G. STEIER

INRA Centre de Traitement de l'Information Génétique 78352 Jouy-en-Josas Cedex

La gestion des populations

La circulation de l'information génétique et sa structuration sous forme d'une base de données

Préambule. *Le Département de Génétique Animale a été précurseur de l'informatique à l'INRA sous l'impulsion de J. Poly, B. Vissac et M. Poutous dans les années 50 ; à cette époque ils réalisaient des calculs sur IBM-704 et 705 et faisaient gérer des fichiers importants sur cartes perforées.*
En octobre 1962, les premiers équipements de calculs internes sont installés sous forme d'un IBM 1620 qui traite non seulement les données zootechniques mais également des données administratives.
En janvier 1968, grâce aux crédits de la nouvelle loi sur l'élevage, est installé l'IBM 360-50, ordinateur de haut de gamme pour l'époque car répondant par exemple aux importants besoins informatiques d'organismes financiers comme la Caisse des Dépôts et Consignations qui en fut dotée au même moment.
Associés à du personnel spécialisé regroupé dans une Unité du Département, dénommée Centre de Traitement de l'information génétique (CTIG), ces moyens modernes de calcul permettaient au Département de remplir ses missions au bénéfice des acteurs de l'Amélioration Génétique. Les biométriciens de l'INRA auront également accès à ce matériel avant d'être doté de moyens propres en 1979.
Le Département de génétique animale a donc joué un rôle moteur dans le développement des moyens informatiques de l'Institut. Cela lui permet aujourd'hui, en partenariat avec les Instituts techniques et le Ministère de l'Agriculture, de disposer des outils nécessaires pour améliorer les méthodes de gestion et de sélection des populations d'animaux domestiques.

Toute organisation durable s'appuie sur des règles de fonctionnement. L'informatique appliquée au Centre de Traitement de l'Information Génétique (CTIG) a les siennes, qui, émergeant des balbutiements fragmentaires des années 60, ont pris corps dans l'articulation des différents traitements au cours des quinze dernières années, en dépit de l'évolution considérable des puissances de calcul et des supports de stockage en accès direct.

Le traitement des informations génétiques en France a lieu à plusieurs niveaux ayant chacun ses missions particulières et non concurrentielles grâce au schéma adopté.

Mais, avant de les décrire, il apparaît opportun de citer les objectifs qui ont guidé ceux qui ont eu la responsabilité de faire évoluer le CTIG.

Objectifs d'ordre organisationnel

Une base de données est une supermémoire au service des décideurs et des acteurs ; elle doit être capable de les renseigner sur la situation actuelle et sur l'historique des faits qui l'ont concernée et qu'elle a enregistrés.

Les organisations qui fonctionnent de plus en plus sur la base des connaissances indirectes de la réalité doivent s'assurer de la pertinence, de la cohérence et de la fiabilité des représentations de la réalité qu'elles manipulent à partir des données stockées dans la base.

Objectifs d'ordre technologique

Un Système de Gestion de Base de Données (SGBD) doit :

- centraliser la gestion des données pour permettre :

  . la suppression des redondances,

  . l'unicité de la saisie,

  . de centraliser les contrôles,

  . le partage des données en mettant en oeuvre des mécanismes de concurrence qui permettent à plusieurs utilisateurs de manipuler simultanément les mêmes collections de données.

- garantir l'intégrité des données et leur sécurité contre les erreurs de manipulations, les pannes, la malveillance, et assurer la restauration dans un état valide,

- permettre la confidentialité par la privatisation des accès et la manipulation de certaines données,

- assurer l'indépendance données-traitement en proposant des mécanismes qui permettent à différents

programmes d'applications d'avoir différentes vues d'une même donnée.

