SLAM3-2 Modélisation

Conception et adaptation d’une base de données

Organiser des données

Organiser des données : liste

Organiser des données, où est le problème, direz-vous ? Y a-t-il vraiment besoin de faire des études pour cela ? Il suffit d'être un peu soigneux, et il n'y a aucune raison que ça se passe mal.

Eh bien si, en fait. Les données, quand elles se présentent en (très) grandes quantités, posent des problèmes auxquels on ne s'attend pas. C'est bien pour cela qu'on a mis au point un certain nombre de règles et de méthodes qui, si elles ne peuvent dispenser personne de réfléchir, aident néanmoins à réfléchir en évitant les catastrophes les plus courantes.

Voyons les problèmes les plus évidents qui se posent lorsqu'on veut organiser des informations. Pour cela, imaginons que nous voulions informatiser les rayons d'une discothèque… – enfin, disons les dix premiers CD du rayonnage, parce que sinon, ça va être un peu longuet.

Faisons simple pour commencer : nous ne noterons, pour chaque CD, que le titre, l'année, le nom de l'artiste et le genre musical. Cela nous donne la liste suivante :

Nursery Cryme, Genesis, 1972, rock progressif
Foxtrot, Genesis, 1972, rock progressif
Selling England by the Pound, Genesis, 1973, rock progressif
Symphonie n°2, Sibelius, 1985, classique
Symphonie n°7, Sibelius, 1987, classique
Concerto pour violon, Mendelssohn, 1992, classique
Crime passionnel, Guidoni, 1982, chanson française
5th Gear, Brad Paisley, 2007, country
Thick as a Brick, Jethro Tull, 1973, rock progressif
Purpendicular, Deep Purple, 1996, rock

Organiser des données : tableau

Il va de soi que pour représenter de telles informations, une liste c'est bien, mais qu'un tableau, ce serait beaucoup mieux. Les conventions étant en accord avec ce que nous suggère l'intuition, on mettra en ligne, les uns en-dessous des autres, les différents disques (autrement dit, en quelque sorte, les « individus » de notre base). Et on portera en colonne les différents renseignements dont on dispose pour chacun de ces individus. Naturellement, il est préférable –- et, en réalité, obligatoire - de nommer ces colonnes : Ma discothèque prendrait ainsi la forme du tableau suivant :

TITRE	ARTISTE	ANNEE	GENRE
Nursery Cryme	Genesis	1972	rock progressif
Foxtrot	Genesis	1972	rock progresif
Selling England by the Pound	Genesis	1973	rock progressif
Symphonie n°2	Jean Sibelius	1985	classique

Symphonie n°7	Jean Sibelius	1987	classique
Concerto pour violon	Felix Mendelssohn	1992	Classique
Crime passionnel	Jean Guidoni	1982	chanson française
5th Gear	Brad Paisley	2007	country
Thick as a Brick	Jethro Tull	1973	rock progressif

Or, ce petit tableau, à lui seul, fait apparaître au moins deux problèmes majeurs.

le premier, c'est que des informations identiques s'y répètent : en l'occurrence, celles concernant les artistes, et encore plus, le genre. En termes savants, on parle de redondance des informations. Imaginons que ma discothèque comporte plusieurs centaines de CD, il y a fort à parier que la mention « classique » ou « rock progressif » va se retrouver répliquée des dizaines de fois. Or, tout cela, il va bien falloir le stocker quelque part sous forme de bits et d'octets, et des informations inutilement répétées, ce sont des octets inutilement occupés…

le second problème, lié au premier, est pour sa part rédhibitoire. C'est que vu la manière dont j'ai bâti mes informations, je ne suis pas à l'abri d'une erreur, ou même d'un simple manque d'homogénéité, dans la saisie. Par exemple, lorsque j'ai tapé le genre du CD Foxtrot, j'ai oublié les deux « s » de « progressif », ce qui est une faute de frappe assez classique. De même, le genre « classique » est orthographié tantôt avec une majuscule, tantôt sans majuscule. Tout cela est fort préjudiciable pour la suite des événements. Si je fais une recherche, par exemple, sur le genre « rock progressif » ou « classique », il manquera des CDs qui auraient dû y figurer. Cette organisation laisse donc la possibilité qu'existe une hétérogénéité des données, un des pires cauchemars de l'informaticien (qui pourtant n'en manque pas) .

