

Traitement des données en Psychologie

UV PSY38X2 - Licence de Psychologie

Informatique

F.G. Carpentier - 2001-2002

1 Généralités sur les bases de données

1.1 Introduction

1.1.1 Fichiers manuels et bases de données

Parmi les activités humaines de traitement de l'information, l'une des plus anciennes, mais aussi l'une des plus fastidieuses, est la tenue de *fichiers* (constitution, mise à jour, consultation, etc.). Le mot *fichier* est ici pris dans un sens "extra-informatique". Il renvoie à l'objet *boîte de fiches*. On utilise souvent le terme de *fichier manuel* pour désigner un tel objet.

Or, l'un des objectifs fondamentaux de l'informatique est d'automatiser de tels traitements répétitifs, fastidieux, non intelligents. Dès les débuts de l'informatique, on s'est donc intéressé à la *gestion de fichiers*.

Cependant, on s'est assez vite rendu compte que :

- la mise sous forme électronique amplifie les problèmes liés à la gestion de tels fichiers : incohérences, difficultés de mise à jour, difficultés d'accès aux données, etc ;
- les possibilités de traitement offertes par l'ordinateur permettent d'envisager une gestion plus élaborée que la simple transposition des tâches effectuées manuellement. Mais la mise en œuvre de telles possibilités exige une structuration des données sous une forme élaborée, celle d'une *base de données*.

C'est ainsi que les informaticiens ont cherché à développer des méthodes visant à identifier et à mémoriser de façon efficace les informations qui circulent entre les différents acteurs d'un domaine ou d'une organisation : *méthodes d'étude d'un système d'information*, et des modèles d'organisation des données : *modèles de données*.

Dans ce cadre, deux types d'approche peuvent être envisagés, correspondant aux deux types de problèmes énoncés ci-dessous :

- On veut étudier une organisation, et les informations qui circulent dans cette organisation. C'est l'aspect "système d'information".
- Dans le cadre d'un travail d'étude, nous avons recueilli (ou allons recueillir) des données formant des lots homogènes (données répétitives). Nous nous posons le problème de mémoriser ces données de façon efficace. C'est ce dernier aspect qui sera traité dans ce chapitre et le suivant.

1.1.2 Exemples

Exemple 1. Dans le cadre d'un travail de recherche, vous avez étudié un corpus de 200 publicités télévisées concernant un thème de consommation donné. Cet exemple sera développé ultérieurement.

Exemple 2. Au sein d'une équipe de psychologues, vous vous intéressez aux liens familiaux existant dans un groupe humain donné.

Exemple 3. Vous travaillez dans une cinémathèque qui souhaite mettre à disposition du public une base documentaire relative à un certain nombre de films.

Dans chacun des cas cités, nous poursuivons des buts identiques:

- Nous voulons mémoriser, de façon efficace, les données recueillies.
- Nous voulons que cette mémorisation respecte la *sémantique* que nous accordons à ces données.
- Nous voulons disposer de possibilités d'interrogation permettant d'accéder à tout ou partie de ces données.

1.1.3 Remarques

Ces problèmes de gestion de lots de données homogènes rejoignent ceux que l'on rencontre en statistiques, traitement d'enquêtes, gestion de systèmes documentaires, etc. Cependant, la problématique posée dans l'approche "bases de données" est spécifique à cette approche:

- En statistiques descriptives, les traitements auront pour but d'obtenir une vue résumée, synthétique de ces données. En particulier, on rend "anonymes" les sujets observés.
- Dans un système documentaire, c'est le temps d'accès aux données qui est un critère essentiel. On s'efforcera de mémoriser les données de façon que l'utilisateur puisse obtenir des réponses à ses requêtes dans des délais brefs. Ce but est en général atteint au prix d'une certaine *redondance* des données stockées (tables inverses...)
- Dans une base de données, on souhaite structurer les données de façon à conserver un accès aux données conforme à leur sémantique et éviter toute redondance, source d'incohérences et de difficultés dans les mises à jour.

1.2 Définitions

1.2.1 Base de données

Une **base de données** est un ensemble de données structurées, fiables, cohérentes, partagées par des utilisateurs ayant des besoins différents.

La définition donnée ici est très générale et s'applique aussi bien à de "petites" bases de données telles que celles des exemples ci-dessus et à de très grosses bases de données (par

exemple dans une grande entreprise). En particulier, une base de données est généralement destinée à être exploitée par plusieurs utilisateurs; elle doit alors être accessible à distance, grâce à un réseau de transmission de données.

