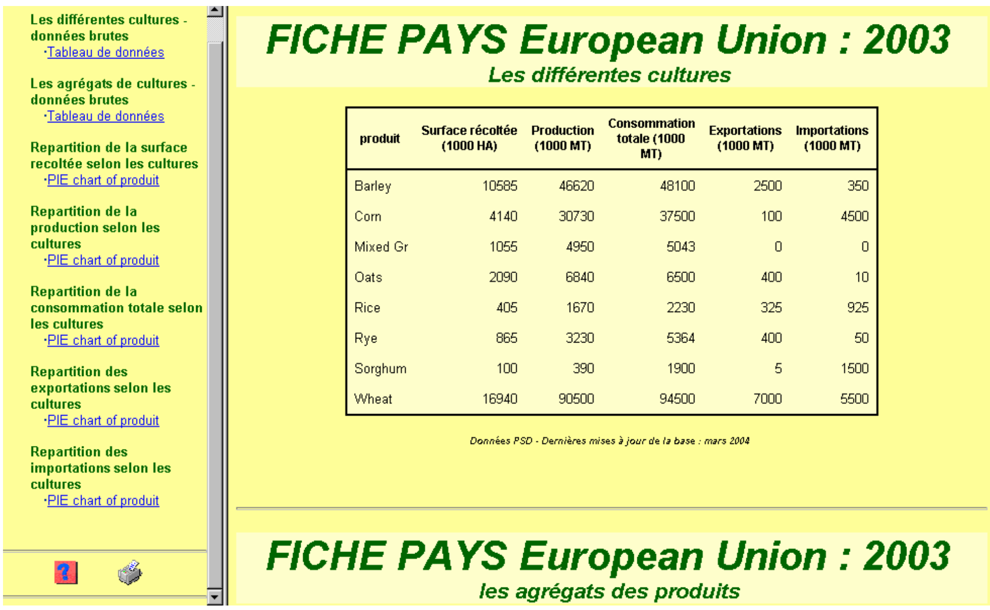


Figure 5 : *Création d'une fiche Pays*

2.2.e. <u>Consultation des données pour un pays</u>

Si l'utilisateur souhaite obtenir une synthèse de l'état économique des produits céréaliers, il cliquera le second choix de l'interface décrite dans la **Figure 4**. Ce choix va déclencher l'ouverture d'une page (**figure 5**) lui permettant de choisir le pays et la période de la synthèse souhaitée.

La sélection de « *European Union* » dans la liste des pays et la sélection de « *2003* » dans la liste des années donne comme résultat la fenêtre présentée par la **Figure 6**.



produit	Surface récoltée (1000 HA)	Production (1000 MT)	Consommation totale (1000 MT)	Exportations (1000 MT)	Importations (1000 MT)
Barley	10585	46620	48100	2500	350
Corn	4140	30730	37500	100	4500
Mixed Gr	1055	4950	5043	0	0
Oats	2090	6840	6500	400	10
Rice	405	1670	2230	325	925
Rye	865	3230	5364	400	50
Sorghum	100	390	1900	5	1500
Wheat	16940	90500	94500	7000	5500

Figure 6 : *Extrait d'une fiche Pays*

On dispose des données numériques ainsi que les graphiques associés. L'aide disponible en cliquant sur décrit les possibilités apportées par le pilote graphique ActiveX. L'icône explique comment imprimer les résultats contenus dans le fenêtre de droite.

2.2.f. Consultation des données pour un produit

On peut obtenir une fiche de synthèse pour un produit donné selon le même principe. Un formulaire **(Figure 7)** va permettre à l'utilisateur de choisir la catégorie de céréales et la période qui l'intéressent. Le nombre de pays étant conséquent dans la base, il est possible d'afficher sur les graphiques que les « n » plus grandes surfaces récoltées, les « n » premiers exportateurs, importateurs, consommateurs et producteurs du type de céréales en question.
Ce « n » correspond au chiffre que l'utilisateur doit choisir dans la liste « nombre de pays ».