Objectifs d'ordre économique

L'amélioration du ratio "Etendue des services rendus par les données/volume des dépenses de mise en oeuvre" doit être une préoccupation permanente à travers :

- la durée de vie des applications qui découle de l'indépendance des modes de stockage et des structures de traitement,

- la multiplication des usages et des utilisateurs via les traitements par lots ou l'accessibilité par des processus de télécommunications courants (minitel, micros.),

- la disponibilité des données,

- la réduction des coûts de saisie et le renforcement des contrôles d'intégrité et de compatibilité,

- la mise en évidence qu'il existe des possibilités de réutilisation transversale de procédures qui peuvent être communes à des "types de contrôle " à travers les différentes espèces,

- un effort de sémantique qui facilite la maintenance en dépit du "turn-over" des hommes ou permet le redéploiement de fonctions si de nouvelles techniques portent à la redistribution des responsabilités.

Evolution du système national de contrôle de performances

La modélisation du système a été réalisée, au début des années 60, au sein du Département de Génétique Animale, où tout se faisait, de la méthode de collecte des informations à la sortie des résultats, en passant par la saisie des données et l'émission des listes-échange avec les éleveurs. A partir de 1968, le succès de l'action attirant de nouveaux adhérents et impliquant des engagements financiers qui ne pouvaient qu'être partagés avec les professionnels, il y eut répartition des tâches avec la mise en place des Centres Régionaux Informatiques (CRI) afin que la collecte des données et la diffusion des résultats se fassent au plus près possible de l'éleveur grâce aux progrès de l'informatique et à la diminution des coûts.

Au maître d'oeuvre unique succédait une collégialité de partenaires et d'intérêts qu'il fallait coordonner, d'où la création du Groupe Elevage-Informatique (GELI) qui validait les choix informatiques en fonction de la politique définie par la Commission Nationale d'Amélioration Génétique (CNAG) et des ingénieurs-pivot. Rattachés en majorité aux Instituts techniques ayant en charge l'espèce et la production (lait ou viande) qui leur étaient confiées, ces derniers devaient jouer un rôle très important de suivi des actions informatiques. Ainsi ils devaient s'assurer que les fonds investis dans les systèmes de traitement respectaient en qualité et délais les normes édictées par les cahiers des charges acceptés en commun. Mais ils devaient aussi jouer un rôle de coordination entre les acteurs professionnels et la Recherche pour faire évoluer les systèmes informatiques.

Les évolutions successives ont eu pour préoccupation de faciliter le recueil de l'information, d'en accroître la fiabilité et la valorisation pour l'ensemble des acteurs (de l'éleveur aux Unités de Sélection).

Au cours de ces évolutions la circulation des informations génétiques s'est toujours appuyée sur 3 niveaux (figure 1) : départemental (Etablissement Départemental de l'Elevage), régional (12 CRI) et national (CTIG).

Rôle du CTIG au niveau national

De par sa position, le CTIG est une plaque tournante qui reçoit des informations brutes des CRI et renvoie des informations élaborées et des index vers les CRI et des partenaires nationaux et raciaux : Instituts techniques, Unions pour la Promotion de Races (UPRA).

En 1990 le CTIG a reçu 45 millions de Vecteur Standard Informatique (VSI) et en a réexpédié à peu près autant tout au long des 52 semaines de traitement.

Pour clarifier ses actions, le CTIG a dû définir des standards pour :

- appréhender les données brutes , ou émettre des résultats sur supports magnétiques afin que l'impression décentralisée évite les coûts de transport,

- façonner en standard les données brutes dans les phases primaires, ou les données élaborées dans les phases ultimes,

- traiter de façon commune les variables expurgées avant qu'elles soient intégrées aux bases de données, et rendre "publiques" les descriptions d'ensembles d'informations sous forme de répertoires de données régulièrement tenus à jour et partiellement ou totalement édités à l'intention des acteurs du système : CRI, ingénieurs-pivot, chercheurs, informaticiens du CTIG (suivi de projet).

Ainsi, ZOOREP11, qui couvre le secteur "bovins-lait", est un document de 145 pages de 100 lignes potentielles.