Organiser des données : système de bases de données relationnelles

Ces deux problèmes possèdent une solution commune, très simple mais extrêmement efficace, qui consiste à recenser séparément les CD et les genres :

TITRE	ARTISTE	ANNEE
Nursery Cryme	Genesis	1972
Foxtrot	Genesis	1972
Selling England by the Pound	Genesis	1973
Symphonie n°2	Jean Sibelius	1985
Symphonie n°7	Jean Sibelius	1987
Concerto pour violon	Felix Mendelssohn	1992
Crime passionnel	Jean Guidoni	1982

GENRE
chanson française

classique
country
rock progressif

Il ne reste plus à présent qu'à préciser à quel genre appartient chaque CD. Pour ce faire, on identifie chaque genre par un code unique, et on reporter ce code dans le tableau des CDs.

CODE	GENRE
1	chanson française
2	classique
3	country
4	rock progressif

TITRE	ARTISTE	ANNEE	CODE
Nursery Cryme	Genesis	1972	4
Foxtrot	Genesis	1972	4
Selling England by the Pound	Genesis	1973	4
Symphonie n°2	Jean Sibelius	1985	2
Symphonie n°7	Jean Sibelius	1987	2
Concerto pour violon	Felix Mendelssohn	1992	2
Crime passionnel	Jean Guidoni	1982	1
5th Gear	Brad Paisley	2007	3
Thick as a Brick	Jethro Tull	1973	4

Nous venons de faire d'une pierre deux coups :

on a économisé de la place en mémoire, car stocker un code mobilise beaucoup moins d'octets que stocker un intitulé

surtout, on a pris une garantie contre les informations hétérogènes : le code correspond toujours au même intitulé, et on ne peut plus guère imaginer se retrouver avec un même genre musical orthographié différemment.

Ce que nous venons de faire, c'est – en tout petit, petit, petit – le fond de la question en matière de modélisation de l'information : nous venons de créer une relation entre deux tables ; voilà pourquoi on parle de Systèmes de Bases de Données Relationnelles. Tout ce qu'on va voir ensuite, ce ne sont pour ainsi dire que des complications et des raffinements à partir de cette base simple.

Cela dit, jusqu'à maintenant, nous avons procédé pour ainsi dire uniquement par intuition – or, vous vous en doutez bien, il existe un certain nombre de formes pré-établies, de standards (tant de

méthode que de représentation) pour venir à bout des problèmes les plus complexes. Ce sont ces formes et ces standards que nous allons aborder à présent.

Remarque capitale : Il suffit d'observer la situation à laquelle nous sommes parvenus pour constater que nous n'avons parcouru que la moitié du chemin. Nous avons certes éliminé une source de redondances en créant la table Genres. Mais nous en avons laissé une deuxième : celle liée aux artistes. Il est donc essentiel de comprendre que ce qui est présenté ici n'est que le premier pas, et que notre modélisation est irrecevable en l'état.

Exercice: Trouvez une solution pour éliminer la seconde source de redondance (=duplication d'information).

Merise et les diagrammes « entités-associations »

Merise est le nom d'une méthode, ou d'un ensemble de méthodes, développée en France dans les années 1970, et qui a été très largement employée depuis lors. Depuis une quinzaine d'années, Merise laisse peu à peu place à UML (une autre norme).

Le cadre général

Merise constitue donc un ensemble très riche de méthodes et de représentations, dont nous ne verrons ici qu'une petite partie - mais la plus cruciale.