Les utilisateurs d'une base de données ont des besoins et des compétences variées. Il est indispensable d'instaurer une certaine hiérarchie: la base doit posséder un *administrateur* et des utilisateurs, avec des profils variés.

Par exemple, vos camarades de licence LEA travaillent sur un lexique de néologismes en langue espagnole.

- Plusieurs étudiants peuvent travailler simultanément sur la base de données (accès concurrents) qui est enregistrée sur le serveur des salles de travaux dirigés.
- Chaque étudiant est identifié par un nom et un mot de passe.
- Chaque étudiant peut modifier les fiches qu'il a lui-même saisies, mais peut seulement consulter les fiches saisies par ses collègues.
- Il existe également un "propriétaire" de la base de données: les enseignants responsables de cet enseignement peuvent s'identifier avec un couple nom/mot de passe particulier et peuvent alors modifier toutes les données, modifier la structure de la base de données, etc.

1.2.2 Système de gestion de bases de données

Un **système de gestion de bases de données** (SGBD) est un logiciel permettant de gérer et de manipuler une base de données.

Les termes "gérer" et "manipuler" cités dans la définition précédente recouvrent à la fois les activités de création et de maintenance de la base (créer la structure, définir les droits d'accès,...) et les activités de manipulation des données elles-mêmes (lecture, insertion, modification, destruction de données).

1.3 Mise en œuvre d'un SGBD

Il s'agit, étant donné un lot de données issues du monde réel, d'organiser et de structurer ces données dans une base de données afin que l'utilisateur final puisse se servir d'un SGBD pour accéder aux données qui l'intéressent.

Cette mise en œuvre présente-t-elle des difficultés plus grandes que celle d'autres logiciels de bureautique (traitement de textes, tableur, ...)? En fait, lorsque nous travaillons avec un logiciel tel qu'un traitement de textes ou un tableur, nous nous appuyons largement sur nos connaissances antérieures et extra-informatiques: nous avons une connaissance explicite ou implicite des règles de présentations d'un document imprimé, de la structure d'un tableau de données, etc.

Certes, nous pouvons nous représenter un SGBD comme une métaphore d'une boîte de fiches. Mais cette métaphore est beaucoup trop réductrice. En fait, un SGBD tel qu'Access ou Oracle s'appuie sur un modèle théorique bien défini, *l'algèbre relationnelle*. Il paraît illusoire de penser apprendre à manier un tel logiciel sans avoir:

- acquis quelques rudiments d'algèbre relationnelle;
- appris à utiliser le modèle relationnel pour structurer des données issues du monde réel.

Jointure entre deux tables selon un attribut commun

La jointure est l'opération fondamentale grâce à laquelle le modèle relationnel peut représenter les associations entre les objets du monde réel. Ainsi, considérons l'exemple traité en travaux dirigés dans lequel on dispose de deux tables : *Intervenants* et *Interventions*, chaque intervention ayant à sa tête un responsable choisi parmi les intervenants. On peut représenter l'information :

Tel intervenant est responsable de telle intervention

en dotant la table *Interventions* d'un attribut *Code-Resp* et en effectuant une jointure entre les deux tables.

Code-Res	Nom	Code-Intv	Objectif	Code-Res
1	Dupond	A	Irrigation	1
2	Durand	B	Ecole	1
3	Martin	C	Vaccination	2

La jointure entre les deux tables produit alors la table suivante :

Code-Res	Nom	Code-Intv	Objectif
1	Dupond	A	Irrigation
1	Dupond	B	Ecole
2	Durand	C	Vaccination

D'autres opérations sont définies par le modèle relationnel, telles que la réunion, l'intersection, le produit cartésien, la jointure selon une condition, etc.

1.4.3 Pourquoi développer des modèles conceptuels des données ?

Il ne suffit malheureusement pas de structurer des données issues du monde réel sous forme de tables pour garantir une structure de base de données correcte. Dans la plupart des situations réelles, il faut répartir les données entre plusieurs tables, et le choix des tables à créer doit nécessairement s'appuyer sur la *sémantique* des données manipulées. Il faut donc sortir du cadre purement formel fourni par l'algèbre relationnelle et construire un modèle de nos données respectant leur signification, un *modèle conceptuel des données*.

Parmi les méthodes de modélisation qui ont été proposées, l'une des plus courantes et des plus simples à mettre en œuvre est le modèle *entités-associations* dont nous donnerons une idée dans le chapitre suivant.