Standard VSI

Entre le CTIG et ses partenaires le support initial des informations a été la carte perforée et, si physiquement elle disparaît progressivement au début des années 70, son format unitaire de 80 colonnes a été utilisé jusqu'à la proposition du VSI qui répondait à un besoin de normalisation pour :

a/ formaliser les apports en enregistrements fixes de 128 caractères pour les données et de 164 pour les impressions.

b/ sécuriser les apports en nombre par l'encadrement des données grâce à :

- l'enregistrement "début" qui identifie l'émetteur et la date d'envoi,

- des enregistrements "détail" qui supportent des codes dont la description est largement diffusée,

- l'enregistrement "fin" qui décompte les enregistrements "détail" et assure que le nombre d'apports est incontesté par l'émetteur et le récepteur , toute différence donnant lieu à un rejet global des apports.

c/ décrire les données globalement, dans un "répertoire de données" mis à jour par le CTIG pour chaque

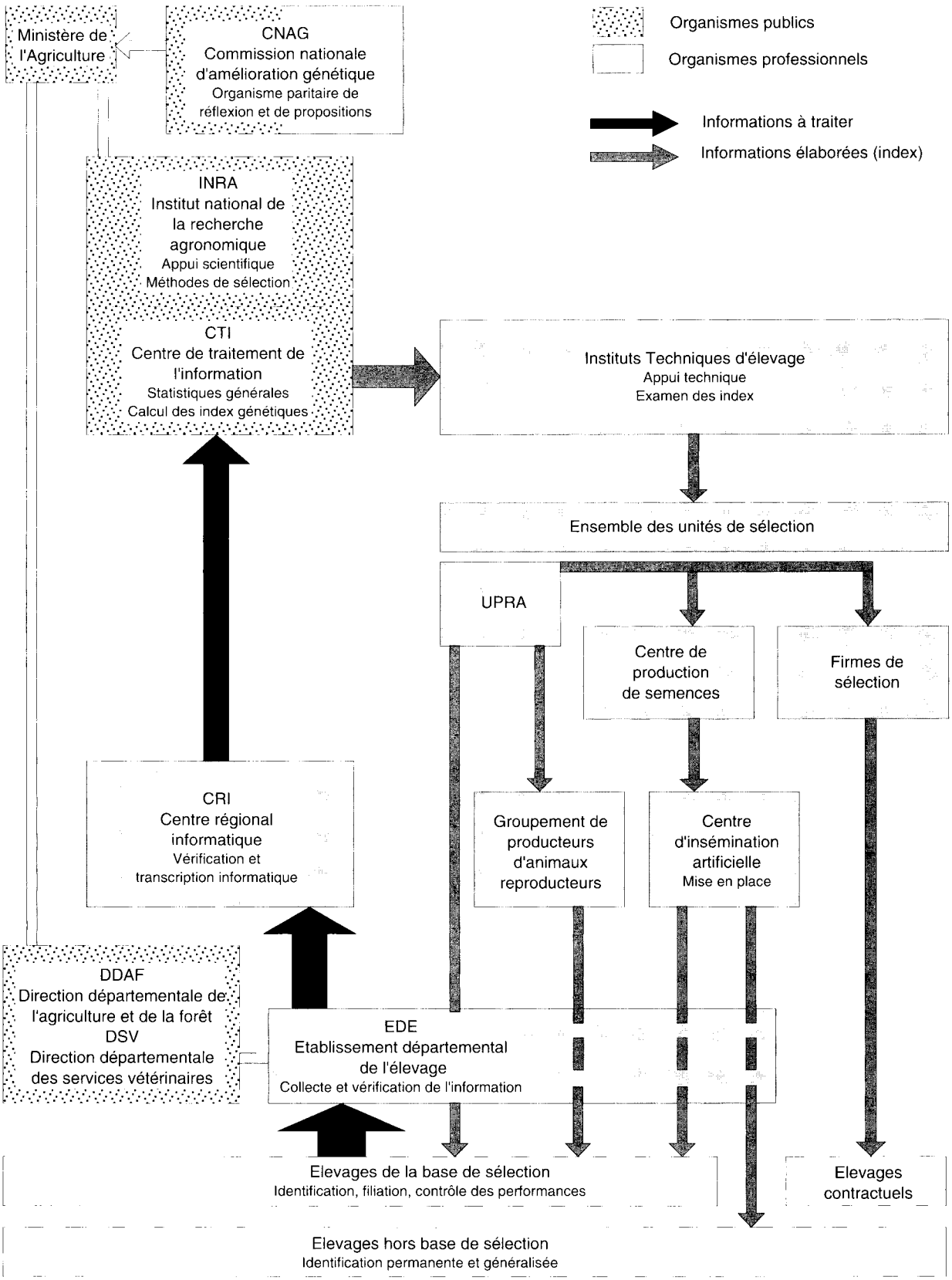

Figure 1. *Relations entre les organismes s'intéressant à l'amélioration génétique des animaux (source : Bulletin de l'élevage français, 1987, numéro spécial).*