Toute base de données va donner lieu à une double représentation :

- le plan le plus abstrait (mais qui contient déjà toutes les informations indispensables pour la construction de la base de données) . Ce plan porte le nom de fort poétique de **Modèle Conceptuel de Données (MCD)**.
- un plan plus proche de ce que sera la base effective, telle qu'elle sera réalisée sur machine : le **Modèle Logique des Données (MLD)**.

Le point crucial à enregistrer dès maintenant, c'est que le MLD se déduit strictement du MCD d'après des règles formelles. Autrement dit, une fois le MCD réalisé, il n'y a plus besoin de réfléchir une seule seconde pour produire le MLD : tout se fait par automatismes. La meilleure preuve, c'est qu'il existe des logiciels qui se proposent de réaliser le MLD d'un clic de souris, d'après le MCD. En revanche, il n'existe rien de tel pour concevoir le MCD : le seul ingrédient qui entre dans sa composition est l'huile de neurones. N'oubliez pas de faire quelques provisions…

Le Modèle Conceptuel de Données, les entités et les relations

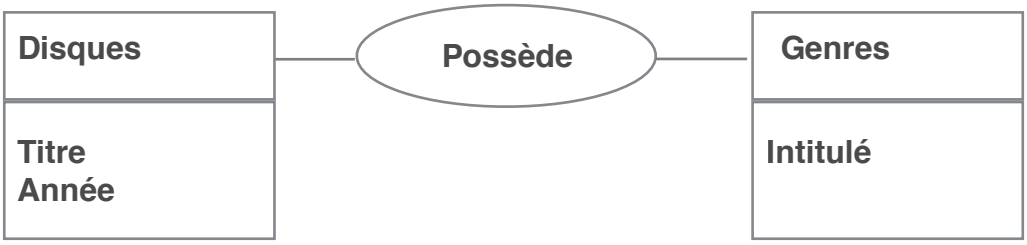
Les informations à traiter doivent être regroupées en ensembles cohérents, comme dans les tableaux que nous avons constitués il y a un instant. Dans les conventions de Merise, ces ensembles s'appellent des **entités**, et sont symbolisés par des **rectangles**. Chaque entité porte un **nom**, qui l'identifie de manière unique. Ce nom sera obligatoirement un substantif au pluriel : pour notre discothèque, on propose très logiquement « Disques » et « Genres ».

Les entités comprennent toujours un certain nombre d'éléments appelés **propriétés** (on parlera aussi d'**attributs**). Il s'agit des différentes rubriques qui devront être renseignées pour chaque individu.

Chaque entité, lorsqu'on passera au MLD (puis à la réalisation concrète de la base) donnera lieu à un tableau (on parle plus volontiers de **tables**). L'entité Disques du MCD produira donc une table Disques dans le MLD, et l'entité Genres, une table Genre. Les différentes propriétés de l'entité, qui sont donc écrites les unes sous les autres, deviendront les titres des colonnes de ces tables. Et dans ces colonnes, on fera figurer les différentes valeurs que prennent ces propriétés pour chacun des éléments de nos tables. Si vous trouvez cette explication un peu compliquée, pensez tout simplement à l'entité / table Disques : le titre, l'année et l'artiste sont disposés les uns sous les autres lorsqu'on parle de l'entité, et les uns à côté des autres (ce sont des en-têtes de colonne) lorsqu'on la représente comme une table.

Il ne reste plus à signifier que pour que chaque CD possède un genre (et pas n'importe lequel), mes deux entités doivent se trouver en **relation** l'une avec l'autre. Cette relation (on peut aussi parler **d'association**) sera symbolisée par un **ovale**, et sera nommée (par un verbe).