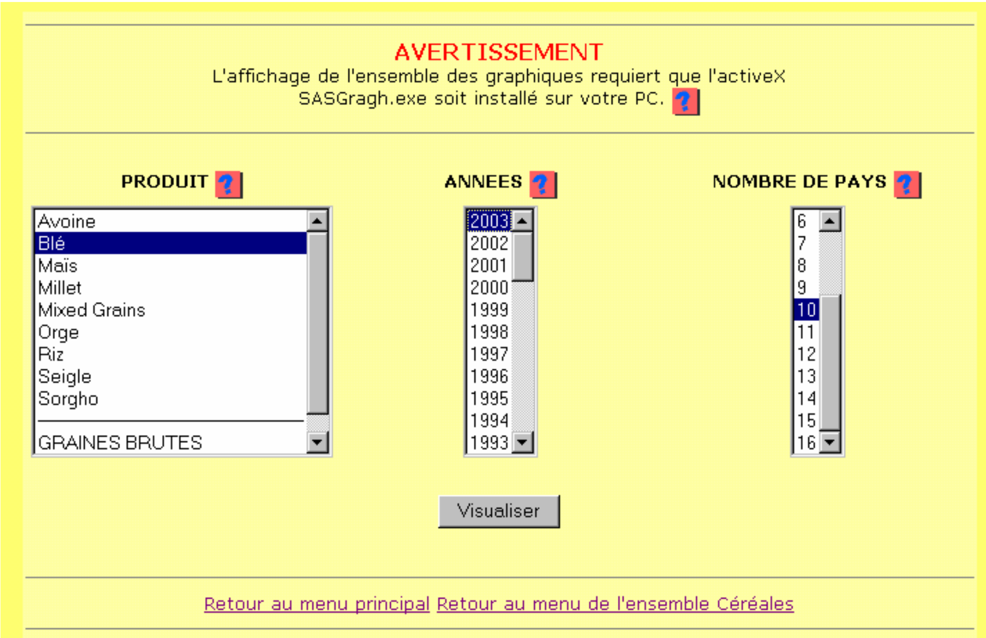


Figure 7 : *Création d'une fiche Produit*

Un extrait du résultat obtenu est présenté sur la **Figure 8**. L'aide disponible est la même que celle disponible sur une fiche pays.
Les graphiques sont interactifs. Ainsi l'utilisateur peut modifier le style de graphique, le sauvegarder, etc. en faisant un « clic droit » lorsqu'il est positionné dessus. Ainsi, il pourra insérer les graphiques qu'il souhaite dans ses publications **(Figure 9)**.