espèce-type de contrôle à l'occasion d'une création ou d'un aménagement de traitement soutenu par un cahier des charges, et individuellement, pour décrire les contraintes d'intégrité qui s'y rattachent :

- plages de valeurs autorisées,
- numériques ou alphanumériques,
- présence obligatoire ou facultative.

Standard de mise en forme

A partir des standards de format et de description, on pouvait mettre en place des standards de traitement itératif en tableaux qui permettent facilement l'ajout de variables dans un code ou de modifier une plage de valeurs et d'obtenir en sortie un VSI Interne ou VSII.

Entre les contraintes externes imposées par les utilisateurs et la nécessité de réduire les coûts de maintenance liés à ces évolutions externes à l'informatique, le VSII est une interface qui permet, par exemple, d'assurer une ergonomie correcte même dans la présentation des rejets aux gestionnaires des CRI ou des syndicats de contrôle.

Le VSII contient :

- les indicatifs normalisés pour un tri logique, dissociant ainsi le code externe de sa séquence naturelle dans un ensemble d'apports,

- une zone de 64 bits qui indique potentiellement 63 types d'erreurs de validité dans un apport,

- la codification de l'émetteur utilisé en cas de rejet,

- la date d'émission pour traiter dans l'ordre chronologique d'apport,

- l'image de l'apport à réémettre en cas de rejet,

- les données en mode interne conforme à leur format de stockage dans la base de données, ou codées spécifiquement pour signaler l'absence.

Standard de mise à jour de la base de données

A partir de ce niveau, les traitements sont plus hétérogènes car ils doivent s'adapter à la spécificité de l'application dans son cadre zootechnique.

Le standard prend une autre forme grâce à l'emploi d'un générateur de programme en langage PL/I : XPL.

Ce générateur mis au point par mes soins en 1972 découle des méthodes Warnier-Corig et du principe de la programmation "topdown" qui permet une plus grande lisibilité des programmes, associée à la modularité qui, sous réserve d'une attention particulière aux interfaces, permet l'introduction de nouveaux codes VSII.

De plus, les descriptions de bases de données installées dans des bibliothèques centralisées permettent des mises à niveaux par simple compilation des programmes et sans remise en cause de l'existant.

Pour la mise en oeuvre sous CMS par une procédure interactive, le programmeur décrit le contexte de son programme :

- nombre de niveaux de rupture (ex : syndicat, éleveur, animal)

- le nombre de fichiers en entrée ou écriture et pour chacun d'eux le format (fixe, variable...), la méthode d'accès (séquentielle, directe,..)

A ces choix explicites, la procédure ajoute des fonctions implicites qui :

- permettent des contrôles de séquence de traitement,

- encadrent les accès aux modules,

- introduisent des procédures de reprise de traitement,

- documentent les sorties sur les "post-list" d'exploitation avec la date de dernière maintenance du programme exploité, les date et heure de début et fin de traitement, les comptages sur les entrées-sorties et les ensembles traités.

Le tout conduit à la mise en place de lignes d'inclusion par "%INCLUDE ..." choisis ou imposés, qui lors de la compilation provoqueront l'appel de modules stockées dans une bibliothèque et mis à niveau en fonction des systèmes utilisés dans le temps : OS/VS1,MVS/SP,MVS/XA MVS/ESA.

XPL a donc introduit une sémantique qui prépare à la lisibilité du programme quel qu'en soit l'auteur et par conséquence facilite la maintenance, qu'elle provienne des évolutions de système ou d'aménagements suscités par les utilisateurs.