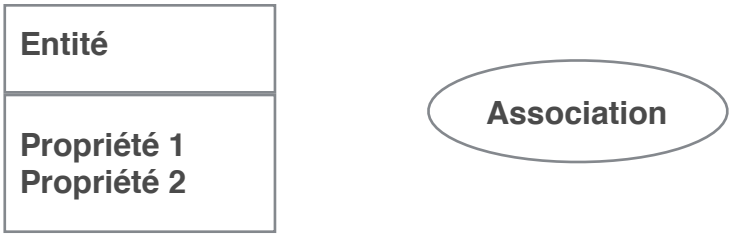
Voilà donc ce que cela donne :



Cette représentation ne se lit pas n'importe comment. Pour être certain de ne pas commettre de contresens, lorsqu'on traduit le schéma ci-dessus, il vaut mieux éviter de dire « les disques possèdent des genres », ou pire encore « la table disque possède certains genres » La bonne traduction, celle qui vous évitera au maximum de commettre des erreurs, consiste à dire que « Chaque élément de la table Disques possède un (ou plusieurs ?) genres ». En prenant l'affaire par l'autre bout, on peut tout aussi bien dire (même si c'est un peu laid à l'oreille) : « Chaque genre est possédé par un (ou plusieurs ?) disques » (on verra un peu plus loin comment en avoir le cœur net sur ces points d'interrogation).

Résumons-nous :

- les données doivent être systématiquement regroupées de manière à éviter les **redondances**, source de gâchis et, surtout, d'erreurs. Ce regroupement est la base de la modélisation.
- chacun des groupements (qui correspond, dans le MCD, à une entité) se traduira par une table dans la Base de Données.
- toute **entité est symbolisée par un rectangle**. Elle est nommée par un substantif au pluriel, désignant les éléments qu'elle contient
- les **propriétés** d'une entité correspondent aux **colonnes** de la table qui sera déduite de cette entité.
- les entités (c'est-à-dire : les individus présents dans les entités) peuvent être mises en rapport via des relations (ou associations)
- toute **relation est symbolisée par un ovale**. Elle est nommée par un verbe (on parle aussi d'**association**)



Vocabulaire et types
Enregistrements et champs

Dans une table, on évite de parler de « lignes » et de « colonnes ».

Les lignes correspondent aux différents individus, ou aux différents objets individuels, répertoriés dans une table : dans la table Disques, chaque ligne correspond à un de mes CD. Ces différents éléments individuels qui correspondent aux lignes sont appelés **enregistrements** (ou tuples).

Les colonnes, qui correspondent aux propriétés de l'entité dans le MCD, sont appelées des champs.

Typage des propriétés

Les informations contenues dans les entités (donc, dans les tables) vont devoir au bout du compte être codées numériquement afin d'être stockées sur un support informatique, sous un nom qui est celui de la propriété.

Pour chaque enregistrement, celle-ci se comporte donc comme un nom de variable… ce qui est somme toute logique, car à de menus détails près, c'en est une. Tout ceci nous amène au fait que les propriétés, à l'instar des variables, relèvent de certains types.

Dans le détail, les types disponibles pour les propriétés varient légèrement d'un système de gestion de bases de données à l'autre. En ce qui nous concerne, nous pouvons en rester à un niveau assez général, en considérant les types les plus courants :

- **numérique** : on y distingue systématiquement l'entier (« Integer ») du nombre à virgule (« Float »). Le type « AutoIncrement », souvent utilisé pour gérer les clés primaires, correspond à un entier dont la valeur est automatiquement attribuée à la création d'un nouvel enregistrement. Les bases de données proposent également toujours au moins un type Date/Heure.

- **texte** : on aura éventuellement différents types correspondant à différentes longueurs maximales du texte.

- **booléen** : 0 (FAUX) ou 1 (VRAI)

Les documents de modélisation, MCD et MLD, devront faire apparaître, pour chaque entité, le type de chaque propriété.

Exercice: Dans notre exemple de CDthèque déterminez les différentes entités (avec leurs propriétés et leurs types) et les relations correspondantes.