FICHE PRODUIT Blé : 2003 : Les 10 premières productions dans le monde

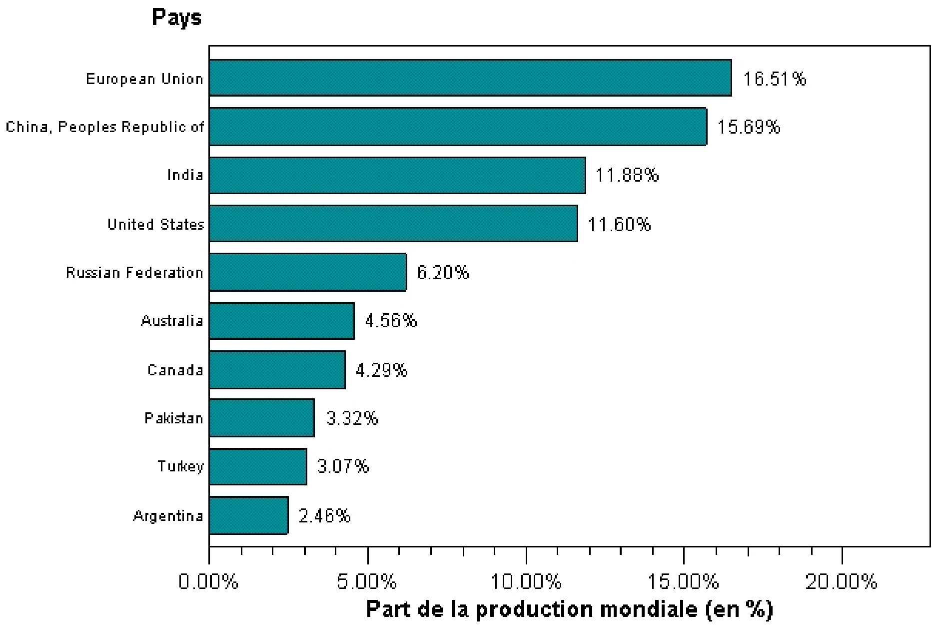


Figure 8 : *Extrait d'une fiche produit*

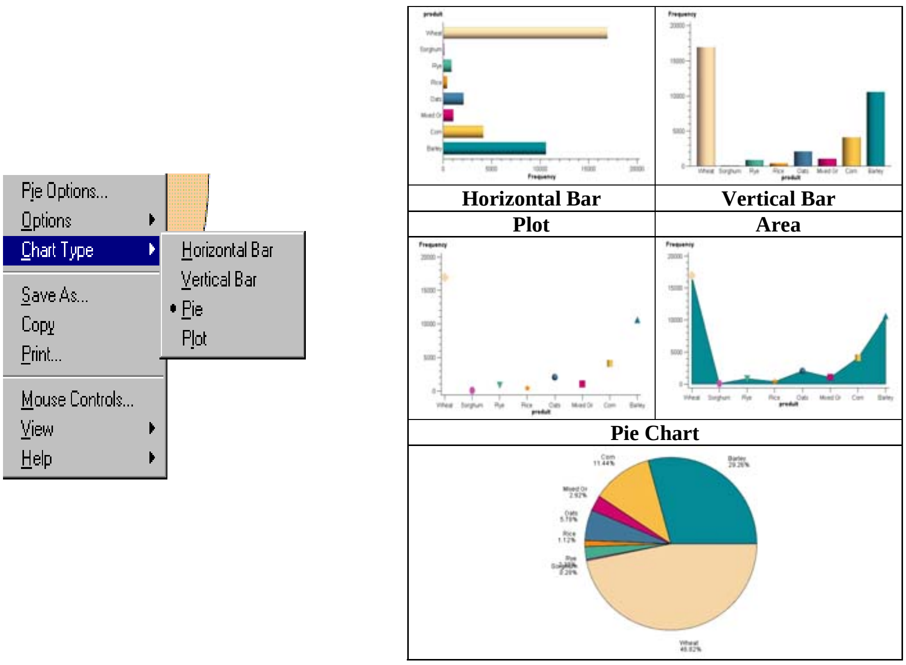
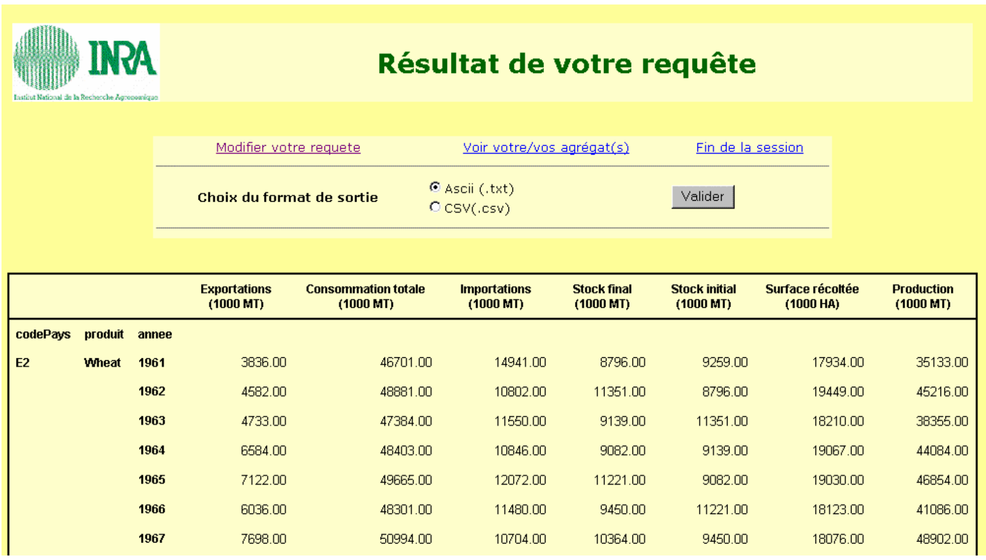


Figure 9 : *Graphique interactif*

2.2.g. <u>Extraction de données</u>.