Les fichiers ont des formats standards sous forme de "segments" à longueur variable de 1536 octets au maximum ; le dossier d'un animal ou d'un éleveur étant composé de n segments à présence variable ; seul un premier dit "témoin" ou "S000" étant toujours présent pour indiquer quels segments sont présents à sa suite.

Ces bases ont également des descriptions centralisées dans des bibliothèques auxquelles on fait appel par des "%INCLUDE ...".

Cet agencement permet l'ajout de données et de traitements supplémentaires sans remettre en cause les données déjà stockées et nécessite une maintenance allégée. Par ailleurs il s'adaptait sans problème à une organisation indexée de type VSAM ; il peut servir de support à une évolution facilitée vers les organisations de bases relationnelles de type DB2.

Des standards d'émission des rejets s'appuyant sur des tables externes, des standards de communications permettant d'extraire des données, sont autant d'éléments qui sont similaires à travers les différentes chaînes.

Standard de contrôle de flux

Chaque mise à jour de base de données donne lieu à l'émission de "tableaux de sécurité" qui parviennent à l'ingénieur-pivot.

Ces tableaux permettent une observation des délais de stockage dans la base depuis l'événement zootechnique (naissance, pesée, etc) ; ceci est très important quant à la prise en compte dans les index ; ils donnent un aperçu de la qualité de l'information par le % de rejets (en moyenne 1 % pour les bovins-lait) et suscitent une enquête auprès de l'émetteur si le seuil d'alerte est atteint.

Standard de fourniture d'informations aux chercheurs

Les généticiens qui ont en charge la fourniture d'index pour la sélection des reproducteurs sont les destinataires privilégiés des informations collectées, passées au crible de nombreux tests de vraisemblance et de compatibilité, et dont ils reçoivent des extractions selon des rythmes et des modalités qu'ils définissent aux informaticiens du CTIG.

Leurs calculs achevés, ils retournent au CTIG des fichiers qui servent à créer des VSI destinés aux intervenants économiques.

Le CTIG agit donc "en frontal" de l'action scientifique, évitant ainsi les doublons dans l'action informatique.

Effet des standards sur les conditions d'exploitation

Il découle de la mise en oeuvre de standards à tous les niveaux que les procédures de traitement mises en oeuvre périodiquement par les techniciens de l'exploitation (pupitreurs et préparateurs) sont relativement homogènes et d'une complexité rapidement assimilable.

Conclusion

En informatique, la technique évolue, la puissance des ordinateurs ne cesse de croître et le prix de la puissance en Mips (millions d'instructions par seconde) ou de l'octet stocké ne cesse de baisser.

Les délais de réalisation sont toujours trop longs ; le temps écoulé entre la formulation du besoin informatique et sa mise en place se chiffre parfois en années ; dans ces conditions, les résultats conviennent plus ou moins aux utilisateurs dont les besoins ont changé entre-temps. C'est pourquoi le temps maximum entre la formulation d'un projet et sa mise en oeuvre ne doit pas dépasser 18 mois. A défaut, il faut découper en sous-projets entrepris l'un après l'autre.

Les programmes ont un cycle de vie avec des crises de jeunesse et des défaillances séniles. Il faut les entretenir, les maintenir, les adapter, les corriger des erreurs de programmation. Cette maintenance est coûteuse en temps et rend les informaticiens moins disponibles au préjudice du temps consacré au développement. Il devient ainsi difficile de répondre à toutes les exigences des utilisateurs.

Malgré ces contraintes il faut noter qu'un système cohérent d'informations est mis au service des organisations d'élevage et des éleveurs adhérant au contrôle de performances. Ce système n'est pas figé, mais les évolutions qu'il a connues ont su préserver sa cohérence. Des applications mises sous forme de base de données relationnelles comme la base de données "taureaux" matérialisent bien ces évolutions qu'il faut poursuivre avec les différents partenaires. Il n'y a pas de raison pour que dans le futur les différents maîtres d'oeuvre informatiques concernés par les données de l'élevage ne continuent pas à oeuvrer dans ce sens.