Après avoir consulté les fiches statistiques, l'utilisateur peut approfondir son analyse et produire ses propres statistiques. Pour cela, il doit interroger la base de données pour extraire ses données et pouvoir ensuite les traiter avec son logiciel préféré. Dans un premier temps, il va sélectionner les pays, les produits céréaliers, les variables économiques et la période sur laquelle on souhaite obtenir des informations pour visualiser les données à extraire (**Figure 10**). Dans un deuxième temps, il choisit l'organisation de son tableau en spécifiant les éléments à mettre en ligne et en colonne. Les données sont ensuite affichées dans une page HTML sous forme d'un tableau généré par le système SAS. (**Figure 11**)

Figure 10 : *Interface d'extraction des données sur l'ensemble céréales*



codePays	produit	annee	Exportations (1000 MT)	Consommation totale (1000 MT)	Importations (1000 MT)	Stock final (1000 MT)	Stock initial (1000 MT)	Surface récoltée (1000 HA)	Production (1000 MT)
E2	Wheat	1961	3836.00	46701.00	14941.00	8796.00	9259.00	17934.00	35133.00
		1962	4582.00	48881.00	10802.00	11351.00	8796.00	19449.00	45216.00
		1963	4733.00	47384.00	11550.00	9139.00	11351.00	18210.00	38355.00
		1964	6584.00	48403.00	10846.00	9082.00	9139.00	19067.00	44084.00
		1965	7122.00	49665.00	12072.00	11221.00	9082.00	19030.00	46854.00
		1966	6036.00	48301.00	11480.00	9450.00	11221.00	18123.00	41086.00
		1967	7698.00	50994.00	10704.00	10364.00	9450.00	18076.00	48902.00

Figure 11 : *Résultat d'une requête d'extraction*

L’utilisateur peut modifier sa requête et la re-exécuter, consulter ou modifier ses agrégats. Enfin, il peut récupérer les données extraites de la base au format ascii ou csv et enregistrer le fichier ainsi téléchargé dans un répertoire personnel.

2.2.h. Création d’un agrégat
Selon les études menées, l’utilisateur peut avoir besoin de constituer différents agrégats. Par exemple, s’il souhaite comparer l'Union européenne « à 25 » avec les Etats-Unis, il lui faut ajouter un agrégat composé de l’UE « à 15 » et des 10 nouveaux membres. Lorsqu'un agrégat est ajouté son nom s'affiche dans la fenêtre d'extraction des données (**Figure 12**)

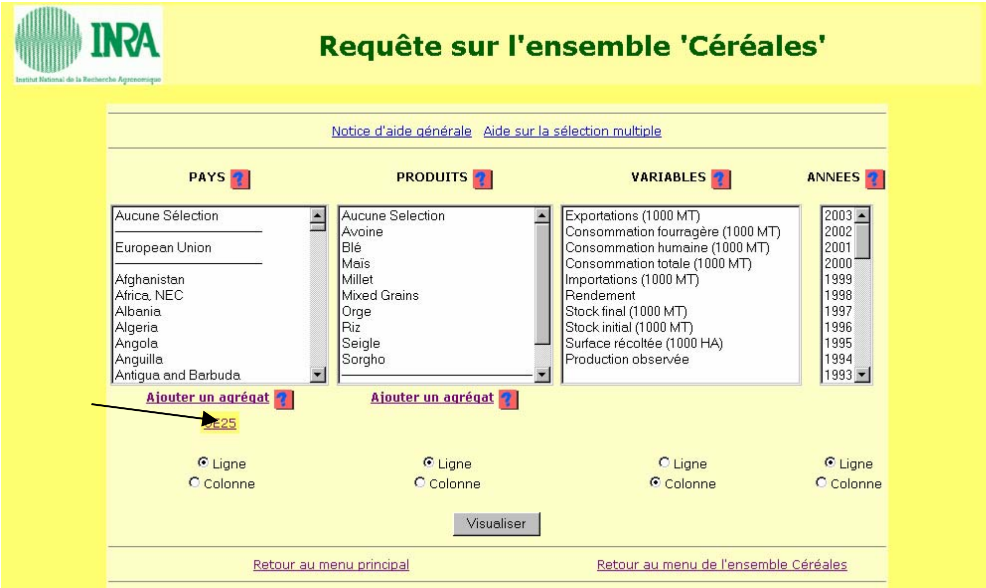


Figure 12 : *Interface d’extraction avec affichage d'un agrégat ajouté*

Le nom de l'agrégat est cliquable et permet d'ouvrir une nouvelle fenêtre permettant à l'utilisateur de le modifier, de le supprimer ou de l'enregistrer.

2.2.i. Fin de la session
A la fin de la session un script Perl efface tous les fichiers temporaires qui ont été nécessaires

3. Conclusion et perspectives

Le système d’information mis à disposition des utilisateurs permet donc de :
- consulter une base de données cohérente mise à jour annuellement,
- retracer l’évolution d’une série,
- disposer d’études statistiques sur les données,
- extraire des données afin de les utiliser pour des travaux de recherche.

Plusieurs élargissements du système d’information sont envisagés :
Tout d’abord, les données actuellement disponibles concernent les céréales et les oléagineux à l’échelle mondiale et la source de données utilisée (PSD) ne contient pas le détail des pays membres de l'Union européenne mais seulement une information globale au niveau des quinze pays constituant l'UE avant l'entrée des dix derniers membres (Pologne, République

Tchèque, Hongrie, Slovaquie, Lituanie, Lettonie, Estonie, Slovénie, Chypre, Malte). Aussi, nous utilisons les données fournies par Eurostat (base Newcronos) pour le détail des pays de l'Europe. Nous avons déjà créé la base de données « Europe » selon le même modèle décrit dans ce document. L'intégration de cette base dans notre système d'information est en cours et demande des contrôles de cohérence supplémentaires, notamment la vérification de l'égalité entre la somme des données pour les pays de la base « Europe » et l'agrégat « Europe » stocké dans la base actuelle.

Par ailleurs, de nouveaux besoins en terme de données sont apparus. Ainsi d'autres produits devront être intégrés à la base de données (animaux, produits laitiers, etc.). D'autre part, chaque variable renseignée pourrait être complétée par les prévisions générées à partir des modèles développés au sein de l'unité (horizon de 10 ans).

Enfin, nous envisageons la mise à disposition de l'outil à des personnes externes à l'institut. Actuellement le système mis en place n'est accessible qu'aux personnes de l'institut (accès via l'intranet et l'utilisation de SAS nécessite l'achat d'une licence). Or nous travaillons de plus en plus avec des partenaires externes, notamment dans le cadre de contrats européens. Aussi, il est envisagé de faire migrer ce système d'information vers une architecture permettant à ces partenaires d'y accéder.

Bibliographie

Diodat C., 2004, Mise en place d'un système d'information, Rapport de stage.

Silviu Herchi, 2001, Réalisation d'applications basées sur les technologies du Web, Rapport de stage.

Schwartz R.L., Christiansen T., 1998, Introduction à Perl, 2ème édition, O'Reilly, 303p.

Chaléat P., Charnay D., 1999, Programmation HTML et Javascript, 3ème édition, Eyrolles, 450p.

SAS Institute Inc aux USA : http://www.sas.com/index.html

SAS Institute France : http://www.sas.com/offices/europe/france/index.html

SAS à l'INRA : http://merlin.lusignan.inra.fr/webpsi/website/pourmoi/offreslogiciels/distribution/sas

Tukey J., 1977, Exploratory data analysis, Addison-Wesley.

USDA, Foreign Agricultural Service, Production, Supply and Distribution on line, http://www.fas.usda.gov/psd/

EUROPA, Eurostat, Agriculture and Fishery, http://epp.eurostat.cec.eu.int/portal/page?_pageid=0,1136206,0_45570467&_dad=portal&_schema=PORTAL